Planning Complexity Registers as a Cost in Metacontrol

Wouter Kool, Samuel J. Gershman*, and Fiery A. Cushman*

Abstract

■ Decision-making algorithms face a basic tradeoff between accuracy and effort (i.e., computational demands). It is widely agreed that humans can choose between multiple decision-making processes that embody different solutions to this tradeoff: Some are computationally cheap but inaccurate, whereas others are computationally expensive but accurate. Recent progress in understanding this tradeoff has been catalyzed by formalizing it in terms of model-free (i.e., habitual) versus model-based (i.e., planning) approaches to reinforcement learning. Intuitively, if two tasks offer the same rewards for accuracy but one of them is much more demanding, we might expect people to rely on habit more in the difficult task: Devoting significant computation to achieve slight marginal accuracy gains would not be "worth it." We test and verify this prediction in a sequential reinforcement learning task. Because our paradigm is amenable to formal analysis, it contributes to the development of a computational model of how people balance the costs and benefits of different decision-making processes in a task-specific manner; in other words, how we decide when hard thinking is worth it. ■

INTRODUCTION

It is not always obvious how hard to think. Whether planning a route home, writing a shopping list, or estimating the financial returns of an investment, we face a basic tradeoff: Thinking harder about a task means doing better at it (a benefit), but it takes time and also diverts attention from other tasks (costs). Thus, many psychological theories agree that humans perform some kind of cost–benefit analysis when allocating "mental effort" to a task (Shenhav et al., 2017; Kurzban, Duckworth, Kable, & Myers, 2013).

To refine these theories, researchers have devoted increasing attention to developing experimental paradigms in which (1) mental effort is linked to both costs and benefits, (2) the costs and benefits can be exogenously manipulated, or (3) their tradeoff is amenable to formal analysis, for instance, within the reinforcement learning (RL) framework. Progress has been made on several of these fronts independently (Kool, Gershman, & Cushman, 2017; Kool, Cushman, & Gershman, 2016; Boureau, Sokol-Hessner, & Daw, 2015; Kool, McGuire, Rosen, & Botvinick, 2010). The aim of this study is to accomplish them simultaneously. Specifically, we assess whether people flexibly adjust the degree of advantageous planning effort devoted to an RL task as the complexity of the planning required is manipulated.

An RL Approach

Several theories in psychology and neuroscience (Kahneman, 2011; Dickinson, 1985) have proposed that there exist two systems that we can use to evaluate the available actions: a slow and deliberative goal-directed system that plans actions so as to obtain a desired goal and a fast and automatic system that relies on habit, associating rewards directly to the actions that produced them without considering the structure of the environment.

Contemporary research has formalized the distinction between habit and planning using RL theory (Dolan & Dayan, 2013; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005), a computational approach that describes how agents ought to choose between actions to maximize future cumulative reward. In this dual-system theory, the habitual system corresponds to model-free RL, which reinforces actions that previously led to reward (Thorndike, 1911). This system is computationally cheap but inflexible, because it needs direct experience to incrementally update its value function to accommodate sudden changes. The goal-directed system corresponds to model-based RL and achieves flexibility by planning in an explicit causal model of the environment. This system is comparatively flexible because sudden changes can directly be incorporated into the causal model, but this comes at the cost of increased computational costs.

Following a seminal paper (Daw et al., 2011), a variety of related sequential decision-making tasks have emerged as the standard behavioral paradigm to dissociate model-free and model-based control strategies in humans (for a review, see Kool, Cushman, & Gershman, in press). This paradigm has afforded rapid progress in determining

This paper is part of a Special Focus deriving from a symposium at the 2017 annual meeting of Cognitive Neuroscience Society, entitled, "Multiple neurocomputational, motivational, and mnemonic mechanisms for decision-making."

Harvard University

*These authors contributed equally to the study.

doi:10.1162/jocn_a_01263

the neural correlates of the two systems (Doll, Duncan, Simon, Shohamy, & Daw, 2015; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013; Wunderlich, Smittenaar, & Dolan, 2012; Daw et al., 2011), the cognitive mechanisms that implement them (Gillan, Otto, Phelps, & Daw, 2015; Otto, Skatova, Madlon-Kay, & Daw, 2015; Otto, Raio, Chiang, Phelps, & Daw, 2013), and their clinical implications (Patzelt, Kool, Millner, & Gershman, submitted for publication; Gillan, Kosinski, Whelan, Phelps, & Daw, 2016).

