An automatic method for discovering rational heuristics for risky choice

Falk Lieder[1] (falk.lieder@berkeley.edu)
Paul M. Krueger[1] (pmk@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)

Department of Psychology,University of California Berkeley, Berkeley, CA 94720 USA

[1] These authors contributed equally.

Abstract

What is the optimal way to make a decision given that your time is limited and your cognitive resources are bounded? To answer this question, we formalized the bounded optimal decision process as the solution to a meta-level Markov decision process whose actions are costly computations. We approximated the optimal solution and evaluated its predictions against human choice behavior in the Mouselab paradigm, which is widely used to study decision strategies. Our computational method rediscovered well-known heuristic strategies and the conditions under which they are used, as well as novel heuristics. A Mouselab experiment confirmed our model's main predictions. These findings are a proof-of-concept that optimal cognitive strategies can be automatically derived as the rational use of finite time and bounded cognitive resources.

Keywords: Decision-Making; Heuristics; Bounded Rationality; Strategy Selection; Rational Metareasoning

Introduction

Some situations require us to decide quickly whereas others call for careful consideration of all available options and potential consequences. People seem to master this challenge by choosing adaptively from a toolbox of diverse decision strategies (Payne, Bettman, & Johnson, 1988; Gigerenzer & Selten, 2002). This toolbox is assumed to include fast-and-frugal heuristics (Gigerenzer & Goldstein, 1996) as well as slower and more effortful strategies. Fast-and-frugal heuristics include Take-The-Best (TTB), which chooses the alternative that is favored by the most predictive attribute and ignores all other attributes, satisficing (SAT) (Simon, 1956), which chooses the first alternative whose expected value exceeds some threshold, and random choice; slower strategies include the Weighted-Additive Strategy (WADD), which computes all gambles' expected values based on all possible payoffs. Except for WADD, all of these strategies are heuristics: they solve some problems very efficiently but err on others.

The systematic errors that result from people's use of heuristics are inconsistent with classic notions of rationality such as logic, probability theory, and expected utility theory (Tversky & Kahneman, 1974). Making good decisions is remarkably constrained: decisions have to be made in a finite amount of time, people's cognitive resources are limited, and maximizing expected utility entails intractable computational problems. This makes expected utility theory an unrealistically high bar for human rationality. According to a more realistic normative standard, people should decide in a way that makes the best possible use of their limited cognitive resources (Griffiths, Lieder, & Goodman, 2015). Previous research has applied this *resource-rational* approach to numerical estimation (Lieder, Griffiths, & Goodman, 2012), availability biases (Lieder, Hsu, & Griffiths, 2014), and strategy selection (Lieder, Plunkett, et al., 2014). However, this approach has not been applied to the domain in which heuristics have perhaps been studied in greatest detail: multi-alternative risky choice. Work on risky choice suggests that people adaptively switch between multiple different strategies depending on how much time is available and whether one of the outcomes is much more likely than the others (Payne et al., 1988). Yet, it remains unclear how people's decision processes compare to resource-rational behavior.

To answer these questions, we model the decision process as a sequence of costly computations and formalize the optimal decision process as the solution to a meta-level Markov decision process. We combine this theory with an algorithm for approximating the optimal solution to create a computational method that can automatically derive optimal cognitive strategies. These rational heuristics can be interpreted as a fair normative standard for human decision making that takes into account that people's time is costly and that their cognitive resources are bounded. We are optimistic that this novel approach will lead to new insights about how decision-makers cope with limited time and bounded computational resources, and advance the debate about human rationality.

We illustrate our approach in multi-alternative risky choice and test its predictions using the Mouselab paradigm that is widely used to study decision strategies (Johnson, Payne, Bettman, & Schkade, 1989). Two known heuristics, TTB and random choice, emerged from our theory as resource-rational strategies for low-stakes decisions with high and low dispersion of their outcome probabilities, respectively. In addition, our computational method discovered a novel heuristic that combines TTB with satisficing. Our experiment demonstrated that people do indeed use the newly discovered heuristic and confirmed our rational model's predictions of when people use which strategy: people used simple heuristics more frequently when the stakes were low, employed fast-and-frugal heuristics less frequently when all outcomes were almost equally likely (low dispersion), and invested more time and effort when the stakes were high. This is the first demonstration that rational meta-reasoning can be used to automatically discover decision strategies used by people.

