Rational use of cognitive resources in human planning

Frederick Callaway[a,1], Bas van Opheusden[a], Sayan Gul[b], Priyam Das[c], Paul Krueger[d], Thomas L. Griffiths[a,d,2], and Falk Lieder[e,2]

[a]Department of Psychology, Princeton University, Princeon, NJ
[b]Department of Psychology, University of California, Berkeley, CA
[c]Department of Cognitive Sciences, University of California, Irvine, CA
[d]Department of Computer Science, Princeton University, Princeon, NJ
[e]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[1]Corresponding author: fredcallaway@princeton.edu
[2]These authors share senior-authorship.

December 29, 2021

Abstract

Making good decisions requires thinking ahead, but the huge number of actions and outcomes one could consider makes exhaustive planning infeasible for computationally constrained agents, such as humans. How people are nevertheless able to solve novel problems when their actions have long-reaching consequences is thus a long-standing question in cognitive science. To address this question, we propose a model of resource-constrained planning that allows us to derive optimal planning strategies. We find that previously proposed heuristics such as best-first search are near-optimal under some circumstances, but not others. In a mouse-tracking paradigm, we show that people adapt their planning strategies accordingly, planning in a manner that is broadly consistent with the optimal model but not with any single heuristic model. We also find systematic deviations from the optimal model that might result from additional cognitive constraints that are yet to be uncovered.

One of the hallmarks of human intelligence is our ability to act adaptively in novel and complex environments. It is widely agreed that this ability depends critically on our ability to plan, that is, to use a model of the world to simulate, evaluate, and select among different possible courses of action. Research in psychology [1–7], economics [8–10] and computer science [11] has formalized planning as search over a "decision tree", where every decision one might have to make is represented as a branching point (see Figure 1A). In principle, one can identify the best plan by considering every possible decision point. However, traversing the full decision tree is infeasible because the size of the tree grows exponentially with the number of steps that one looks ahead.

The question of how people are able to effectively plan in the face of such formidable computational obstacles is of great interest for both researchers who wish to understand human intelligence and those who wish to recreate it [12]. In fact, one of the earliest attempts to replicate human-like intelligence in a computer, conducted by Newell and Simon, focused on problems that require thinking multiple steps ahead [11, 13, 14]. Even at this early stage, it was immediately recognized that the success of human planners (and any hope for success of artificial planners) depended critically on the use of heuristics to circumvent the exponential growth of search trees. Recent work on human planning has largely followed a similar vein, proposing and testing different possible heuristics people could be using to reduce the cost of planning. For example, people might limit the depth of their search [4–7], "prune" away initially unpromising courses of action [1, 2], or avoid planning altogether by relying on habit or "memoization" [1, 15]. Each of these models provides insight into how people circumvent the computational intractability of planning.

Despite these successes, the approach of postulating and testing specific heuristics faces several challenges. First, it is limited by the creativity of the researchers who must generate hypotheses about different possible heuristics people could be using. Second, it does not provide a straightforward way to predict which heuristics will be employed in new situations, or how each individual heuristic should be parameterized (e.g., How deep will someone plan in this environment? How large of a punishment will lead to a branch being pruned?) Finally, although these models are intuitively motivated as making planning more efficient, they do not provide a formal answer to the question of why people use these heuristics [16].

These challenges—hypothesis generation, generalizable prediction, and functional explanation—are not unique to planning; indeed, they arise in nearly all areas of cognition. In many domains, progress in addressing these challenges has been made by analyzing optimal solutions to the problem a cognitive system is meant to solve [17, 18]. This approach has generated insight into a wide range of problems, including decision-making [19], generalization [20], categorization [21, 22], perception [23], and information-seeking [24, 25]. More recently, the notion of optimality has been extended to account not only for the demands imposed by the external environment but also the demands imposed by our own cognitive limitations [26–30]. This approach dates back to Simon [31] and has been especially useful in the domain of decision-making, where it has been used to explain both how long people deliberate [32–36] and also what people think about [37, 38] while making "simple" (i.e., non-sequential) choices. However, to the best of our knowledge, there has been no such analysis in the domain of planning, despite the especially critical role that computational limitations play in this case (but c.f. [39, 40] for closely related efforts, which we discuss further below).

In this work, we propose an optimal model of planning under computational constraints. Drawing on the field of rational metareasoning in artificial intelligence [41–43], we formalize planning as a sequential decision problem in which an agent executes a sequence of cognitive operations to construct a decision tree. Formalizing planning in this way allows us to identify the optimal planning strategy for a given environment as the one that maximizes the expected utility of executing the resulting plan minus the cost of each cognitive operation used to make that plan. This also provides a flexible framework for specifying heuristic planning strategies in a highly precise and composable way. Every model we consider specifies an explicit distribution over the sequence of planning operations that will be executed in any given environment.

To rigorously test the fine-grained predictions of the optimal and heuristic models, we develop a novel process-tracing paradigm that externalizes the cognitive operations underlying planning as mouse clicks, extending the widely used Mouselab paradigm [44] to sequential decision-making problems. In a series of four experiments, we find that our participants use planning strategies that are largely consistent with optimal planning strategies, using previously proposed heuristics when they are adaptive, but adjusting their strategies when the structure of the environment changes. However, we also find systematic deviations from optimal planning, in particular a bias towards considering states in the order in which they would be traversed (forward search). Based on these results, we conclude that human planners use highly adaptive planning strategies, but that these strategies are also shaped by additional constraints that may reflect the specific cognitive mechanisms underlying human planning.

Results

Model

How can we formally characterize the problem of resource-constrained planning? One intuitive way to conceptualize this problem is in terms of a cost-benefit tradeoff [15, 45–48] in which an agent must find an optimal balance between the mental effort or time spent planning and the quality of the resulting decision. This type of model predicts, for example, that people will reduce the depth of planning under time pressure [5]. However, this one-dimensional simplification cannot capture the full range of different planning strategies people might employ. In particular, a planning strategy specifies not only the amount but also the direction of planning, that is, which courses of action are explored deeply and which are hardly considered at all [39]. To further complicate matters, it is not sufficient (or perhaps even possible) to determine in advance the amount and direction of planning. An adaptive planning strategy will dynamically adjust both based on the partial results of previous planning; for example, one can only prune away a branch of a decision tree after discovering a large punishment early on that branch [2].

To summarize, the problem of planning involves balancing between costs and rewards attained at different time points, by determining in which direction to plan (or to stop planning) based on the outcome of previous planning. That is, in addition to being a method for solving sequential decision problems, planning is itself a sequential decision problem (Figure 1). This insight has been formalized in the field of rational metareasoning [43], which casts planning (and reasoning more generally) as a sequential decision problem in which an agent performs a sequence of cognitive operations to update its beliefs about the world. In particular, we apply the framework of metalevel Markov decision processes (metalevel MDPs [49]). This allows us to apply the familiar conceptual and technical tools for MDPs to formalize the problem of planning and to identify optimal solutions to that problem. We provide a technical description of this model in Methods, giving an overview below.

A Markov decision process (MDP) is the standard formalism for modeling a sequential interaction between an agent and a stochastic environment. It is defined by a set of states, $\mathcal{S}$, a set of actions, $\mathcal{A}$, a transition function, T, and a reward function, r. The transition function specifies the dynamics of the environment (i.e., how taking actions moves the agent from

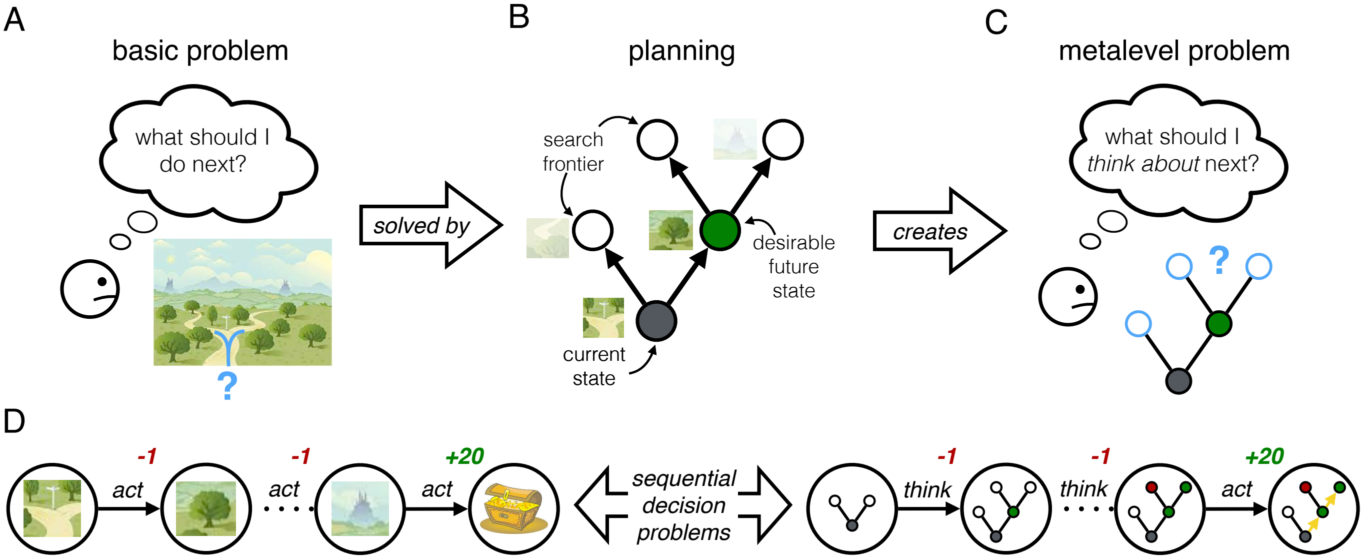

Figure 1: Formalizing planning under computational constraints. (A) The basic problem facing an intelligent agent is to take actions that maximize long-run reward. If the agent can predict the consequences of their actions, they can solve this problem by planning. (B) In one version of planning, the agent constructs a decision tree, where nodes (circles) represent possible future states of the world and edges (arrows) represent possible actions the agent could take. The agent constructs the tree by iteratively considering possible future states, estimating the reward to be gained there, and expanding the search frontier to include states that could be visited next. Eventually, this procedure will reveal the sequence of actions that maximize reward. But for an agent with limited cognitive resources, exploring the entire tree is usually infeasible. This creates the metalevel problem: (C) Which states should the agent consider—or ignore—in order to achieve the best tradeoff between the costs and benefits of planning? (D) The key observation underlying our model is that the basic problem and the metalevel problem are both sequential decision problems. That is, they require the agent to make a sequence of choices, in which the outcome of each choice depends on which choices were made previously. But while the basic problem is defined by states of the world, physical actions, and external rewards, the metalevel problem is defined by decision trees and the mental operations that build them; the metalevel rewards capture both cognitive costs and also the external reward gained by executing the chosen plan. By formalizing planning in this way —concretely, as a Markov decision process (MDP)—we can use standard MDP-solving techniques to identify optimal planning strategies.

one state to another) and the reward function specifies the goal, giving a scalar state-dependent reward for each action that the agent takes. The agent chooses actions to maximize cumulative reward using a policy, π, which specifies which action to take based only on the current state.