Here, we adapt this family of tasks to investigate whether people become less likely to devote profitable mental effort to model-based control due to increases in the complexity of the planning task (in our task, due to the increased depth of a decision tree).

Allocation of Mental Effort

In recent years, researchers have devoted increasing attention to the question of how, from moment to moment, people decide to allocate mental effort. Several foundational studies established that people assign a subjective cost to allocating cognitive control and that this can be offset by the prospect of reward (Dixon & Christoff, 2012; Kool et al., 2010; Botvinick, Huffstetler, & McGuire, 2009). Most importantly, Westbrook, Kester, and Braver (2013) showed that participants' willingness to perform a cognitive task decreases as its effort demands increase. These studies manipulate the demand of mental effort by imposing working memory or executive function engagement (for reviews, see Kool, Shenhav, & Botvinick, 2017; Shenhav et al., 2017; Botvinick & Braver, 2015) but do not make direct contact with the RL framework.

Meanwhile, several other studies implicate a key role for cognitive control in model-based action selection. For example, model-based control is significantly reduced under cognitive load (Otto, Gershman, Markman, & Daw, 2013), and the degree to which people are prone to use model-based strategies correlates with measures of cognitive control ability such as working memory capacity (Otto, Raio, et al., 2013) and performance on response interference tasks (Otto et al., 2015). These findings suggest that the exertion of model-based control is itself dependent on executive functioning or cognitive control and therefore carries an effort cost.

Recently some effort has been made to integrate these literatures by exploring sensitivity to costs and benefits of cognitive control within RL tasks. Here, the rationale is that people attach an intrinsic cost to model-based control through its reliance on cognitive control and that this cost is factored into a cost–benefit analysis that determines the allocation of metacontrol. Initial evidence for this hypothesis came from a study in which people exerted more model-based control in response to amplified reward, but only when this strategy was likely to earn more reward than model-free control (Kool, Gershman, et al., 2017). These results suggest that people adaptively arbitrate between model-free and model-based control through cost–benefit analysis. However, this study tested only sensitivity to increased benefits. Keramati, Smittenaar, Dolan, and Dayan (2016) have provided some initial evidence that people are able to use a mixture of planning and habit to navigate multistage decision-making tasks and that increased time pressure reduces the influence of the goal-directed system on this spectrum. However, it remains unclear whether this balance between habit and planning merely reflects the capacity to engage in model-based control or whether it is determined by a value-based, cost–benefit tradeoff.

We hypothesize that by increasing the demands on planning, by increasing the depth of the causal structure, participants will be less willing to incur the increased costs of model-based control and thus rely on the less accurate model-free system.

EXPERIMENT 1

Participants completed a novel multistage decision-making task in which planning demands, but not available rewards, varied from trial to trial. We hypothesized that participants would show a reduced willingness to exert model-based control in response to increased planning complexity.

Participants

One hundred one participants (range = 22–64 years, mean = 36 years, 44 women) were recruited on Amazon Mechanical Turk to participate in the experiment. Participants gave informed consent, and the Harvard Committee on the Use of Human Subjects approved the study.

Participants were excluded from analysis if they timed out on more than 20% of all trials (more than 40), and we excluded all trials on which participants timed out (average 4.2%). After applying these criteria, data from 98 participants were used in subsequent analysis for the multistage paradigm.

Multistage Decision-making Paradigm

Materials and Procedure

The experiment was designed to test whether choice behavior shows a reduction in model-based control in response to an increase in the complexity of the planning demands. Our paradigm (Figure 1A) was an extended form of a recently developed two-step task (Kool et al., 2016). This task can dissociate model-free and model-based control by capitalizing on the ability of the model-based system to plan over an internal model of the task toward goals, whereas the model-free system requires direct experience of response–reward associations to inform its decisions.