Background

We will formulate our theory using the mathematical frameworks of Markov decision processes, bounded optimality, and

rational metareasoning, introduced in this section.

Markov Decision Processes

Each sequential decision problem can be modeled as a *Markov Decision Process* (MDP)

$$M = (\mathcal{S}, \mathcal{A}, T, \gamma, r, P_0), \tag{1}$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T(s,a,s')$ is the probability that the agent will transition from state s to state s' if it takes action a, $0 \leq \gamma \leq 1$ is the discount factor on future rewards, $r(s,a,s')$ is the reward generated by this transition, and P_0 is the probability distribution of the initial state S_0 (Sutton & Barto, 1998). A *policy* $\pi : \mathcal{S} \mapsto \mathcal{A}$ specifies which action to take in each of the states. The expected sum of discounted rewards that a policy π will generate in the MDP M starting from a state s is known as its *value function*

$$V_M^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right]. \tag{2}$$

The optimal policy $\pi_M^{\star}$ maximizes the expected sum of discounted rewards, that is

$$\pi_M^{\star} = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right]. \tag{3}$$

Bounded optimality and rational metareasoning

People and robots have to make decisions in a limited amount of time and with bounded cognitive resources. Given that these resources are scarce, which strategy should a decision-maker employ to use its resources most effectively? The theory of bounded optimality and rational metareasoning (Russell & Wefald, 1991; Russell & Subramanian, 1995) was developed to answer this question for rational agents with limited performance hardware. It frames this problem as selecting computations so as to maximize the sum of the rewards of resulting decisions minus the costs of the computations involved.

Concretely, the problem of choosing computations optimally can be formulated as a meta-level MDP (Hay, Russell, Tolpin, & Shimony, 2012). A meta-level MDP

$$M_{\text{meta}} = (\mathcal{B}, \mathcal{C}, T_{\text{meta}}, r_{\text{meta}}) \tag{4}$$

is a Markov decision process whose actions $\mathcal{C}$ are cognitive operations, its states $\mathcal{B}$ represent the agent's probabilistic beliefs, and the transition function T_{meta} models how cognitive operations change the agent's beliefs. In addition to a set of computations C that update the agent's belief, the cognitive operations also include the meta-level action $\perp$ that terminates deliberation and translates the current belief into action. The meta-level state b_t encodes the agent's probabilistic beliefs about the domain it is reasoning about.The meta-level reward function r_{meta} captures the cost of thinking (Shugan, 1980) and the reward r the agent expects to receive from the environment when it stops deliberating and takes action. The computations C do not yield any external reward. Their only effect is to update the agent's beliefs. Hence, the meta-level reward for performing a computation $c \in C$ is $r_{\text{meta}}(b_t, c) = -\text{cost}(c)$. By contrast, terminating deliberation and taking action ($\perp$) does not update the agent's belief. Instead, its value lies in the anticipated reward for taking action, that is

$$r_{\text{meta}}(b_t, \perp) = \arg\max_a b_t^{(\mu)}(a), \tag{5}$$

where $b_t^{(\mu)}(a)$ is the expected reward of taking action a according to the belief b_t.

Adaptive strategy selection in risky choice

Consistent with rational metareasoning, people flexibly adapt their decision processes to the structure of the problem they face. Concretely, Payne et al. (1988) found that people use fast-and-frugal heuristics, like TTB, more frequently when they are under time pressure and when one outcome is much more likely than the others. In this research, participants were given the choice between gambles $g_1, \cdots, g_n$. Each gamble was defined by the payoffs it assigns to each of four possible outcome whose probabilities are known ($P(O)$). Participants could inspect a payoff matrix $V_{o,g}$ with one row for each outcome o and one column for each gamble g. Critically, each payoff is only revealed when the participant clicks on the corresponding cell of the payoff matrix using a mouse; this task is hence referred to as the *Mouselab* paradigm (see Figure 1).

The adaptiveness of people's strategy choices in the Mouselab paradigm suggests that their decision processes are efficient and effective. But it is difficult to test whether they are optimal, because it is unclear what it means to decide optimally when one's time is valuable and one's cognitive resources are limited. To clarify this, the following section develops a normative theory of resource-bounded decision making in the Mouselab paradigm.