While a standard MDP formalizes the interaction between an agent and its external environment, a *metalevel* MDP formalizes the interaction between an agent and its internal, computational environment. The states correspond to beliefs, actions correspond to computations, and rewards capture both computational cost and decision quality. We now define our metalevel MDP model for planning.

A state in a metalevel MDP captures the agent's current knowledge about the problem being solved. To avoid confusion with the world state, we will refer to metalevel states as belief states. Following previous work [1–3, 39], we begin by assuming that the belief state corresponds to a decision tree; we relax this assumption later. As illustrated in Figure 1B, a decision tree represents a set of possible action sequences as a tree-structured directed graph, in which nodes correspond to hypothetical future states and edges correspond to actions that bring the agent from one state to another. Each node in the tree also corresponds to a plan to take the sequence of actions leading to that state and then act randomly until reaching a terminal state.

An action in a metalevel MDP corresponds to an elementary computation the agent can execute. In decision tree search, this computation is node expansion, an operation that determines the cost or reward for visiting a state, integrates that value into the total value of the path leading to that state, and adds the immediate successors of the target state to the search frontier, that is, the set of nodes that can be expanded on the next iteration. Thus, node expansion updates the decision tree in the same way that a physical action can change the state of the world. These dynamics (including the distribution of rewards that could be revealed at each node) constitute the metalevel transition function. In addition to expanding a node, the agent can decide to terminate planning at any moment. Upon terminating planning the agent executes an action sequence that has maximal expected value according to the decision tree it has built up until that point.

Finally, the metalevel reward function captures both the cost of computation and the quality of the decision that is ultimately made. Specifically, we assume that node expansion has a fixed cost, corresponding to the effort and time spent executing the operation. To capture decision quality, the reward for the termination action is the expected value of the external rewards one will attain while executing the chosen plan. The expected value of a plan is the sum of rewards up to and including the associated node, plus (for an incomplete plan) the expectation of the unknown future rewards. The chosen plan is the one that maximizes this expected value. Thus, the reward for the termination action is equal to the maximal expected value of any plan.

We have now specified all four components of a metalevel MDP for decision-tree planning. However, there are countless possible planning algorithms consistent with this general class. To create a complete model, we must specify one additional component: the strategy one uses to select which nodes to expand, and when to stop expanding nodes. Formally, this corresponds to a policy for the metalevel MDP, a distribution over computational actions for each possible belief state.

One policy of particular interest is the optimal policy, that is, the one that maximizes the expected total metalevel reward. On a given trial, the total metalevel reward is the external reward attained by executing the chosen plan minus the cost of the node expansions used to construct the plan. The optimal policy thus balances the costs and benefits of search, expanding the nodes that are most likely to improve one's ultimate decision, and only doing so when the expected improvement in decision quality outweighs the cost of expansion. In the terminology of MDPs, the optimal policy selects actions that maximize the optimal state-action value function, $Q(s,a)$. This function specifies the expected total reward an agent will receive (including both cost and decision quality) if she executes the node expansion action a in belief state s and continues selecting actions optimally until termination. Importantly, this function depends on the cost of node expansion; the optimal model's behavior is thus governed by one key free parameter (not including parameters of the error model used to fit human data; see Methods).

Exactly computing Q for a large MDP is very computationally intensive. Early work in rational metareasoning proposed that this function can be approximated by a myopic one-step lookahead [43]. This myopic policy chooses the planning operation that would be most helpful if the agent had to select a plan immediately afterward. Like the optimal model, this model has one key free parameter, the cost of node expansion.

We additionally consider "heuristic" policies based on three classical planning algorithms [50]. Breadth-first search first considers all immediate successors of the current state, then the successors of those states, and so on. That is, it prioritizes nodes that are close to the initial state. In contrast, depth-first search constructs a full plan to a terminal state before considering any alternative; it prioritizes nodes that are far from the current state. Finally, best-first search prioritizes nodes on promising paths, that is, nodes that lie on the frontier of plans with high expected value.

These classical algorithms specify the order in which nodes are expanded, but are agnostic about how people might decide when to stop planning. Previous research has proposed a number of heuristics people might use to reduce the amount of planning they must do to reach a decision. We consider four such heuristics. The "satisficing" heuristic terminates planning as soon as it finds a path whose expected value exceeds some predefined threshold [31]. The "best vs. next" heuristic terminates planning when one path's expected value is sufficiently greater than any other path's [51]. As discussed below, these two terms respectively correspond to absolute and relative stopping rules in evidence accumulation models. The pruning heuristic stops considering paths once their value falls below a predefined threshold [2]. The "depth limits" heuristic only considers states that can be reached in some predefined number of steps [4–7]. For brevity, we will refer to these heuristic mechanisms for limiting the amount of planning as simply "heuristic mechanisms". We assume that people could use any combination of these four mechanisms, resulting in $3 \times 2^4 = 48$ heuristic planning models (three search orders and sixteen combinations of heuristic mechanisms for each). The heuristic models have between 3 and 9 parameters depending on which mechanisms are included (see Methods).

Task: Mouselab-MDP

All the models we consider make precise predictions about the exact sequence of node expansion operations a person will execute while planning. The ideal way to test these predictions would be to compare them directly to the node expansion operations performed by people. Unfortunately, this is impossible because those operations are internal and unobservable.

Early work on human planning addressed this challenge using "think aloud" protocols in which participants narrate their planning process [14, 52, 53]. However, verbal reports are only indirectly related to the cognitive operations involved in planning and do not lend themselves well to precise quantitative modeling.

More recently, researchers have tried to infer people's planning algorithms based only on their external actions [1–3, 7, 45, 51]. However, the precise nature of a person's planning algorithm is generally only weakly constrained by their actions alone, because there are usually many sequences of planning operations that are consistent with each possible choice. Concretely, in the task illustrated in Figure 2 there are 8 possible choice sequences and over 2.7 billion node expansion sequences.

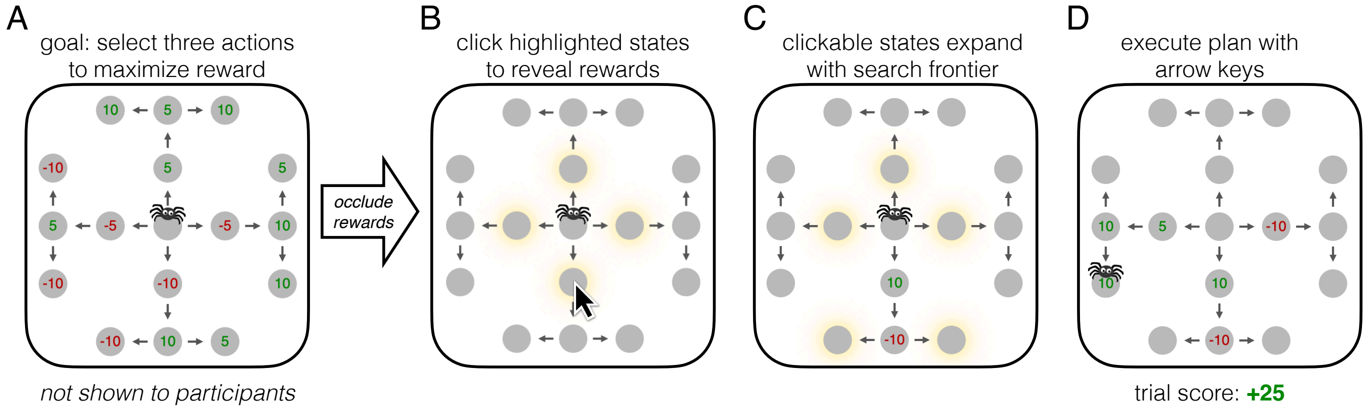
A
goal: select three actions to maximize reward
10
5
10
-10
5
5
5
-5
-5
10
-10
-10
10
-10
10
5
not shown to participants
occlude rewards
B
click highlighted states to reveal rewards
C
clickable states expand with search frontier
10
-10
D
execute plan with arrow keys
10
5
-10
10
10
-10
trial score: +25

Figure 2: Experimental task. (A) Participants are presented with a sequential decision problem displayed as a graph. Gray circles indicate states, arrows indicate actions, and green and red numbers indicate rewards and punishments. (B) Rewards are initially occluded, but can be revealed by clicking on the corresponding state. Only highlighted states can be clicked. (C) The clickable states expand with the search frontier, which includes all states adjacent to either the initial state or an already-clicked state. (D) At any point, participants can execute a plan by pressing a sequence of three arrow keys.

How can we collect fine-grained and precise data on human planning processes? A similar problem faced researchers studying how people make non-sequential decisions. To address this challenge, Payne and colleagues developed the Mouselab paradigm [44, 54], which traces participants' decision-making processes by requiring them to click to reveal decision-relevant information. In the original paradigm, participants clicked on cells in a table to reveal the payoffs associated with different outcomes of risky gambles. Here, we apply the same idea to multi-step decision problems, with participants clicking to reveal rewards at hypothetical future states.

The task, "Mouselab-MDP", is illustrated in Figure 2. On each trial, participants are presented with a route-planning problem, displayed as a graph. Each vertex in the graph (gray circles) corresponds to a future state the participant could visit, and harbors a reward or punishment (-10, -5, +5, or +10 with equal probability). The edges in the graph correspond to actions the participant can take to travel between states. The goal is to select a sequence of three actions that maximize the total reward. The potential gains and losses are initially occluded, but the participant can reveal them by clicking on the corresponding state, with the constraint that they can only click on states adjacent to the initial state or a previously revealed state. This constraint ensures that participants follow a forward-planning strategy, as has often been assumed in the literature [1–7]; we remove the constraint in Experiment 3. Each click was followed by a three-second delay.

Importantly, the task involves two types of sequential decision problems, both of which can be modeled as MDPs. The problem of moving the spider in the web is modeled as an MDP with 17 states (gray circles), four actions (key presses), and four possible rewards (-10, -5, +5, and +10). In contrast, the problem of selecting which potential rewards to consider when planning a route is modeled as a *metalevel* MDP, with over four billion possible states (patterns of revealed rewards), 16 actions (one for revealing each reward), and fourteen possible rewards (one implicit cost for the delay and thirteen possible path values, i.e., -30 to 30 in steps of 5).

Like its predecessor, Mouselab-MDP externalizes the core representations and operations underlying a cognitive process. In particular, our paradigm externalizes the decision tree as the graphical display, the node expansion operation as clicking, and the cognitive cost of that operation as the delay. While it is possible that externalizing a cognitive process in this way might alter the strategy people adopt, the extensive use of the original Mouselab paradigm [54–58] and the early advances made possible by a less structured form of process tracing [14, 52, 53] provide support for using this approach. We return to this point in the Discussion.

Experiment 1

In our first experiment, we sought to test the extent to which human planning is consistent with an optimal planning strategy in a relatively unstructured environment, illustrated in Figure 2A.