Each trial of the task involved either one or two choices between several stimuli, "space stations" or "spaceships," that appeared on a blue earth-like planet background. As

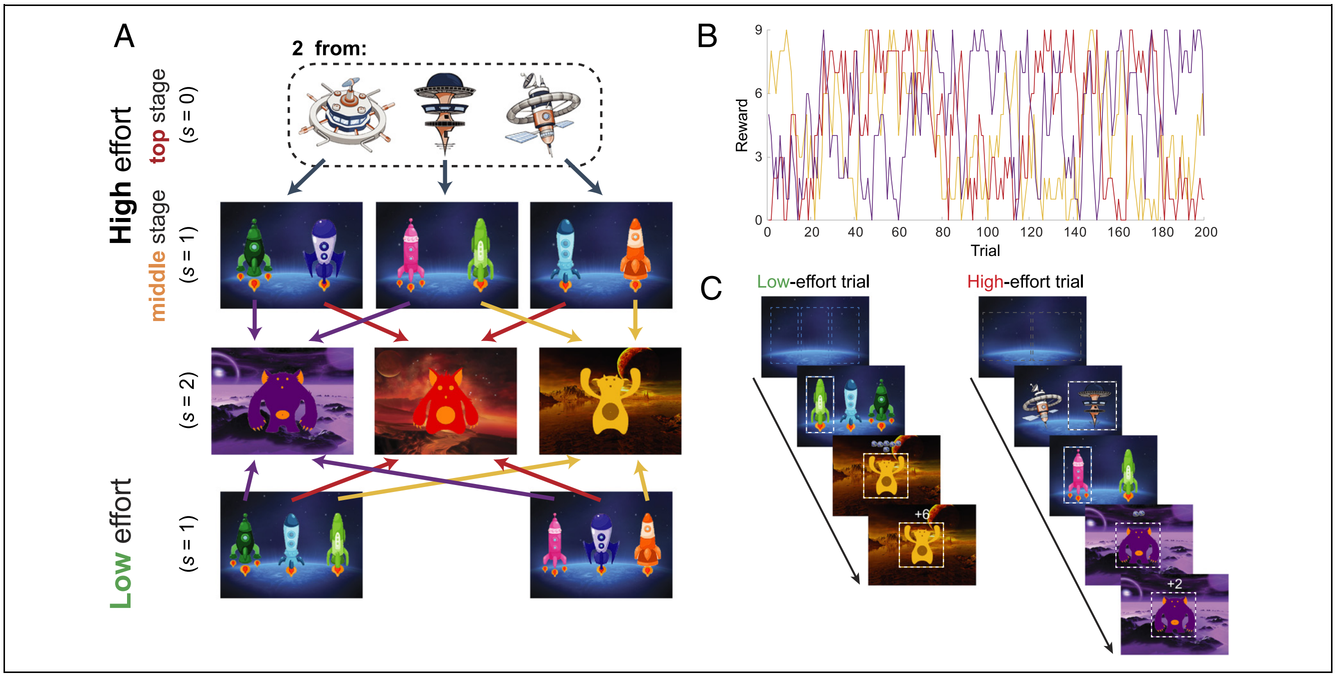

Figure 1. Design of Experiment 1. (A) State transition structure. Low-effort trials (bottom) require a choice between three spaceships that deterministically transition to one of three final-stage states. High-effort trials first require a choice between two randomly selected space stations that deterministically transition to a pair of spaceships. These spaceships then transition to the same final-stage states as in the low-effort trials. Each final stage is associated with a scalar reward. For each level of the transition structure, it is indicated how the stage is indexed by the computational model. (B) The rewards at each final-stage state changed across the duration of the experiment according to a Gaussian random walk with $\sigma = 2$ between 0 and 9. (C) Timeline of events for low- and high-effort trials. At the start of each trial, empty containers indicate whether the following trial would be either a low-effort (three containers) or a high-effort (two containers) trial. After transitioning to the final-stage state, the participant is provided with a scalar reward.

explained in detail below, sometimes these choices involved three options, and at other times they involved two options. The choices were presented side by side, and each choice option had an equal probability of appearing in any position on the screen. Choices between two options had to be made using the "F" or "H" button keys for the left- and right-hand options, and the "G" button key if there was a third option in the middle. All choices had to be made within a response deadline of 2 sec. The selected spaceship and alien were highlighted for the remainder of the response period.