Boundedly-optimal decision-making

To model the meta-decision problem posed by the Mouselab task, we characterize the decision-maker's belief state b_t by probability distributions on the expected values $e_1 = \mathbb{E}[v_{O,g_1}], \cdots e_n = \mathbb{E}[v_{O,g_n}]$ of the n available gambles $g_1, \cdots, g_n$. Furthermore, we assume that for each element $v_{o,g}$ of the payoff matrix V there is one computation $c_{o,g}$ that inspects the payoff $v_{o,g}$ and updates the agent's belief about the expected value of the inspected gamble according to Bayesian inference. Since the entries of the payoff matrix are drawn from the normal distribution $\mathcal{N}(\bar{v}, \sigma_v^2)$, the resulting posterior distributions are also Gaussian. Hence, the decision-maker's belief state b_t can be represented by $b_t = (b_{t,1}, \cdots, b_{t,n})$ with

$$b_{t,g} = \left(b_{t,g}^{(\mu)}, b_{t,g}^{(\sigma^2)}\right), \tag{6}$$

where $b_{t,g}^{(\mu)}$ and $b_{t,g}^{(\sigma^2)}$ are the mean and the variance of the probability distribution on the expected value of gamble g of the belief state b_t.

Given the set O_t of the indices $(k_o^{(1)}, k_g^{(1)}), \cdots, (k_o^{(t)}, k_g^{(t)})$ of the t observations made so far, the means and variances characterizing the decision-maker's beliefs are given by

$$b_{t,g}^{(\mu)} = \sum_{(o,g)\in O} p(o) \cdot v_{o,g} + \sum_{(o,g)\notin O} p(o) \cdot \bar{v} \tag{7}$$

$$b_{t,g}^{(\sigma^2)} = \sum_{(o,g)\notin O} p(o)^2 \cdot \sigma_v^2. \tag{8}$$

The meta-level transition function $T(b_t, c_{o,g}, b_{t+1})$ encodes the probability distribution on what the updated means and variances will be given the observation of a payoff value $V_{o,g}$ sampled from $\mathcal{N}(\bar{v}, \sigma_v^2)$. The meta-level reward for performing the computation $c_{o,g} \in C$ encodes that acquiring and processing an additional piece of information is costly. We assume that the cost of all such computations is an unknown constant λ. The meta-level reward for terminating deliberation and taking action is $r_{\text{meta}}(b_t, \perp) = \max_g b_t^{(\mu)}(g)$.

Approximating the optimal meta-level policy: Bayesian value function approximation

Unfortunately, computing the optimal policy for the meta-level MDP defined above is intractable. However, it can be approximated using methods from reinforcement learning. We initially used the semi-gradient SARSA algorithm (Sutton & Barto, 1998) with limited success. We therefore developed a new algorithm that replaces the gradient descent component of that algorithm by Bayesian linear regression.

Our algorithm learns a linear approximation to the meta-level Q-function

$$Q_{\text{meta}}(b,c) \approx \sum_k w_k \cdot f_k(b,c), \tag{9}$$

whose features $\mathbf{f}$ include a constant, features of the belief state b_t, and features of the computation c_t. The features of the belief state were the expected value of the maximum of the gambles' expected values ($\mathbb{E}[\max_g E_g | b_t]$) and the decision-maker's uncertainty about it ($\sqrt{\text{Var}[\max_g E_g | b_t]}$). The largest posterior mean ($\max_g b_{t,g}^{(\mu)}$) and its associated uncertainty ($\sqrt{\mu_{t,g^\star}^{(\sigma^2)}}$ where $g^\star = \arg\max_g b_{t,g}^{(\mu)}$), the second largest posterior mean and the decision-maker's uncertainty about it, and the expected regret $\mathbb{E}[\text{regret}(g)|b_t]$ that the decision-maker would experience if they chose based on their current belief (where $\text{regret}(g) = \max_g E_g - \max_g b_{t,g}^{(\mu)}$ for $E_i \sim \mathcal{N}(b_{t,i}^{(\mu)}, b_{t,i}^{(\sigma)})$ for all gambles i). The features of the computation $c_{o,g}$ were its myopic value of computation ($\text{VOC}(b_t, c_{o,g})$; see Russell & Wefald, 1991), the current uncertainty about the expected value of the inspected gamble ($b_{t,g}^{(\sigma)}$), the probability of the inspected outcome, the difference between the largest posterior mean and the posterior mean of the inspected outcome, a binary variable indicating whether the computation acquired new information, and the expected reduction in the expected regret $\text{ER}(b)$ minus its cost (i.e. $\mathbb{E}[\text{ER}(B_{t+1})|b_t, c] - \text{ER}(b_t) - \lambda$, where B_{t+1} is the unknown belief state resulting from performing computation c in belief state b_t and $\text{ER}(b_t) = \mathbb{E}[\text{regret}(\arg\max_g b_{t,g}^{(\mu)})|b_t]$).