To evaluate participants' performance, we must consider both the scores they achieved as well as the amount of planning effort (i.e., clicking) that they expended. Figure 3A thus shows the average reward and number of clicks each participant made per trial. The blue line shows the Pareto front, the maximum average reward attainable for a given average number of clicks. On average, participants earned 0.92 fewer points than they could have with the same number of clicks. They earned 4.94 more than clicking randomly (95% CI [4.43, 5.44];

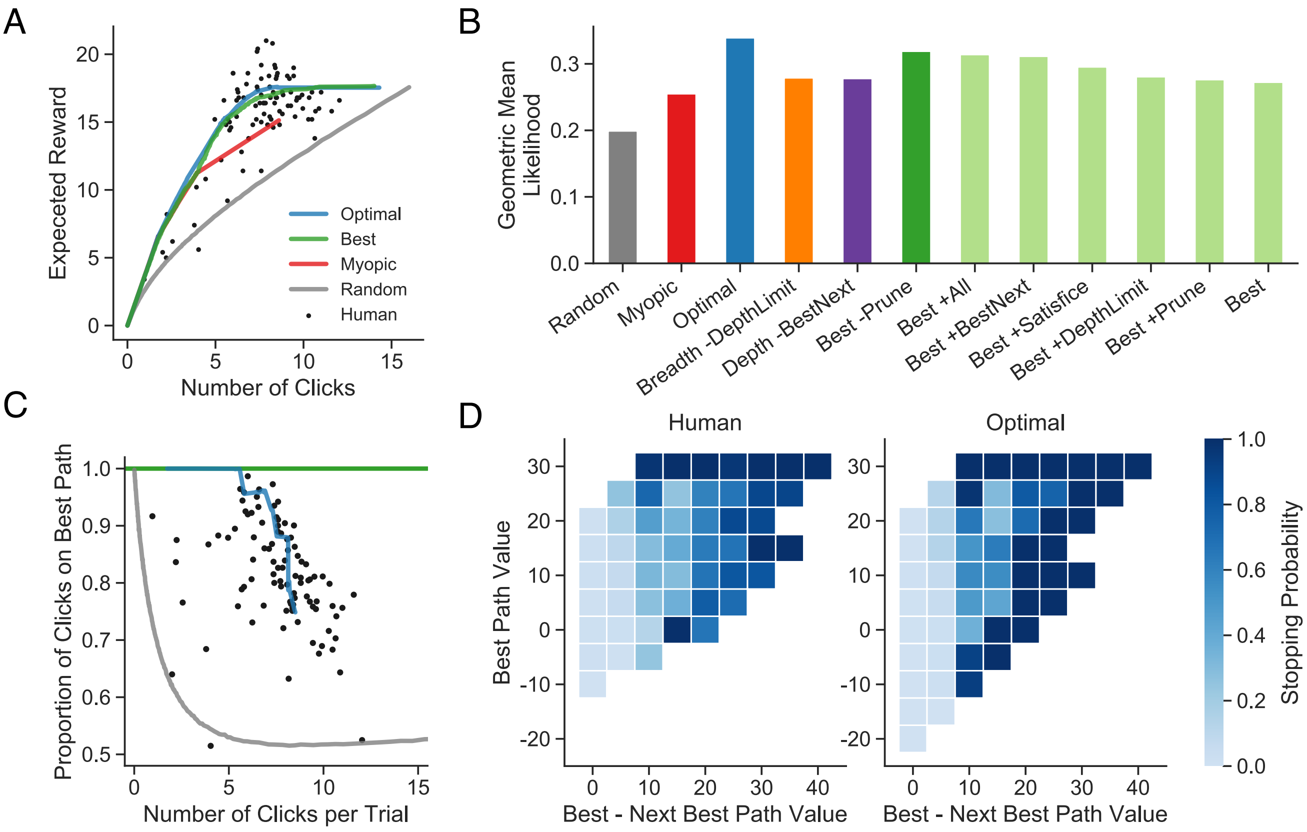

Figure 3: Experiment 1 results. (A) Pareto curves. Each point shows the average reward attained and number of clicks made by a participant (black dots) or model (colored lines). Note that with a small number of trials, it is possible to exceed the expected performance of the optimal model by getting lucky. (B) Model comparison. Bars show geometric mean likelihood (the total log-likelihood divided by the number of observations and then exponentiated) estimated on out-of-sample data. For the heuristic models, we indicate which heuristic components are present: +All indicates that all mechanisms are included, -Prune indicates that all mechanisms except for pruning are included. The best-fitting versions each heuristic model are shown in dark bars. Alternative best-first search models are shown in light green. Note that any visually detectable difference corresponds to a large difference in likelihood. (C) Selection rule. Proportion of clicks following a best-first strategy as a function of the average number of clicks per trial for each participant. Colors match panels A and B. Model predictions are made without fitted noise parameters. (D) Stopping rule. Probability that planning is terminated as a function of the value of the best path found yet and the difference in values of the best and next best paths. The right panel shows simulations from the noise-free optimal model. Cases in which all nodes have been clicked and termination is required are excluded.

Wilcoxon test: $z = 8.40, p < .001$). Confidence intervals are boot-strapped over participants and p-values are two-tailed (see Methods).

Selection rule: cost-dependent best-first We first considered the order in which the model expands nodes. Inspecting simulations of the optimal planning strategy across a range of costs (0.05 to 3.75, the maximum cost for which any planning occurs), we found that the optimal model expands a node on a path that has maximal expected value between 74.6% and 100% of the time, compared to 51.7% in the random clicking model. That is, optimal planning in this environment resembles best-first search. Consistent with this prediction, participants expanded a path with maximal expected value on average 81.5% of the time (95% CI [79.6, 83.3]; Wilcoxon test vs. chance: $z = 8.46, p < .001$).

However, the degree to which optimal planning conforms to best-first search depends on the cost parameter, with a closer match for higher costs. Intuitively, this is because the optimal planning policy expands nodes that are likely to lead to a quick decision. When the cost is high, a plan can be chosen when it is only moderately better than its competitors; the path that currently has maximal value is the most likely candidate. When the cost is low, however, a plan must be exceptionally good to justify stopping early; a path with moderately high value is actually less likely to provide such an outcome, compared to a completely unexplored path. As a result, the optimal model predicts that the degree to which people follow best-first search will decrease with the average number of clicks they make (the most direct behavioral correlate of the cost parameter). Figure 3C confirms this prediction (Spearman's $\rho = -0.481$, 95% CI [-0.66, -0.28], $p < .001$). The correlation also arises in the random model because all paths are "best" on the first click. However, controlling for the best-first rate of the random model, we still find a significant correlation ($\rho = -0.347$, 95% CI [-0.56, -0.12], $p < .001$).

Stopping rule: both absolute and relative By inspecting simulations of the optimal model with a range of costs matching that inferred from human participants, we found that the model was more likely to stop planning when it had found a path with high expected value, consistent with satisficing. However, its stopping decisions were more strongly influenced by the difference between the value of the best path and the next best path. That is, the optimal stopping rule depends primarily on the best path's relative value, but also on its absolute value.

As illustrated in Figure 3D, our participants' decisions to terminate planning were also sensitive to both the absolute and relative value of the best path. A mixed-effects logistic regression with random intercepts and slopes for each participant revealed significant effects of both terms (best path value: $\beta = 0.82$, 95% CI [0.69, 0.94], $z = 12.89$, $p < .001$; best vs. next: $\beta = 1.68$, 95% CI [1.52, 1.84], $z = 20.70$, $p < .001$). However, compared to the coefficients for the optimal model ($\beta = 0.99$, 95% CI [0.84, 1.15] and $\beta = 4.64$, 95% CI [4.02, 5.26]), people appear to be under-sensitive to relative value (note that the confidence intervals for the optimal model are not negligible due to the mixed-effects structure; predictors are standardized by their mean and SD in the human data).

These results are broadly consistent with evidence accumulation models of non-sequential decisions, where relative stopping rules (specifically best vs. next) generally perform better, both in terms of fitting data [59, 60] and maximizing accuracy [32, 61]. However, although both the model's and our participants' stopping decisions were primarily driven by relative value, absolute value also played a role. This raises the intriguing possibility that people could be using a hybrid stopping rule in simple value-based choices as well.

Model comparison Having characterized the qualitative matches and mismatches between participant and optimal behavior in the task, we next sought to quantify the ability of the optimal and heuristic models to predict human behavior quantitatively. We fit our models to participants at the individual level and obtained out-of-sample predictions using five-fold cross-validation. We used the total log-likelihood (LL) across all five folds as a measure of model performance. Note that this metric accounts for the flexibility of the different models without relying on parameter counting (as do AIC and BIC), which can be a poor measure of flexibility [62]. Differences in this cross-validated log-likelihood (Δ_{LL}) can be interpreted similarly to differences in AIC: $\Delta_{\text{LL}} = 1$ is roughly equivalent to $\Delta_{\text{AIC}} = 2$.

Figure 3B shows the predictive accuracy achieved by each of the models. The optimal model clearly outperforms the random, myopic, breadth-first, and depth-first models (all $\Delta_{\text{LL}} > 3981$). In terms of total likelihood, it also outperformed best-first search (all $\Delta_{\text{LL}} > 1250$), although 41 participants were best fit by the one of the best-first models vs. 45 by the optimal model (9 by some other model). Importantly, given that the best-first model achieved a near-optimal reward-effort trade-off (Figure 3A), a substantial majority of participants were best fit by an optimal or near-optimal model.

Experiment 2: Adapting to the environment

In Experiment 1, we found that participants seemed to use a best-first search strategy that was well-suited to the task environment. However, this does not mean that people always plan in this way. On the contrary, a key prediction of the optimal model is that people adapt their strategy to the structure of the environment. We tested this prediction in Experiment 2.

To investigate the effect of environment structure on human planning strategies, we constructed three new experimental environments (see Figure 4A). The environments have the same transition structure (four independent paths with five steps each) but different reward distributions. In the "constant variance" environment all states had the same reward distribution, as in Experiment 1. In the other two environments, most states had low variance; extreme rewards could only be found in one state on each path. In the "decreasing variance" environment extreme rewards were possible only in the first state on each path. In the "increasing variance" environment extreme rewards were possible only in the last state.

We designed these environments to produce clear qualitative differences in the predictions of the optimal model. Specifically, in each environment, the optimal planning strategy resembles a different classical planning algorithm: breadth-first for decreasing variance, best-first for constant variance, and depth-first for increasing variance. As illustrated in Figure 4B, each algorithm is approximately optimal in its respective environment, but suboptimal in the other two.