At the start of each trial, it was randomly determined whether it would involve either high- or low-effort demands. Low-effort trials started randomly in one of two possible first-stage states, each of which featured a choice between a triplet of spaceships (Figure 1A, bottom part). This choice deterministically controlled which final-stage state (a purple, red, or yellow planet) would be visited. In each first-stage state, there was always one spaceship that led to each of the three planets.

High-effort trials followed a similar, but slightly different, logic. Here, each trial began with a choice between two of three randomly selected space stations (Figure 1A, top part) in a "zeroth" stage. The zeroth-stage choice deterministically controlled which of three possible first-stage states would be visited. Each of these involved a choice between two spaceships. This choice then determined which of the three final-stage planets would be visited. The first-stage spaceships on the high-effort trials were the same as those on the low-effort trials and transitioned to the same planets as on the low-effort trials. Spaceships that appeared in the same first-stage state on the low-effort trials did not appear together on the first-stage states on high-effort trials. As can be seen in Figure 1A, each possible choice between space stations afforded the possibility to visit any of the final-stage planets. This meant that, on each trial of the task, any planet could be visited.

At the start of each trial, the effort condition was cued by a number of empty containers at the locations where the choices were to appear. These were presented for 1 sec. Each final-stage state was associated with reward. Specifically, on each planet, participants found a single alien, and they were told that this alien "worked at a space mine." They were instructed to press the space bar within the time limit to receive the reward. Participants were told that sometimes the aliens were in a good part of the mine and they paid off a high number of points or "space treasure," whereas at other times the aliens were mining in a bad spot and this yielded fewer pieces of space treasure. The payoffs of these mines changed over the course of the experiment according to independent random walks. One of the alien's reward distributions was initialized randomly within a range of 1–3 points, one within a range of 4–6 points, and the last within a range of 7–9 points. They then drifted according to a Gaussian random walk ($\sigma = 2$), with reflecting bounds at 0 and 9 (for an example, see Figure 1B). New sets of drifting reward sequences were generated for each participant. Participants were given 1¢ for every 10 points. The running score was always presented in the top right corner of the screen.

Each participant completed 25 practice trials followed by 200 rewarded trials (see Figure 1C for an example sequences). Before these, participants were instructed about the reward distributions of the aliens. Next, they practiced traveling to each of the three planets from the low-effort arm and then from the high-effort arm. Specifically, they were required to transition to each planet 10 times in a row separately for the low- and high-effort transition structures. In these practice sessions, there was no time limit for responding.

Experimental Logic

This paradigm is able to distinguish between model-free and model-based influences on choice. To see this, consider the first stage of the low-effort arm in Figure 1A. Crucially, the choices between the three spaceships are equivalent between the two first-stage states. For each triplet, one spaceship always led to the purple planet, one always to the red planet, and one always to the yellow planet. Only the model-based value update capitalizes on this equivalency because it recomputes the expected value of all actions in a manner sensitive to the representation of terminal rewards. In contrast, the model-free update applies only to the specific sequence of actions that preceded reward (Doll et al., 2015). Therefore, on trials that start in a different starting state than the previous trial, only a model-based agent's action values will reflect the reward outcome of the previous trial, because it plans toward the final-stage actions. The model-free system, on the other hand, relies purely on locally learned action–reward associations and is therefore not able to generalize between starting states.

Similar logic applies to the high-effort condition (Figure 1A). Because the model-based system evaluates actions by planning toward the final-stage actions, it can recompute the value of all actions upon learning new information about their rewards. Therefore, if the space station selected on the previous trial is not present in the current trial (two of three space stations get randomly selected on each high-effort trial), only the model-based system will be able to use the previous reward outcome to inform choice. Using the full structure of the experiment, the model-based system is even able to transfer reward information learned in the low-effort condition to the high-effort condition and vice versa, because these conditions share the same final-stage planets and spaceships.