The weights $\mathbf{w}$ are learned by Bayesian linear regression of the bootstrap estimate $\hat{Q}(b,c)$ of the meta-level value function onto the features $\mathbf{f}$. The bootstrap estimator is

$$\hat{Q}(b_t, c_t) = r_{\text{meta}}(b_t, c_t) + \hat{w}_t' \cdot \mathbf{f}(b_{t+1}, c_{t+1}), \tag{10}$$

where $\hat{w}_t$ is the posterior mean on the weights w given the observations from the first t trials, and $\mathbf{f}(b_{t+1}, c_{t+1})$ is the feature vector characterizing the subsequent belief state b_{t+1} and the computation c_{t+1} that will be selected in it.

Given the learned posterior distribution on the feature weights $\mathbf{w}$, the next computation c is selected by contextual Thompson sampling (Agrawal & Goyal, 2013). Specifically, to make the t^{th} meta-decision, a weight vector $\tilde{w}$ is sampled from the posterior distribution of the weights given the series of meta-level states, selected computations, and the resulting value estimates experienced so far, that is

$$\tilde{w} \sim P(\mathbf{w} | (b_1, c_1, \hat{Q}(b_1, c_1)), \cdots, (b_{k-1}, c_{k-1}, \hat{Q}(b_{k-1}, c_{k-1}))).$$

The sampled weight vector $\tilde{w}$ is then used to predict the Q-values of each available computation $c \in \mathcal{C}$ according to Equation 9. Finally, the computation with the highest predicted Q-value is selected.

Application to Mouselab experiment

As a proof of concept, we applied our approach to the Mouselab experiment described below. The experiment comprises 50% high-stakes problems and 50% low-stakes problems. Since participants are informed about the stakes, we learned two separate policies for high-stakes and low-stakes problems, respectively. Half of each of those problems had nearly uniform outcome probabilities ("low dispersion") and for the other half one outcome was much more likely than all others combined ("high dispersion"). The parameters of the simulated environment were exactly equal to those of the experiment described below. Our model assumed that people play each game as if they receive the payoff of the selected gamble. We estimated the cost per click to be about $\lambda = 3$ cents. This value was selected to roughly match the average number of acquisitions observed in the experiment.

To approximate the optimal meta-decision policy for this task, we ran our feature-based value function approximation method for 4000 low-stakes training trials and 4000 high-stakes training trials, respectively.

Model predictions

The meta-level MDP described above formalizes the costs and benefits of acquiring and processing additional pieces of information: acquiring additional information can improve the decision that will be taken later on but also incurs an immediate cost. Hence, the optimal solution approximated by our computational method executes a cognitive operation or sequence of operations if and only if the resulting improvement in decision quality is larger than cost of those

Figure 1: The Mouselab paradigm, showing an example sequence of clicks generated by the SAT-TTB strategy, which was discovered through approximate rational metareasoning.

operations. Intuitively, this means that the decision process prescribed by our model achieves the optimal tradeoff between decision quality versus decision time and mental effort. This tradeoff depends on the stakes of the decision such that higher stakes usually warrant more deliberation. Likewise, since processing probable outcomes is more likely to improve the quality of the resulting decision than processing improbable outcomes, we expect our model to prioritize probable outcomes over less probable outcomes—especially in high-dispersion trials.

Our computational method automatically discovered strategies that people are known to use in the Mouselab paradigm as well as a novel strategy that has not been reported yet. Our method rediscovered TTB, WADD, and the random choice strategy. In addition, it discovered a new hybrid strategy that combines TTB with satisficing (SAT-TTB). Like TTB, SAT-TTB inspects only the payoffs for the most probable outcome. But unlike TTB and like SAT, SAT-TTB terminates as soon as it finds a gamble whose payoff for the most probable outcome is high enough. On average, this value was about \$0.15 when the payoffs ranged from \$0.01 to \$0.25 (i.e., low-stakes trials). Figure 1 illustrates this strategy.