If people indeed adapt their planning strategy to the environment, we should find that, out of these three classical search models, the model that achieves the best reward-effort trade-off should also predict human behavior best. Figure 4C confirms this prediction (all $\Delta_{\text{LL}} > 446$). For the classical search models, we used the combination of heuristic mechanisms that achieved the best likelihood across all conditions; however, we excluded depth limits from this analysis because they allow the best-first and depth-first models to mimic breadth-first search. With the unrestricted set of heuristic models, the optimal model best predicts human

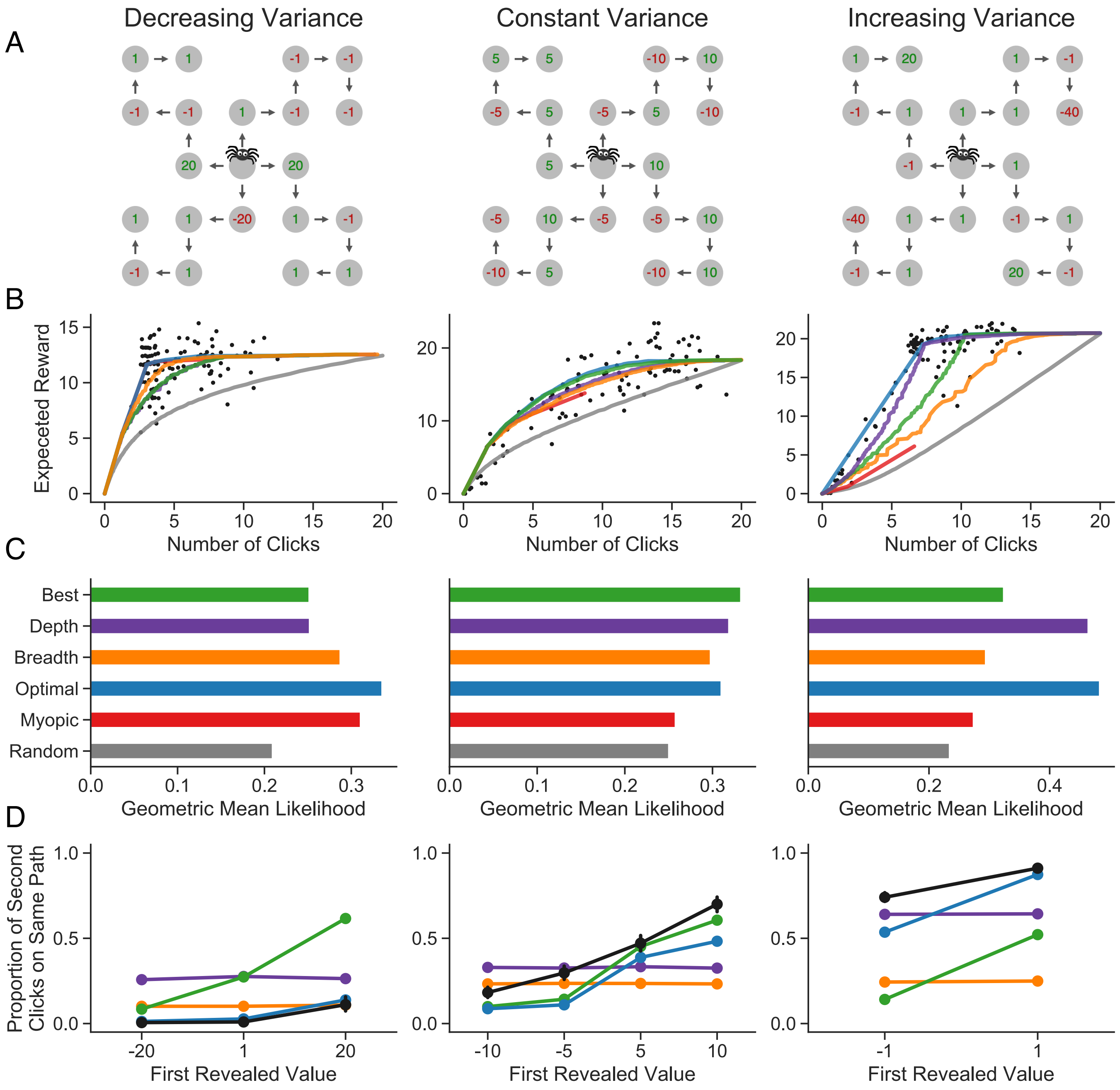

Figure 4: Experiment 2. Each column shows one experimental condition. (A) Example trials. Each condition is characterized by a different location-dependent reward distribution in which large values are found at the beginning of each path, at any location, or at the end of each path, respectively. (B) Pareto curves. Each point shows the average reward attained and number of clicks made by a participant (black dots) or model (colors match panel C). In each condition one classical algorithm achieves near optimal performance. (C) Model comparison. Best, Depth, and Breadth refer to the versions of the model that performed best in Experiment 1, as shown in Figure 3. Of these classical algorithms, the one that achieves the best reward-click trade off (shown in panel B) also best predicts human behavior. Depth limits are excluded because they allow the best-first and depth-first models to mimic breadth-first search. (D) Behavioral indicator of planning strategy. Each panel shows the probability of making a second click on the same path as the first, depending on the value revealed by that first click. Human data is in black and model colors match panel C. Error bars show bootstrapped 95% confidence intervals. In each condition one of the classical planning strategies captures the qualitative behavioral pattern, but only the optimal model captures the pattern in every condition. All heuristic mechanisms are excluded from this plot. See Supplementary Figure 1 for the same plot with the full models (including Myopic).

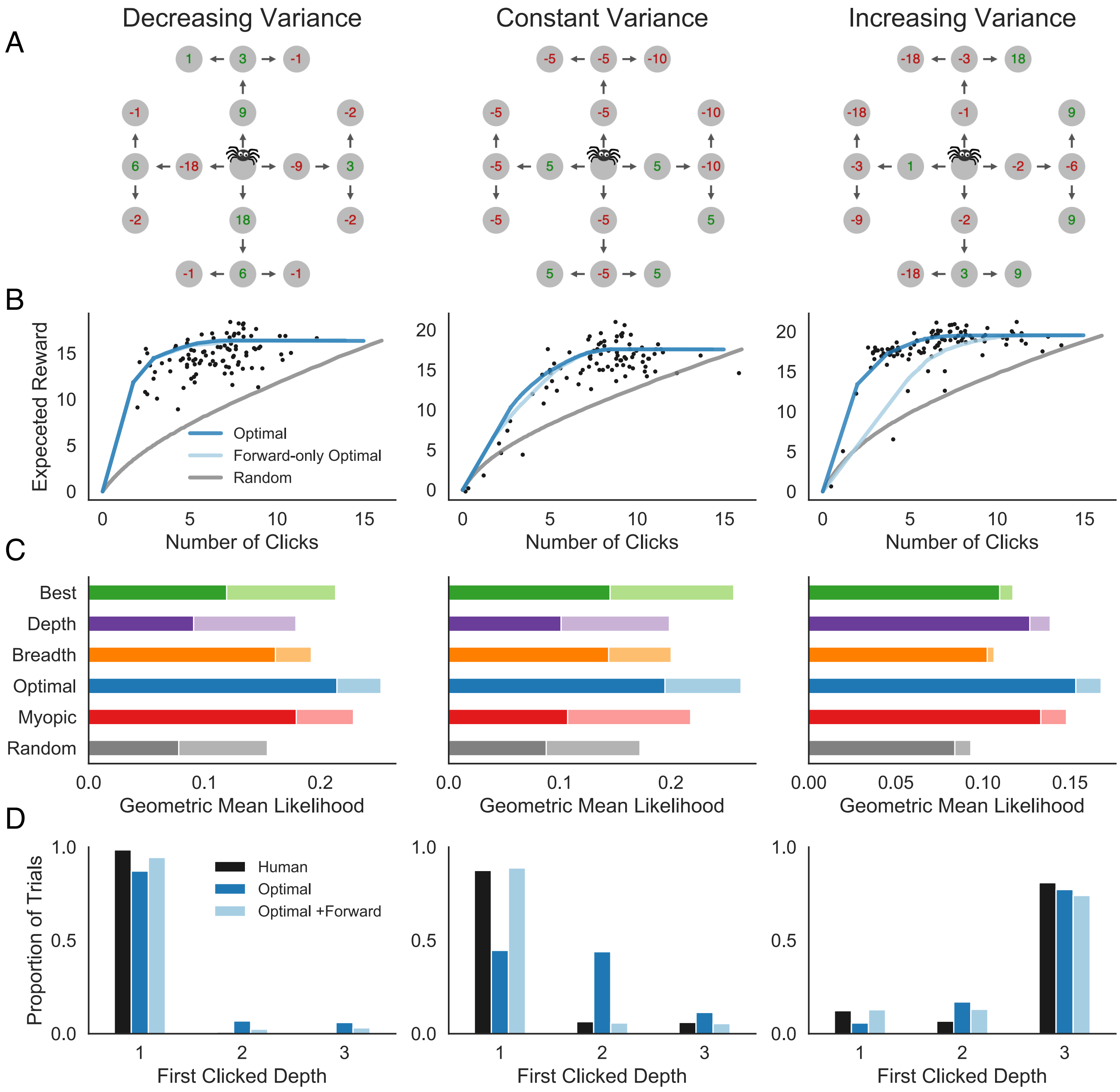

Figure 5: Experiment 3 results. (A) Example trials. Each condition is characterized by a different location-dependent reward distribution with standard deviation linearly increasing, decreasing, or remaining constant with depth. (B) Pareto curves. The light blue line shows the optimal model restricted to plan forwards. (C) Model comparison. Light bars show the performance of the corresponding model with a fitted degree of forward-search bias (including the no-bias model and forward-only model as special cases). (D) Behavioral indicator of forward and backward planning. Each panel shows a histogram of the depth of the first clicked state, in the data and in simulations from the optimal model with and without a forward-search bias. Although participants use forward-search by default (center), they switch to backward search when the environment encourages this strategy (right).

behavior in the increasing ($\Delta_{\text{LL}} = 606$) and decreasing ($\Delta_{\text{LL}} = 1276$) conditions; the best-first model with best vs. next fits best in the constant condition (compared to optimal: $\Delta_{\text{LL}} = 2150$).

Figure 4D demonstrates the shift in planning strategy with a simple behavioral measure. Considering only trials on which at least two clicks were made, we can ask how often people use their second click to continue down the path that they began with their first, depending on the value revealed by that first click. An overall tendency to continue down the same path is consistent with a depth-first strategy, the reverse tendency is consistent with a breadth-first strategy, and high sensitivity to the revealed value is consistent with a best-first strategy; we illustrate this by plotting the predictions of the basic search models without any heuristic mechanisms. Participants in each condition show the same pattern as the adaptive search order.

Experiment 3: Backwards planning

In the previous experiments, we constrained participants' planning strategies to variations of decision tree search by only allowing them to click on states adjacent to the initial state or a previously-clicked state. However, people may sometimes use planning strategies that are not constrained in this way. For example, they may plan backward from a goal as in means-ends analysis [14] or they may even consider states in arbitrary order [63]. Experiment 3 thus investigated a broader class of possible planning algorithms by lifting the forward-planning constraint, allowing participants to click any state at any point.

As in Experiment 2, we used environments with decreasing, constant, and increasing variance. For this experiment, we employed the transition structure from Experiment 1 and decreased or increased the reward variance exponentially with depth. The constant variance condition used the same reward distribution as Experiment 1. See Figure 5A for examples.

The key prediction of the optimal model is that participants will adopt a backward-planning strategy in the increasing variance condition, considering terminal states first and then working towards the initial state. Consistent with this prediction, participants in this condition were most likely to click a terminal state first (Figure 5D, right).