Dual-system RL Model

To estimate the probability of model-free versus model-based control at each choice point, we used an established and validated dual-system RL model (Daw et al., 2005, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010). This model consists of a model-free system and a model-based system that both represent values for the actions at the zeroth and first stages. The systems differ in the way they estimate those values. The model-free system learns "cached"values for all actions in all stages through a simple temporal difference learning algorithm (Sutton & Barto, 1998). In essence, this system simply increases the value of actions that lead to outcomes that are more positive than expected and decreases the value of actions that lead to outcomes that are less positive than expected. The model-based system plans through an internally represented model of the experiment to find the expected final-stage outcomes for each action. The model includes three weighting parameters (w_{low}, $w_{\text{high,top}}$, and $w_{\text{high,middle}}$) that encode the probability of choosing model-based (vs. model-free) value estimates on the first stage of the low-effort arm and on the zeroth and first stages of the high-effort arm, respectively. We predicted a decreased probability of model-based control at the start of high-effort trials as compared with the low-effort trials, reflecting the increased demands of goal-directed planning and, by hypothesis, increased subjective effort cost.

Our multistage decision-making task consists of 12 possible actions distributed across three stages. Low-effort trials start at the first stage ($s = 1$) with three available actions the identity of which is determined by the first-stage state ($\{a_{1,\text{A}}$, $a_{1,\text{B}}$, and $a_{1,\text{C}}\}$ or $\{a_{1,\text{D}}$, $a_{1,\text{E}}$, and $a_{1,\text{F}}\}$; see bottom part of Figure 1A) and then deterministically transitions to one of the final-stage ($s = 2$) states with one available action. High-effort trials involve an additional zeroth stage ($s = 0$) with two randomly selected actions out of a set of three possible actions $\{a_{0,\text{A}}$, $a_{0,\text{B}}$, $a_{0,\text{C}}\}$ before transitioning to the first stage, where there are two available actions the identity of which is determined by the stage 0 choice ($\{a_{1,\text{A}}$, $a_{1,\text{E}}\}$, $\{a_{1,\text{B}}$, $a_{1,\text{F}}\}$, or $\{a_{1,\text{C}}$, $a_{1,\text{D}}\}$; see top part of Figure 1A). Our models consist of model-based and model-free strategies that both learn a function $Q(s, a)$ mapping each stage–action pair to its expected future return (value). On trial t, the zeroth-, first-, and final-stage actions are denoted by $a_{0,t}$, $a_{1,t}$, and $a_{2,t}$, and each stage's rewards as $r_{0,t}$, $r_{1,t}$ (always zero, there is only reward on the final stage), and $r_{2,t}$.

Model-free Strategy

The model-free agent uses the SARSA(λ) temporal difference learning algorithm (Rummery & Niranjan, 1994), which updates Q value for each chosen stage–action pair (s, a) at stage s and trial t according to

$$Q_{\mathrm{MF}}(s, a_{s,t}) = Q_{\mathrm{MF}}(s, a_{s,t}) + \alpha\delta_{s,t}e_{s,t}(s, a)$$

where

$$\delta_{s,t} = r_{s,t} + Q_{\mathrm{MF}}(s+1, a_{s+1,t}) - Q_{\mathrm{MF}}(s, a_{s,t})$$

is the reward prediction error, $a_{s,t}$ is the chosen action at stage s and trial t, α is the learning rate parameter, and $e_{s,t}(s, a)$ is an eligibility trace set equal to 0 at the beginning of each trial and updated according to

$$e_{s,t}(s, a_{s,t}) = e_{s-1,t}(s, a_{s,t}) + 1$$

before the Q-value update. The eligibilities of all state–action pairs are then decayed by λ after the update.