Furthermore, our model makes intuitive predictions about the contingency of people's choice processes on stakes and outcome probabilities. First, our model predicts that people should use fast-and-frugal heuristics more frequently in high-dispersion trials. This is intuitively rational because high dispersion means that one outcome is much more likely than all others and fast-and-frugal heuristics ignore all outcomes except for the most probable one(s). Concretely, our model generated TTB as the strategy of choice for 100% of the high-dispersion problems with low-stakes, but for low-dispersion problems with low-stakes the model considered the random choice strategy to be optimal in the majority (56%) of cases; it used the SAT-TTB hybrid strategy for 24% of such trials, and it indicated the TTB strategy only for the remaining 20%.

Second, our model predicts that people should use simple heuristics, like TTB, SAT-TTB, and random choice, primarily when the stakes are low. This, too, is intuitively rational because fast and frugal heuristics tend to be faster but less accurate than more effortful strategies. Our model used these heuristics for 100% of the low-stakes problems. But for high-stakes problems, the model never used any of these or other frugal strategies. Instead, the model typically inspected the vast majority of all cells ($24.8/28$ for low-dispersion problems and $23.7/28$ for high-dispersion problems). The few cells that it did not inspect were mostly the payoffs of less-likely outcomes of the best gamble when its inspected payoffs for the most likely outcome(s) were high enough to guarantee that it would be optimal.

Third, our model predicts that when the stakes are high people should invest more time and effort ($F(1,396) = 9886.8, p < 0.0001$) to reap a higher fraction of the highest possible expected payoff ($F(1,339) = 135.24, p < 0.0001$). This, too, is consistent with the rational speed-accuracy tradeoff inherent in our theory. When the stakes were low the model inspected only 4.3 payoffs on average and reaped only 87% of the possible reward; but when the stakes were high the model inspected 24.3 of the 28 possible payoffs and reaped 99% of the best expected payoff on average. In 97% of these trials, the model achieved this near-maximal performance while being more efficient and more frugal than the WADD strategy which it employed for only 3% of these problems.

Experimental test of novel predictions

To test the predictions of our model, we conducted a new Mouselab experiment that manipulated the stakes and dispersion of outcome probabilities within subjects in an identical manner to the model simulations.

Methods

Participants We recruited 200 participants on Amazon Mechanical Turk. The experiment took about 30min. Participants received a base pay of \$1.50, and one of their twenty winnings was selected at random and awarded as a bonus to motivate them to take each trial seriously (avg. bonus \$3.53).

Procedure Participants performed a variation of the Mouselab task (Payne et al., 1988). Participants played a series of 20 games divided into two blocks. Figure 1 shows a screenshot of one game. Every game began with a 4×7 grid of occluded payoffs: there were seven gambles to choose from (columns) and four possible outcomes (rows). The occluded value in each cell specified how much the gamble indicated by its column would pay if the outcome indicated by its row occurred. The outcome probabilities were described by the number of balls of a given color in a bin of 100 balls, from which the outcome would be drawn. For each trial, participants were free to inspect any number of cells before selecting a gamble, with no time limit. The value of each inspected cell remained visible onscreen for the duration of the trial. Upon selecting a gamble, the resulting reward was displayed.

Experimental design The experiment used a 2×2 within subjects design. Each block of ten trials was either low-stakes

or high-stakes, with block order randomly counterbalanced across participants. In games with low-stakes, the possible outcomes ranged from \$0.01 to \$0.25, while in high-stakes games, outcomes ranged from \$0.01 to \$9.99. The payoffs were drawn from a truncated normal distribution with mean $\frac{r_{\max}+r_{\min}}{2}$ and standard deviation $0.3 \cdot (r_{\max} - r_{\min})$. Within each block, there were five low-dispersion trials and five high-dispersion, ordered randomly. In low-dispersion trials, the probability of each of the four outcomes ranged from 0.1 to 0.4, whereas in high-dispersion trials, the probability of the most likely outcome ranged from 0.85 to 0.97.

Strategy identification We identified six different decision strategies, in humans and in simulations, using the following definitions: TTB was defined as inspecting all cells in the row corresponding to the most probable outcome and nothing else. SAT occurs when one gamble's payoffs are inspected for all four outcomes, potentially followed by the inspection of all outcomes of another gamble, and so on, but leaving at least one gamble unexamined. The hybrid strategy, SAT-TTB, was defined as inspecting the payoffs of 1 to 6 gambles for the most probable outcome and not inspecting payoffs for any other outcome. TTB2 was defined as inspecting all fourteen cells of the two most probable outcomes, and nothing else. WADD was defined as inspecting all 28 cells column by column.Random decisions mean zero samples were taken.