However, we also see a systematic deviation from the optimal model predictions. In the constant variance case (Figure 5D, center), the model is completely neutral between depth-one and depth-two states because they provide equivalent information about the optimal path. In contrast, participants showed a strong tendency to click a depth-one state first. More generally, participants in the constant-variance condition showed a consistent bias for forward search, clicking a state whose parent had already been revealed 92.4% of the time compared to 75.5% in the noise-free optimal model simulations (95% CI [86.2, 94.4], Wilcoxon test vs. optimal $z = 5.32, p < .001$). Importantly, however, such a bias was not maladaptive as indicated by the strong performance of a strictly-forward planning strategy (Figure 5B center).

Figure 5B shows that augmenting the models with a forward-search bias improves predictive accuracy considerably. Whether or not we incorporate the bias, the optimal model predicted human behavior best in every condition (with bias: all $\Delta_{\text{LL}} > 509$). Note that it is not clear how to extend pruning and depth limits when non-adjacent nodes on a single path can be expanded; thus, we do not include these mechanisms for this analysis.

Experiment 4: Planning a road trip

In Experiment 4, we tested the ability of the optimal model to generalize to a new task environment. In this new task, illustrated in Figure 6A, participants acted as travel agents, planning a route from an initial city to a goal city and minimizing the price of hotels that must be visited along the way. Participants were informed that hotels could cost \$25, \$35, \$50, or \$100 (with equal probability), but to see the actual price of the hotel in a city they had to type its name into a search box.

Although the task has the same formal structure as that used in Experiment 3 (allowing us to use the same models), there are three important dimensions on which the new task differs from the previous ones. First, rather than allowing participants to plan an arbitrary path, they were required to reach a specific destination; second, the transition structures were not limited to trees—that is, there could be multiple ways to reach a given state; third, the distribution of possible costs did not have a mean of zero, making it necessary to account for expected future cost when estimating the value of an incomplete plan. This task thus provides a non-trivial test of the model's ability to generalize.

As illustrated in Figure 6B, the optimal model most accurately predicted human behavior when the bias for forward search was taken into account ($\Delta_{\text{LL}} = 295$). Interestingly, the forward search bias is so important for capturing behavior that when we remove it, the breadth-first model (which follows forward search by default) performs best.

However, the tendency towards forward search was not without exception. Participants violated forward search by looking up a city without a revealed parent 7.2% of the time. Figure 6C shows that these exceptions were not random: participants were more likely to violate forward search when doing so was more valuable (logistic regression with random slopes and intercepts for each participant: $\beta = 2.48$, 95% CI [1.59, 3.37], $z = 5.48$, $p < .001$).

Discussion

In this paper, we proposed a rational model of resource-constrained planning and compared the predictions of the model to human behavior in a new process-tracing

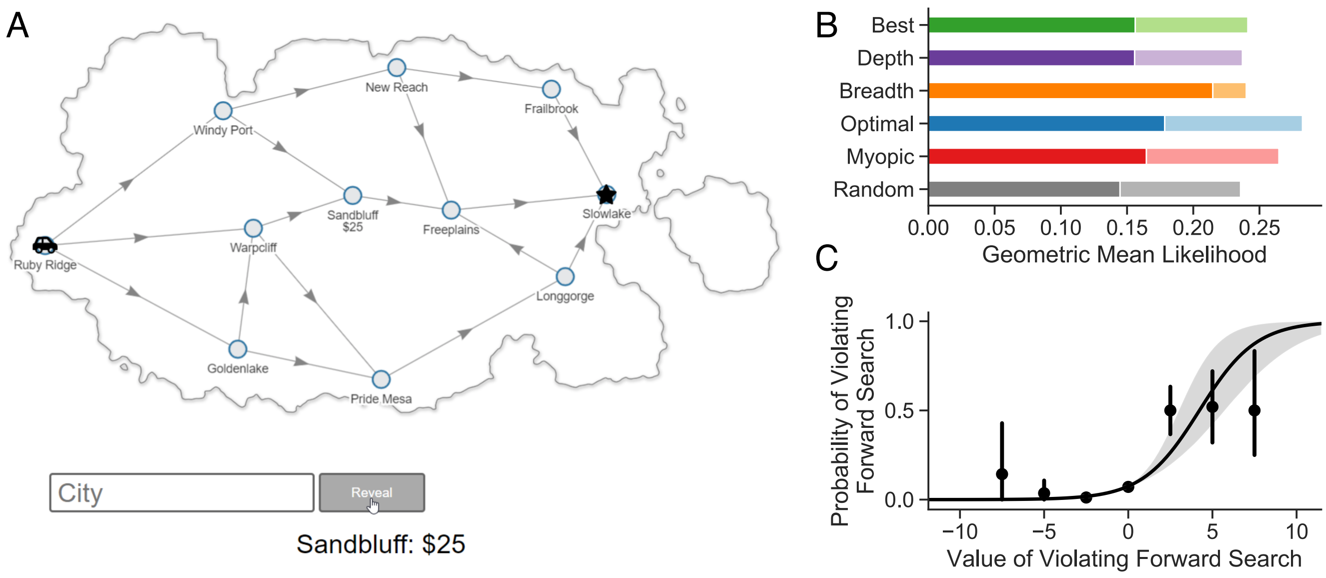

Figure 6: Experiment 4 results. (A) Task: participants acted as a travel agent, attempting to find a low-cost route from a start city to a goal city. They could reveal the price of passing through each city using a textual search interface. (B) Model comparison. Light bars show models augmented with a forward-search bias. (C) The probability of a participant inspecting a city without a revealed parent (i.e., violating forward search) as a function of the value of doing so. This value is defined as the maximal Q value for expanding a node not on the frontier minus the maximal Q value for expanding a node on the frontier (see Methods). The line shows a logistic regression fit and points show binned means. The shaded regions and error bars show 95% confidence intervals.

paradigm. Our results suggest that human planning strategies are highly adaptive in ways that previous models cannot capture. In Experiment 1, we found that the optimal planning strategy in a generic environment resembled best-first search with a relative stopping rule. Participant behavior was also consistent with such a strategy. However, the optimal planning strategy depends on the structure of the environment. Thus, in Experiments 2 and 3, we constructed six environments in which the optimal strategy resembled different classical search algorithms (best-first, breadth-first, depth-first, and backward search). In each case, participant behavior matched the environment-appropriate algorithm, as the optimal model predicted.

The idea that people use heuristics that are jointly adapted to environmental structure and computational limitations is not new. First popularized by Herbert Simon [31], it has more recently been championed in ecological rationality, which generally takes the approach of identifying computationally frugal heuristics that make accurate choices in certain environments [64–67]. However, while ecological rationality explicitly rejects the notion of optimality [68], our approach embraces it, identifying heuristics that maximize an objective function that includes both external utility and internal cognitive cost. Supporting our approach, we found that the optimal model explained human planning behavior better than flexible combinations of previously proposed planning heuristics in seven out of the eight environments we considered (see Supplementary Table 1).

Why did the optimal model generally explain human behavior better than the heuristic models? One possibility is that the optimal model has a more sophisticated stopping rule, informed by the full distribution of possible rewards, not just the expected values of different paths. Indeed, augmenting the heuristic models with distributional variants of the best vs. next and satisficing rules improved fit substantially (see Supplementary Information). However, the optimal model still achieved a better fit in all but two cases (constant variance in Experiments 2 and 3).

The increasing variance environments in Experiments 2 and 3 provide an especially interesting test of the model. In these environments, distal rewards are more extreme than proximal ones, and so the optimal model considers these states as soon as possible. In contrast, a classic finding is that people tend to neglect long-term consequences [69], suggesting that people might fail to consider those distal states in their planning. We found that people's clicking was consistent with the optimal model. In Experiment 2, they ignored small short-term losses to more quickly find large long-term rewards (Figure 4D), and when we lifted the forward-planning constraint in Experiment 3, people considered the final states first (Figure 5D). A potential reason why people were more far-sighted in our experiments than they are in some real-world situations is that our experiment allowed them to learn about the structure of the decision environment and adapt their decision strategy to it through intensive practice with immediate, reliable performance feedback that is often unavailable in the real world [70]. Consistent with this, people did show a strong bias to consider proximal rewards first when the environment did not strongly incentivize a different strategy (Figure 5D, center).

The ways in which our participants deviated from the optimal model are equally—if not more—informative than the ways in which they were consistent [16]. Using the approach of resource-rational analysis [29, 30], we can use the observed discrepancies to generate hypotheses about additional constraints (internal or external) that shape human planning strategies. That is, people's cognitive resources might be more limited than the model assumes and they may be adapted to an environment that differs from our artificial experimental task in important ways.

We found the most striking deviation from the optimal model's predictions in Experiments 3 and 4, where we observed a strong bias for forward search when it was not adaptive (nor clearly maladaptive, see 5B). This suggests that people's default representation of plans is temporally ordered, and that representing or computing information which does not fit this temporal structure is cognitively costly. There are two reasons such a representation might be preferred. First, in many (but not all) natural environments, the set of states one could feasibly reach is not clear in advance; one can only discover such states by forward search. In these cases, the standard assumption that people can only search in the forward direction [2, 3, 5, 7] may be appropriate. Second, in many domains, people likely have generative models of the world [71, 72]; given such models, one can directly simulate the consequences of an action, but one must infer what action could have led to a given consequence. In these cases, forward search will be less costly than backward search, but still possible; this is consistent with Figure 6C.

One important limitation of our work is that externalizing planning, as our task does, may alter the internal process that we wish to measure [73]. Nevertheless, there are at least five reasons to believe that the present results already reveal something important about human planning. First, the paradigm is a direct extension of the Mouselab paradigm, which has been widely used in the multi-attribute and risky-choice literature [54–58]. Second, our Experiment 1 results replicate previous findings that suggest that participants use a best-first strategy [3] (or, similarly, avoid nodes following large losses [2]) in the absence of environmental structure that a different algorithm could exploit. Third, we found that people show a bias for forward search even when the task does not require or even encourage it; this suggests that participants are carrying over a strategy that they have developed for naturalistic planning (where such a bias is likely adaptive, as discussed above). Fourth, recent work has noted a parallel between planning and information-seeking (our task could be characterized as the latter), suggesting that similar neural mechanisms may underlie both behaviors [74]. Finally, measuring how people plan in the absence of working memory constraints provides a useful comparison point for future work investigating how these constraints shape human planning strategies.

Comparing human and optimal planning in a more naturalistic paradigm is thus a critical step in future research. One promising approach is to use reaction time in a secondary task as a signal of previous planning (e.g., choosing between a subset of actions [75], replanning after a random teleportation [76], or determining whether a specific state falls on the optimal path [77]). Another approach would be to use eye- or mouse-tracking with a display that only reveals the reward at future states, but not the transition function. However, deploying these paradigms would also require augmenting the model to account for constraints on working memory and imperfect knowledge of the transition function—important but challenging directions for future work.