Results

Our process tracing data confirmed that people do indeed use the SAT-TTB strategy discovered by our model. Table 1 shows the frequency of various decision strategies, for each of the four different types of trials. Out of 4000 trials across all participants, TTB was the most common strategy overall, accounting for 25.3% of all trials. SAT-TTB was the second most common strategy among those we examined: participants employed this strategy on 10.7% of all trials. In 8.0% of trials participants chose randomly without making any observations—mostly during low-stakes games. Interestingly, we also observed a second novel strategy that we call Take-The-Best-Two (TTB2). This strategy inspects all gambles' payoffs for the *two* most probable outcomes, and was used in 6.3% of trials. The WADD strategy occurred in 4.5% of trials. Finally, the SAT strategy was used in 3.1% of games.

Consistent with our model's first prediction, people used TTB more frequently when the dispersion was high ($\chi^2(1) = 897.9, p < 0.0001$). Consistent with our model's second prediction, participants used simple heuristics more frequently when the stakes were low: the frequency of the random choice—the simplest heuristic—increased significantly from 4.2% on high-stakes problems to 19.9% on low-stakes problems ($\chi^2(1) = 88.2, p < 0.0001$), and so did the frequency of the second simplest heuristic, SAT-TTB ($\chi^2(1) = 86.3, p < 0.0001$), and the third simplest heuristic, TTB ($\chi^2(1) = 20.0, p < 0.0001$). The frequency of SAT also increased from high- to low-stakes games ($\chi^2(1) = 3.4, p < 0.05$, one-tailed). Finally, consistent with our model's third prediction, the frequency of the most effortful and most accurate strategy, WADD, increased with the stakes ($\chi^2(1) = 19.3, p < 0.0001$).

	Frequency				
Strategy	Total	HS-HD	HS-LD	LS-HD	LS-LD
TTB	1012	392	64	449	107
SAT-TTB	412	68	54	140	150
Random	320	41	42	111	126
TTB2	251	34	94	25	98
WADD	178	33	84	19	42
SAT	89	14	22	23	30

HS-HD = High-stakes, high-dispersion HS-LD = High-stakes, low-dispersion
LS-HD = Low-stakes, high-dispersion LS-LD = Low-stakes, low-dispersion

Table 1: Frequency of strategy types for each type of trial.

Together, the strategies reported in Table 1 account for only about half (48.6%) of all trials.To test our model's predictions on all of the trials, we quantified people's decision style by four metrics introduced by Payne et al. (1988): the number of inspected cells (*acquisitions*), the proportion of those inspections that pertained to the most probable outcome (*prioritization*), the degree to which subsequent acquisitions inspected the payoffs of different gambles for the same outcome versus the payoffs of the same gamble for different outcomes (*outcome-based processing*: $\frac{n_{\text{same outcome}} - n_{\text{same gamble}}}{n_{\text{same outcome}} + n_{\text{same gamble}}}$), and the average ratio of the expected value of the chosen gamble over the expected value of the optimal choice (*relative performance*). To further test our model's predictions, we ran a 2-way mixed-effects ANOVA for each of these four metrics.

As shown in Figure 2, the effects of the stakes and outcome probabilities on the four metrics confirmed the model's predictions. Our model's first prediction that high dispersion promotes the use of fast-and-frugal heuristics was confirmed by a decrease in the number of acquisitions ($F(1,3798) = 78.24, p < 0.0001$) in conjunction with an increases in outcome-based processing ($F(1,3432) = 68.31, p < 0.0001$) and prioritization (($F(1,3478) = 280.1, p < 0.0001$)). The increase in prioritization was especially striking: while only 40.4% of participants' clicks inspected the most probable outcome when dispersion was low, they focused 70.6% of their acquisitions on the most probable outcome when dispersion was high. Our model's second prediction that the higher stakes should decrease people's reliance on fast-and-frugal heuristics was confirmed by a significant increases in the number of acquisitions ($F(1,3798) = 281.47, p < 0.0001$) which was accompanied by a decrease in prioritization ($F(1,3478) = 62.42, p < 0.0001$) and an increase in relative performance ($F(1,3798) = 47.62, p < 0.0001$). Consistent with the model's third prediction, the average outcome-based processing metric was lower for high stakes but this effect was not statistically significant ($F(1,3432) = 2.45, p = 0.06$, one-tailed). Our model's third prediction that high-stakes increases time, effort, and performance, was confirmed by a significant increases in the number of acquisitions ($F(1,3798) = 281.47, p < 0.0001$) and relative performance ($F(1,3798) = 47.62, p < 0.0001$) with high stakes.