A second limitation of our work is that we only consider deterministic environments. This assumption greatly simplifies the task of identifying optimal strategies; in particular, it ensures that it is optimal to do all planning before taking any actions, allowing us to avoid the complexities associated with interleaving planning and action. Although we enforced this plan-then-act structure in our main experiments, a follow-up experiment (Supplementary Experiment 5) found that participants rarely violate this ordering when allowed to do so (3.9% of trials). However, in stochastic environments, planning far ahead may be wasteful because an unexpected transition can render much of that planning irrelevant. In such cases, it may be optimal to take an action and see its result before planning further ahead. Investigating how people adapt their planning strategies in unpredictable environments is thus an important direction for future work.

A third limitation is that we only consider problems with small, unstructured state spaces. This contrasts with early work exploring human planning in massive state spaces with rich internal structure, such as propositional logic [14]. Although this limitation applies equally to most recent empirical studies of human planning, future work should explore the strategies people use to plan efficiently in more complex environments.

Taken together, these three limitations put important limits on the conclusions we can draw from our results. Although we have shown that human planning can be quite close to optimal in simple environments without working memory constraints, it remains unclear whether people will be able to plan as effectively in more complex domains when working memory is limited. Nevertheless, our results do suggest that models of efficient use of limited cognitive resources may be a good starting place when developing theories of planning in these more naturalistic conditions.

A final limitation of our work is that we do not provide a process-level theory for how people are able to approximate optimal planning. One plausible hypothesis is that people use a myopic approximation, considering the immediate value of expanding a node while disregarding the potential for future node expansions. Indeed, such an approximation has been employed in two recently proposed models of human planning [39, 40]. However, we found that this model generally performed poorly, both in terms of reward (Figures 3A and 4B) and predicting human behavior. Another hypothesis is that people learn effective planning strategies through experience [78, 79]. However, the mechanisms that allow this learning to proceed so rapidly given the large state spaces of metalevel MDPs are still not well understood.

Over the past decades, the assumption that humans are well-adapted to their environment [17, 18] has facilitated rapid progress in many psychological domains [19–22, 24, 25]. However, the constraints imposed by the external environment are insufficient to explain many key features of human cognition [30, 80]. By additionally considering the constraints imposed by our limited cognitive resources—that is, our internal environments—we can apply the tools of rational modeling to a much broader set of cognitive phenomena [29, 30]. In this work, we have presented the beginning of such an analysis for planning. We anticipate that more precise characterizations of the cognitive constraints that shape planning will yield a correspondingly deeper understanding of this remarkable human ability.

Methods

All experiments can be viewed exactly as they were given to participants and in abbreviated form at https://webofcash.netlify.app. All experiments were approved by the institutional review board of Princeton University, and all participants gave informed consent. Each participant could only participate in one experiment (including pilots). For all experiments, we aimed to collect 100 participants per condition. We did not conduct a formal power analysis because all our hypothesis tests were highly significant in pilot samples. All reported statistics, model comparisons, and figures were pre-registered (Experiment 1: https://aspredicted.org/jd8rs.pdf, Experiment 2: https://aspredicted.org/w4kt2.pdf, Experiment 3: https://aspredicted.org/2cr5k.pdf, Experiment 4: https://aspredicted.org/wq87z.pdf). We describe deviations from the pre-registered analysis plan in the SI.

Experiment 1

We recruited 104 participants from Prolific who reported fluency in English, resided in the United States, and had a 95% approval rating (this number excludes participants who accepted the study but did not move past the second instruction page). We excluded 6 participants because they failed a quiz following the instructions and 3 participants who did not complete the experiment for some other reason, leaving 95 participants in the analysis. Participants who completed the experiment or failed the quiz received \$1.50 for participation. Those who completed the experiment additionally earned a performance-dependent bonus of (mean ± sd) \$2.43 ± \$0.42 for 22.5 ± 6.6 minutes of work.

Main Task In the main task of Experiment 1 (see Figure 2), participants navigated a cartoon spider through a directed graph in which each vertex (the gray circles) harbored a gain or loss, with the goal of maximizing the total payoff accrued along the selected route. All rewards were independently drawn from a discrete uniform distribution over the values $\{-10, -5, +5, +10\}$. At the beginning of each trial, all rewards were occluded; however, participants could click on nodes adjacent to the starting location or to an already-revealed node to reveal the value. After each click, there was a three-second delay during which no additional clicks could be made. To visually convey these constraints, nodes were highlighted whenever they could be clicked. At any point, participants could stop clicking and move the spider from the starting node using the arrow keys. After each arrow key press, the spider moved to an adjacent node, the value of that node was revealed (if not already revealed), and its value was added to a total shown in the top right. Clicking was disabled after the first move, and the trial ended when the participant reached a terminal node (i.e., one with no outgoing edges).

Procedure The experiment began with an instruction phase in which participants completed increasingly complex versions of the task. First, they were told the basic goal of selecting paths to maximize the amount of "money" acquired, and completed three trials with the rewards fully revealed. Second, they were told the reward distribution and shown ten examples where they did not make choices. Third, they completed one trial with the rewards occluded (i.e., guessing randomly). Fourth, they were told that they could click nodes to reveal the values, and completed three trials in which they had to make at least five clicks. Finally, they were told the conversion between in-game currency and their bonus (1 US cent for every 2 points) and completed three practice trials of the full task.

After completing the instructions, participants took a multiple-choice quiz that asked about the reward distribution, the rules for inspecting nodes, and the points-to-bonus conversion. Participants who failed the quiz were shown a review screen with all the necessary information and were given another chance to complete the quiz. If they failed the quiz three times, they were dismissed. Otherwise, they progressed to the main phase of the experiment where they completed 25 trials of the main task. They were given an initial endowment of 100 points to minimize the chance that they would ever have a negative score.

Experiment 2

All aspects of the design were identical to Experiment 1 except where noted otherwise. We recruited 313 participants. We excluded 4 who failed the quiz and 11 who did not complete the experiment, leaving 298 participants in the analysis. Participants received \$1.50 plus a bonus of \$2.18 ± \$0.74 for 23.6 ± 10.4 minutes of work.

Main Task The main task of Experiment 2 had the same basic structure as that in Experiment 1, but with a different graph and reward structure (see Figure 4A). The graph had a single choice point at the first move (four options) followed by four forced moves. The reward distributions depended on a between-participant condition. In the constant variance condition, it was the same as in Experiment 1. In the other two conditions, most nodes were -1 or +1 with equal probability, but four nodes had an extreme distribution. For increasing variance, the terminal nodes (farthest from the initial location) had values of +20 with 2/3 probability and -40 with 1/3 probability. For decreasing variance, the nodes closest to the initial location had value +1 with 3/5 probability, and either +20 or -20 with roughly 1/5 each, slightly skewed towards -20 to make the expected reward 0 (.185 and .215). These distributions were selected to make the optimal planning strategy closely resemble depth-first and breadth-first search in the increasing and decreasing variance conditions respectively.

Procedure The procedure was identical to Experiment 1 except that we replaced the bonus question with a question asking on which nodes the maximal reward could be found.

Experiment 3

All aspects of the design are identical to Experiment 2 except where noted otherwise. We recruited 319 participants. We excluded 11 who failed the quiz and 17 who did not complete the experiment, leaving 291 participants in the analysis. Participants received \$1.50 plus a bonus of \$2.49 ± \$0.43 for 21.3 ± 7.5 minutes of work.

Main Task The task had the same basic structure and graph as Experiment 1. The key difference from previous experiments is that we lifted the restriction that only nodes adjacent to the initial state or already-revealed nodes could be revealed. That is, participants could reveal any unrevealed node at any point. The graph was the same as in Experiment 1. The reward structure varied by condition. In the constant variance condition, it was identical to Experiment 1. In the increasing variance condition, the reward distribution for depth 1 nodes was uniform over the values $\{-2,-1,+1+2\}$. The possible values at later depths were scaled by 3^d; that is, the range and standard deviation increased by a factor of 3 from each depth to the next, up to $\{-18,-9,+9+18\}$ at the depth 3 leaf nodes. In the decreasing variance condition, the situation was exactly reversed: depth 1 nodes could take values in $\{-18,-9,+9+18\}$, and the values decreased by a factor of 3 with each step down to $\{-2,-1,+1+2\}$ at the leaf nodes.

Procedure The procedure was identical to Experiment 2.

Experiment 4

We recruited 137 participants. We excluded 7 who failed the quiz and 37 who did not complete the experiment, leaving 93 in the analysis. Due to a technical error, instruction progress was not recorded, hence the larger number of incompletes. Participants received \$1.75 plus a bonus of \$0.99 ± \$0.13 for 18.2 ± 7.8 minutes of work.

Main Task Participants assumed the role of a travel agent planning a road trip. On each trial, the participant saw a map of an island with eleven cities represented as circles and roads represented as arrows. Participants were instructed that the client wants to travel from a given starting location to a goal location. Each "day" they can move along any single arrow between two cities and each "night" the client has to stay in a hotel at a price that varies between cities. Participants were informed that hotels could cost \$25, \$35, \$50, or \$100, and that all values were equally likely. To reveal the price of the hotel in a given city, participants had to type its name into a text box. They could uncover any number of prices, in any order, and they could submit their recommended route at any moment. At this point the total cost was computed; this value was subtracted from a budget of \$300 and the participant's bonus for the trial was 1 cent for each \$10 remaining.

Procedure The experiment began with an instruction phase in which the task was explained through verbal instructions and images. Participants were required to complete a quiz (in no more than three attempts) before continuing. Each participant then performed 8 trials, the first of which was a practice trial that did not count towards their bonus payment.

Metalevel MDP

An overview of the model is given in the main text; here we provide a formal definition. A metalevel MDP is defined as a tuple $(\mathcal{S},\mathcal{A},T,r)$ where $\mathcal{S}$ is a set of belief states, $\mathcal{A}$ is a set of computational actions, T is the metalevel transition function, and r is the metalevel reward function. We now specify each component in turn.

For the first two experiments, the metalevel state space, $\mathcal{S}$, is the set of all decision trees that can be constructed in a given environment. We make the simplifying assumptions that the external environment is itself tree-structured and known to the agent. Thus, the largest possible decision tree has the same graphical structure as the environment itself. Let N be the size of this tree. We can then represent a decision tree as a vector s of length N where each position corresponds to a node in the tree (and a world state). The values, s_i, specify either the reward that can be attained at the world state i, or a special value, $\oslash$, indicating that the corresponding node has not been expanded yet. In the initial belief state, only the root node (the current world state) has been expanded, always having value 0; all other nodes have value $\oslash$.

The metalevel action space, $\mathcal{A}$, includes an expansion action, a_i, for each node, i, as well as the termination action, $\perp$. Note that in Experiments 1 and 2 an expansion action may only be executed if the corresponding node is in the search frontier, which is defined as the set of unexpanded nodes whose parent node has been expanded, formally

$$\text{frontier}(s) = \{a_i \mid s_i = \oslash \wedge \text{parent}(s_i) \neq \oslash\}. \tag{1}$$

For ease of notation, we define frontier(s) as the set of allowable node expansion actions rather than the nodes themselves.

The metalevel transition function specifies the effect of node expansion on the decision tree. Executing a_i produces s' that is identical to s except that s'_i is set to a reward which is sampled from a node-specific distribution, R_i. Additionally, executing the termination action, $\perp$, always results in a unique terminal belief state, $s' = s_\perp$. Note that this generative specification implicitly defines a probability distribution $T(s' \mid s, a)$.

The reward function, r, captures the cost of each expansion, as well as the expected external reward that will be attained when a plan is executed. Formally,

$$r(s,a) = \begin{cases} \max_{p \in \mathcal{P}} V(s,p) & \text{if } a = \perp \\ -\lambda & \text{otherwise} \end{cases} \tag{2}$$

where λ is the cost of node expansion (a free parameter), p is a complete plan (i.e., a sequence of object-level states beginning with the current state and ending with a terminal state), $\mathcal{P}$ is the set of all such plans, and $V(s,p)$ is the expected value of executing a plan given the current belief state, that is

$$V(s,p) = \sum_{i \in p} \begin{cases} \mathrm{E}[R_i] & \text{if } s_i = \oslash \\ s_i & \text{otherwise.} \end{cases} \tag{3}$$

Model specifications

Each of our candidate models corresponds to a parameterized family of metalevel policies. A policy is defined by a state-conditional distribution over actions, $\pi(a \mid s)$. For all models, this distribution is specified as a four-step generative process. First, if the frontier is empty (i.e., all nodes have been clicked or pruned), the model executes the termination action, $\perp$. Second, if the frontier includes at least one node, then a random legal action is executed with some probability, ε. Otherwise (Step 3), the model executes $\perp$ with probability $p^M_{\text{stop}}(s)$; the form of this function depends on the model, M. Finally (Step 4), if the model did not act randomly or terminate, then it selects a node to expand, each node having probability $p^M_{\text{select}}(s,a)$ The models are thus defined by stochastic stopping and selection rules.

The heuristic models (best-first, depth-first, and breadth-first) all use a common stopping rule that incorporates both the relative and absolute value of the best path identified so far. The stopping probability is a logistic function of a weighted linear combination of these terms, that is,

$$p^H_{\text{stop}}(s) = \frac{1}{1 + \exp\left\{-f_{\text{stop}}(s)\right\}}, \tag{4}$$

where

$$f_{\text{stop}}(s) = \beta_{\text{satisfice}} \cdot V_{\text{best}} + \beta_{\text{bestnext}} \cdot (V_{\text{best}} - V_{\text{next}}) + \theta_{\text{stop}}. \tag{5}$$

V_{best} and V_{next} are the expected values of the best and second-best paths given the current belief state, θ_{stop} sets the midpoint of the logistic function, and the β's control the contribution of each term to its slope. This implementation allows the model to both flexibly interpolate between a relative and an absolute stopping rule and also to vary the precision in the application of the rule. For example, a classic "hard" satisficing rule can be created by setting $\beta_{\text{satisfice}}$ to a very large number, β_{bestnext} to zero, and θ_{stop} to $-\theta \cdot \beta_{\text{satisfice}}$ where θ is the aspiration level. This results in

$$p^{\text{SATISFICE}}_{\text{stop}}(s) = \frac{1}{1 + \exp\{-\beta_{\text{satisfice}}(V_{\text{best}} - \theta)\}}, \tag{6}$$

that is, a logistic function of the expected value of the best path with slope $\beta_{\text{satisfice}}$ and intercept θ.

We defined the selection rule for each heuristic model so that its policy approximates the corresponding classical search algorithm. To do this, we defined

$$p^H_{\text{select}}(s,a) = \frac{\mathbf{1}(a \in \text{frontier}(s)) \cdot \exp\left\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s,a)\right\}}{\sum_{a' \in \text{frontier}(s)} \exp\left\{\beta_{\text{select}} \cdot f^{\text{ALG}}_{\text{select}}(s,a')\right\}}, \tag{7}$$

where $f^{\text{ALG}}_{\text{select}}(s,a)$ denotes a node-scoring function for each algorithm; specifically,

$$\begin{aligned} f^{\text{BEST}}_{\text{select}}(s,a_i) &= V(s,i) = \max_{p \in \{\mathcal{P} \mid i \in p\}} V(s,p) \\ f^{\text{DEPTH}}_{\text{select}}(s,a_i) &= \text{depth}(s,i) \\ f^{\text{BREADTH}}_{\text{select}}(s,a_i) &= -\text{depth}(s,i). \end{aligned} \tag{8}$$

We chose these node scoring functions to ensure that in the limit $\beta_{\text{select}} \to \infty$, the model's selection rule is deterministic and exactly matches the corresponding algorithm. Pure best-first search always expands a node with maximal expected value, pure depth-first search always expands the deepest node in the tree, and pure breadth-first search always expands every node at each depth before expanding any at the next depth. However, To account for variability in human selection decisions, we allow for $\beta_{\text{select}} \in [0, \infty)$.

The random model takes the same form as the heuristic models, with $f^{\text{RAND}}_{\text{select}}(s,a) = 0$ and $f_{\text{stop}}(s) = \theta_{\text{stop}}$. This is equivalent to a fixed stopping probability and random selection. In the random model the probability of choosing computations at random is set to zero ($\varepsilon = 0$) because this step is redundant.

For the optimal model, we define both the stopping and the selection rules using the optimal state-action value function, Q_λ, of the metalevel MDP with computational cost λ. We computed the Q function using dynamic programming. The stopping rule is

$$p^{O}_{\text{stop}}(s) = \frac{\exp\left\{\beta_{\text{stop}} \cdot Q_\lambda(s,\perp)\right\}}{\sum_{a' \in \text{frontier}(s) \cup \{\perp\}} \exp\left\{\beta_{\text{stop}} \cdot Q_\lambda(s,a')\right\}} \tag{9}$$

and the selection rule is

$$p^{O}_{\text{select}}(s,a) = \frac{\exp\left\{\beta_{\text{select}} \cdot Q_\lambda(s,a)\right\}}{\sum_{a' \in \text{frontier}(s)} \exp\left\{\beta_{\text{select}} \cdot Q_\lambda(s,a')\right\}}. \tag{10}$$

Note that if $\beta_{\text{select}} = \beta_{\text{stop}}$, this corresponds to a single softmax over the full action space. However, we use separate inverse temperature parameters for stopping and selection to match the flexibility of the error model used by the optimal model to that of the heuristic models.

The myopic model has the same form, but the Q_λ function is replaced by a myopic one-step approximation [43], which we denote $Q^{\text{myopic}}_\lambda$. For the termination action, this approximation is exact because the trial ends after this action is executed and thus $Q_\lambda(s,\perp) = Q^{\text{myopic}}_\lambda(s,\perp) = r(s,\perp)$. For expansion, the myopic model approximates the Q value as the expected value of stopping at the next time step (after expanding a node) minus the expansion cost, that is

$$Q^{\text{myopic}}_\lambda(s,a) = \mathrm{E}_{s' \sim T(\cdot|s,a)}\left[r(s',\perp)\right] - \lambda. \tag{11}$$

Pruning and depth limits

To model pruning [1, 2] and depth limits [5, 7], we assume that each time a participant expands a node, she may choose to eliminate the corresponding branch from further consideration. Because both mechanisms ultimately involve removing a branch of the decision tree, we refer to them as value-based and depth-based pruning, respectively. If a path is pruned, all unexpanded nodes on that path are removed from the frontier, preventing the model from selecting these nodes. Note that pruning also acts as a secondary stopping rule because all models stop whenever the frontier is empty.

We assume that the value-based and depth-based pruning mechanisms operate independently. For each one, the probability of pruning a just-expanded node is defined as a logistic function of the expected value or tree depth of the node. Value-based pruning is defined

$$p^{\text{VALUE}}_{\text{prune}}(s,i) = \frac{1}{1 + \exp\left\{-\beta^{\text{VALUE}}_{\text{prune}} \cdot \left(\theta^{\text{VALUE}}_{\text{prune}} - V(s,i)\right)\right\}} \tag{12}$$

where $V(s,i)$ is the value of the best path that includes node i, defined in Equation 8. Thus, a path is increasingly likely to be pruned the further below $\theta^{\text{VALUE}}_{\text{prune}}$ its expected value is. Depth-based pruning is defined

$$p^{\text{DEPTH}}_{\text{prune}}(s,i) = \frac{1}{1 + \exp\left\{-\beta^{\text{DEPTH}}_{\text{prune}} \cdot \left(\text{depth}(s,i) - \theta^{\text{DEPTH}}_{\text{prune}}\right)\right\}}. \tag{13}$$

Thus, a path is increasingly likely to be pruned the further past the depth limit it is. Finally, the complete heuristic model contains both forms of pruning operating independently, resulting in

$$p_{\text{prune}}(s,i) = 1 - \left(1 - p^{\text{VALUE}}_{\text{prune}}(s,i)\right) \cdot \left(1 - p^{\text{DEPTH}}_{\text{prune}}(s,i)\right). \tag{14}$$

Unfortunately, implementing this model exactly requires creating (and marginalizing over) a new latent state variable that specifies which nodes have been pruned. To avoid the formidable computational challenges associated with fitting such a model, we follow Huys et al. [1, 2] and use a mean-field approximation. Specifically, we assume that the stochastic decision of whether to prune each branch is resampled at every time step based on its current expected value, treating the set of pruned nodes at each time step as independent. When computing the stopping and selection probabilities (Equations 4 and 7), we marginalize over all possible frontiers that could result from different combinations of pruning decisions, weighing each by its probability according to Equation 14.

Backward planning and forward-search bias

In Experiments 3 and 4, we modified the metalevel MDP to allow planning algorithms that do not correspond to traditional decision-tree search. The formalism described above is maintained with one exception: frontier(s) in Equations 7, 9, and 10 is replaced with unexpanded$(s) = \{a_i \mid s_i = \oslash\}$. Although the metalevel state and action spaces are formally the same, we now interpret a metalevel state as a partially computed value function and a metalevel action as computing the reward at a future world state and also integrating this information into the value of its ancestor states (we assume an acyclic transition function).

However, because we found that participants still showed a strong tendency for forward search, we augmented the selection rule of all models with a forward-search bias, $\beta_{\text{forward}} \cdot \mathbf{1}(a \in \text{frontier}(s))$. For the heuristic models, this term was added to f_{select}. For the optimal and myopic models, it was added inside the exponentiation in the numerator and denominator of Equation 10.

Model fitting and evaluation

We fit all models by maximum likelihood estimation at the individual level, cross-validated across trials. We used five folds in all experiments except Experiment 4, where we used seven folds because there were only seven trials (excluding the practice trial). For each participant, model, and fold, we optimized the model's free parameters by minimizing the negative log-likelihood on the training set, using the L-BFGS algorithm with 100 random starting points sampled from a plausible range. The lapse rate ε was constrained to be no less than .01 to prevent extremely low test likelihoods (a simple form of regularization). For the optimal model, we optimized the cost parameter on a grid (0 to 4 in steps of .05) because dynamic programming is not easily differentiated. We then computed the log-likelihood of each computational action in the test set (node expansions and terminations). The total log-likelihood of the data under each model is the sum of the log-likelihoods in each test set.