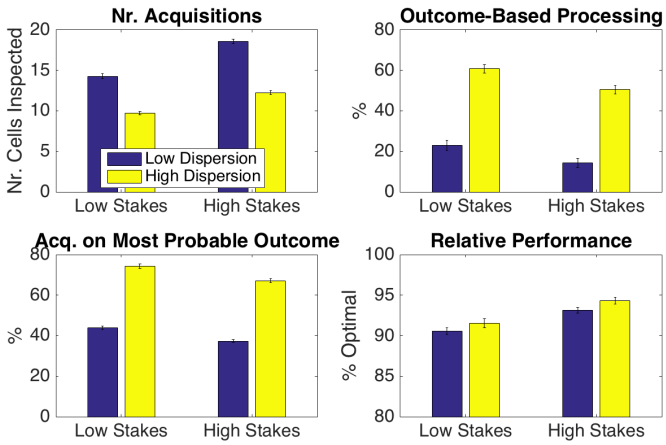

Figure 2: People's decision style by stakes and dispersion of the outcome probabilities.

Despite these qualitative agreements, there were quantitative differences. Most notably, our model predicted a more pronounced effect of the stakes on the number of acquisitions than we observed in people (+19.6 vs. +3.4); the smaller effect in people might reflect their concave utility function.

Discussion

In summary, our resource-rational theory of multi-alternative risky choice predicted some of the main strategies people use in the Mouselab paradigm and the conditions under which they are selected. In addition to automatically discovering known strategies and contingencies, our computational approach also discovered a novel, previously unknown heuristic that integrates TTB with satisficing (SAT-TTB), and our experiment confirmed that people do indeed use SAT-TTB on a non-negligible fraction of problems—especially when the stakes are low.

Tajima, Drugowitsch, and Pouget (2016) solved meta-level MDPs to derive boundedly optimal drift-diffusion models. The strategy discovery method presented here generalizes this approach to more complex decision mechanisms that can process and generate evidence in many different ways.

One limitation of the current work is that we do not know how closely our algorithm approximated the optimal policy, and it is possible that a more accurate approximation would yield somewhat different predictions. Future work will systematically evaluate the accuracy of our approximation method on smaller problems for which the optimal meta-level policy can be computed exactly. Another limitation of the present work is that the cost of computation had to be fit to the participants' responses. Future work will control the cost per click and measure it independently. This will enable a direct comparison of the time and effort people invest against the optimal amount of deliberation. However, a thorough answer to this question will require a more detailed model of people's cognitive architecture including a model of working memory. Another direction for future work is to characterize the decision strategies the model employed on the vast majority of high-stakes problems where it did not use WADD.

Our proof-of-concept study suggests that formulating the problem of making optimal use of finite time and limited cognitive resources as a meta-level MDP is a promising approach to discovering cognitive strategies. This approach can be leveraged to develop more realistic normative standards of human rationality. This might enable future work to systematically evaluate the extent to which people are resource-rational. In the long term, our approach could be used to improve human reasoning and decision-making by discovering rational heuristics and teaching them to people.

Acknowledgments. This work was supported by grant number ONR MURI N00014-13-1-0341 and a grant from the Templeton World Charity Foundation.

References

Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th international conference on machine learning* (pp. 127–135).

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.

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

Hay, N., Russell, S., Tolpin, D., & Shimony, S. (2012). Selecting computations: Theory and applications. In N. de Freitas & K. Murphy (Eds.), *Uncertainty in artificial intelligence: Proceedings of the twenty-eighth conference.* P.O. Box 866 Corvallis, Oregon 97339 USA: AUAI Press.

Johnson, E. J., Payne, J. W., Bettman, J. R., & Schkade, D. A. (1989). *Monitoring information processing and decisions: The mouselab system* (Tech. Rep.). DTIC Document.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, 2699–2707.

Lieder, F., Hsu, M., & Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the 36th annual conference of the cognitive science society.*

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2870–2878). Curran Associates, Inc.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

Russell, S. J., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*(1-3), 361–395.

Shugan, S. M. (1980). The cost of thinking. *Journal of consumer Research*, *7*(2), 99–111.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.