Statistical analyses

Analyses on human data were performed on all test trials for all participants who passed the exclusion criterion. For comparison to the optimal model, we conducted analyses on a simulated dataset using costs fit to participant data, but removing decision noise (setting $\varepsilon = 0$, $\beta_{\text{stop}} = 10^5$, and $\beta_{\text{select}} = 10^5$).

Regression analyses were performed using the "lme4" R package with default settings [81]. We included random intercepts as well as random slopes for each fixed effect. Confidence intervals were produced using the default Wald method. Note that, to allow for direct comparison of the model and participant coefficients, we also use mixed-effects regression for the model; in this case, we used the participant that the model's cost parameter was fit to as the group identifier.

All other analyses were performed over participant means. Thus, we report mean proportions rather than total proportions. Confidence intervals were produced by bootstrapping over participants. Wilcoxon and Spearman tests were performed using the "scipy" Python package with default settings.

Data availability

The anonymized data that support the findings of this study are available on the Open Science Framework (https://osf.io/6venh/).

Code availability

The modeling and analysis code is available on the Open Science Framework (https://osf.io/6venh/).

Acknowledgments

This work was supported by grant number ONR MURI N00014-13-1-0341, grant number AFOSR 9550-18-1-0077, a grant from the Templeton World Charity Foundation, and a grant from Facebook Reality Labs. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Author contributions

FC, FL, BVO, and TLG designed the studies. FC, FL and TLG devised the main model. FC, SG, PD, and BVO devised the alternative models. FC implemented the model, collected data, performed analyses, and drafted the manuscript. TLG and FL supervised all aspects of the project. All authors discussed results and revised the manuscript.

References

[1] Huys, Q. J. M. *et al.* Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3098–103 (2015).

[2] Huys, Q. J. M. *et al.* Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology* **8**, e1002410 (2012).

[3] Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y. & Ma, W. J. A computational model for decision tree search. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2017).

[4] MacGregor, J. N., Ormerod, T. C. & Chronicle, E. P. Information processing and insight: a process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**, 176 (2001).

[5] Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences* **113**, 12868–12873 (2016).

[6] Krusche, M. J. F., Schulz, E., Guez, A. & Speekenbrink, M. Adaptive planning in human search. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2018).

[7] Snider, J., Lee, D., Poizner, H. & Gepshtein, S. Prospective Optimization with Limited Resources. *PLOS Computational Biology* **11**, e1004501 (2015).

[8] Von Neumann, J. & Morgenstern, O. *The theory of games and economic behavior* (Princeton university press, Princeton, NJ, 1944).

[9] Stahl, D. O. & Wilson, P. W. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization* **25**, 309–327 (1994).

[10] Camerer, C. F., Ho, T.-H. & Chong, J.-K. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics* **119**, 861–898 (2004).

[11] Newell, A. & Simon, H. The logic theory machine–a complex information processing system. *IRE Transactions on information theory* **2**, 61–79 (1956).

[12] Griffiths, T. L. *et al.* Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences* **29**, 24–30 (2019).

[13] Newell, A., Shaw, J. C. & Simon, H. A. Report on a general problem solving program. In *IFIP Congress*, vol. 256, 64 (Pittsburgh, PA, 1959).

[14] Newell, A., Simon, H. A. *et al. Human problem solving*, vol. 104 (Prentice-hall Englewood Cliffs, NJ, 1972).

[15] Kool, W., Gershman, S. J. & Cushman, F. A. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science* **28**, 1321–1333 (2017).

[16] Norris, D. & Cutler, A. More why, less how: What we need from models of cognition. *Cognition* **213**, 104688 (2021).

[17] Marr, D. *Vision: A computational investigation into the human representation and processing of visual information* (San Francisco: WH Freeman, 1982).

[18] Anderson, J. R. *The adaptive character of thought* (Psychology Press, Hillsdale, NJ, 1990).

[19] Savage, L. J. *The Foundations of Statistics*. The Foundations of Statistics (John Wiley & Sons, Oxford, England, 1954).

[20] Tenenbaum, J. B. & Griffiths, T. L. Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences* **24**, 629–640 (2001).

[21] Anderson, J. R. The adaptive nature of human categorization. *Psychological Review* **98**, 409–429 (1991).

[22] Ashby, F. G. & Alfonso-Reese, L. A. Categorization as probability density estimation. *Journal of mathematical psychology* **39**, 216–233 (1995).

[23] Knill, D. C. & Richards, W. *Perception as Bayesian Inference* (Cambridge University Press, 1996).

[24] Oaksford, M. & Chater, N. A rational analysis of the selection task as optimal data selection. *Psychological Review* **101**, 608–631 (1994).

[25] Gureckis, T. M. & Markant, D. B. Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science* **7**, 464–481 (2012).

[26] Howes, A., Lewis, R. L. & Vera, A. Rational Adaptation Under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action. *Psychological Review* **116**, 717–751 (2009).

[27] Lewis, R. L., Howes, A. & Singh, S. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science* **6**, 279–311 (2014).

[28] Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).

[29] Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* **7**, 217–229 (2015).

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

[31] Simon, H. A. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* **69**, 99–118 (1955).

[32] Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**, 700–765 (2006).

[33] Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience* **32**, 3612–3628 (2012).

[34] Tajima, S., Drugowitsch, J. & Pouget, A. Optimal policy for value-based decision-making. *Nature Communications* **7**, 12400 (2016).

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

[36] Fudenberg, D., Strack, P. & Strzalecki, T. Speed, accuracy, and the optimal timing of choices. *American Economic Review* **108**, 3651–3684 (2018).

[37] Callaway, F., Rangel, A. & Griffiths, T. L. Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology* **17**, e1008863 (2021).

[38] Jang, A. I., Sharma, R. & Drugowitsch, J. Optimal policy for attention-modulated decisions explains human fixation behavior. *eLife* **10**, e63436 (2021).

[39] Sezener, C. E., Dezfouli, A. & Keramati, M. Optimizing the depth and the direction of prospective planning using information values. *PLOS Computational Biology* **15**, e1006827 (2019).

[40] Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience* **21**, 1609–1617 (2018).

[41] Matheson, J. E. The Economic Value of Analysis and Computation. *IEEE Transactions on Systems Science and Cybernetics* **4**, 325–332 (1968).

[42] Horvitz, E. J. Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, 429–447 (1987).

[43] Russell, S. & Wefald, E. Principles of metareasoning. *Artificial Intelligence* **49**, 361–395 (1991).

[44] Payne, J. W. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance* **16**, 366–387 (1976).

[45] Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8**, 1704–1711 (2005).

[46] Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology* **7**, e1002055 (2011).

[47] Shenhav, A., Botvinick, M. & Cohen, J. The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013).

[48] Kool, W. & Botvinick, M. Mental labour. *Nature Human Behaviour* **2**, 899–908 (2018).

[49] Hay, N., Russell, S., Tolpin, D. & Shimony, S. Selecting Computations: Theory and Applications. In de Freitas, N. & Murphy, K. (eds.) *Proceedings of the 28th conference on Uncertainty in Artificial Intelligence* (AUAI Press, Corvallis, 2012).

[50] Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Prentice Hall, Upper Saddle River, NJ, 2002).

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

[52] De Groot, A. D. *Thought and choice in chess* (Grouton, The Hague, 1965).

[53] Chase, W. G. & Simon, H. A. Perception in chess. *Cognitive Psychology* **4**, 55–81 (1973).

[54] Payne, J. W., Bettman, J. R. & Johnson, E. J. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**, 534–552 (1988).

[55] Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M. & Doherty, M. L. Process tracing methods: Contributions, problems, and neglected research questions. *Organizational behavior and human decision processes* **43**, 75–117 (1989).

[56] Payne, J. W., Bettman, J. R. & Johnson, E. J. *The adaptive decision maker* (Cambridge University Press, Cambridge, UK, 1993).

[57] Gabaix, X., Laibson, D., Moloche, G. & Weinberg, S. Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *American Economic Review* **96 (4)**, 1043–1068 (2006).

[58] Schulte-Mecklenbeck, M., Kuehberger, A. & Johnson, J. G. Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. In Schulte-Mecklenbeck, M., Kühberger, A. & Johnson, J. G. (eds.) *A handbook of process tracing methods for decision research*, 37–58 (Psychology Press, 2011).

[59] Ratcliff, R. & Smith, P. L. A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review* **111**, 333–367 (2004).

[60] Teodorescu, A. R. & Usher, M. Disentangling decision models: From independence to competition. *Psychological Review* **120**, 1–38 (2013).

[61] McMillen, T. & Holmes, P. The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology* **50**, 30–57 (2006).

[62] Piantadosi, S. T. One parameter is always enough. *AIP Advances* **8**, 095118 (2018).

[63] Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, 216–224 (1990).

[64] Gigerenzer, G. Why heuristics work. *Perspectives on psychological science* **3**, 20–29 (2008).

[65] Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annual review of psychology* **62**, 451–482 (2011).

[66] Todd, P. M. & Gigerenzer, G. Bounding rationality to the world. *Journal of economic psychology* **24**, 143–165 (2003).

[67] Gigerenzer, G. & Goldstein, D. G. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* **103**, 650 (1996).

[68] Gigerenzer, G. & Todd, P. M. *Simple Heuristics That Make Us Smart* (Oxford University Press, USA, 1999).

[69] O'Donoghue, T. & Rabin, M. Doing It Now or Later. *American Economic Review* **89**, 103–124 (1999).

[70] Kahneman, D. & Klein, G. Conditions for intuitive expertise: a failure to disagree. *American psychologist* **64**, 515 (2009).

[71] Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences* **110**, 18327–18332 (2013).

[72] Jara-Ettinger, J., Gweon, H., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* **20**, 589–604 (2016).

[73] Lohse, G. L. & Johnson, E. J. A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes* **68**, 28–43 (1996).

[74] Hunt, L. T. *et al.* Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience* **24**, 1051–1064 (2021).

[75] Ongchoco, J. D., Jara-Ettinger, J. & Knobe, J. Imagining the good: An offline tendency to simulate good options even when no decision has to be made. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 904–910 (2019).

[76] Ho, M. K., Abel, D., Cohen, J., Littman, M. & Griffiths, T. The efficiency of human cognition reflects planned information processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 1300–1307 (2020).

[77] Solway, A. *et al.* Optimal Behavioral Hierarchy. *PLOS Computational Biology* **10**, e1003779 (2014).

[78] Lieder, F. & Griffiths, T. L. Strategy selection as rational metareasoning. *Psychological Review* **124**, 762–794 (2017).

[79] Krueger, P. M., Lieder, F. & Griffiths, T. L. Enhancing metacognitive reinforcement learning using reward structures and feedback. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Cognitive Science Society, 2017).

[80] Rahnev, D. & Denison, R. N. Suboptimality in perceptual decision making. *Behavioral and Brain Sciences* **41**, e223 (2018).

[81] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).