We now describe how these learning rules apply specifically to our task. The reward prediction error is different between the stages of the task. Since $r_{0,t}$ and $r_{1,t}$ are always zero, the reward prediction error at the zeroth and first stages are driven by the value of the selected first- and final-stage actions $Q_{\mathrm{MF}}(1, a_{1,t})$ or $Q_{\mathrm{MF}}(2, a_{2,t})$,

$$\delta_{1,t} = Q_{\mathrm{MF}}(2, a_{2,t}) - Q_{\mathrm{MF}}(1, a_{1,t})$$

and for high-effort trials,

$$\delta_{0,t} = Q_{\mathrm{MF}}(1, a_{1,t}) - Q_{\mathrm{MF}}(0, a_{0,t})$$

Because the trial ends after the final stage, the prediction error on this stage is driven by the reward $r_{2,t}$,

$$\delta_{2,t} = r_{2,t} - Q_{\mathrm{MF}}(2, a_{2,t})$$

The first-stage values are updated at the final stage, with the first-stage values receiving the final-stage prediction error down-weighted by the eligibility trace decay, λ. On high-effort trials, the zeroth-stage values are also updated with the final-stage prediction error, but down-weighted by λ^2. Thus, when $\lambda = 0$, only the values of the final stage get updated.

Model-based Strategy

The model-based algorithm works by learning a transition function that maps the first-stage and zeroth-stage actions to the subsequent states and then combining this function with the final-stage model-free values to compute cumulative state–action values by iterative expectation. In other words, the agent first decides which zeroth-stage and first-stage actions lead to which final-stage state and then looks up the reward values for these final-stage actions.

At the final stage, learning of the immediate rewards is equivalent to the model-free update, because those Q values are simply an estimate of the immediate reward $r_{3,t}$. As we showed above, the SARSA learning rule reduces to a delta rule for predicting the immediate reward. This means that the two approaches coincide at the final stage, and so we set $Q_{\mathrm{MB}} = Q_{\mathrm{MF}}$ at this stage.

The model-based values are defined in terms of Bellman's equation (Sutton & Barto, 1998), which specifies the expected values of each first-stage action using the transition structure P (assumed to be fully known to the agent). For the first-stage, where there is only one available action at the next (final) stage, they are defined as

$$Q_{\mathrm{MB}}(1, a_{1,t}) = Q_{\mathrm{MF}}(2, A(1, a_{1,t}))$$

where $A(1, a_1)$ is the final-stage action that becomes available after taking the first-stage action a_1 given the deterministic transition structure P.

For the zeroth stage, where there are two available actions at the next stage, the model-based actions values are defined as

$$Q_{\mathrm{MB}}(0, a_{0,t}) = \max_{a \in \{A(0, a_{0,t})\}} Q_{\mathrm{MB}}(1, a)$$

where $A(0, a_0)$ is the set of actions that becomes available after taking the zeroth-stage action a_0 using transition structure P. We assume that these model-based estimates are recomputed on each trial using the transition structure P and the final-stage reward values.

Decision Rule

To connect the values to choices, for the zeroth and first stage in the paradigm, the model-free and model-based Q values are mixed according to a weighting parameter w (Daw et al., 2011):

$$Q_{\mathrm{net}}(s, a_i) = wQ_{\mathrm{MB}}(s, a_i) + (1 - w)Q_{\mathrm{MF}}(s, a_i)$$

To accommodate our effort manipulation, we defined three different weights that operated on different trial and choice stages. We set $w = w_{\mathrm{low}}$ for the first stage of the low-effort trials, $w = w_{\mathrm{high,top}}$ for the zeroth stage of the high-effort trials, and $w = w_{\mathrm{high,middle}}$ for the first stage of the high-effort trials. There was no choice at the final stage.

We used the softmax rule to translate these Q values to actions. This rule computes the probability for an action, reflecting the combination of the model-based and model-free action values weighted by an inverse temperature parameter. At both states, the probability of choosing action a on trial t is computed as

$$P(a_{s,t} = a|s) = \frac{\exp(\beta Q_{\mathrm{net}}(s, a))}{\sum_{a'} \exp(\beta Q_{\mathrm{net}}(s, a'))}$$

where a' indexes the current available actions in stage s and the inverse temperature β determines the randomness of the choice.

Table 1. Best-fitting Parameter Estimates Shown as Median and Quartiles across Participants and Experiments

Predictor	*B*	α	λ	w_{low}	$w_{high,top}$	$w_{high,middle}$
Experiment 1						
25th percentile	3.40	.30	0.01	0.48	0.24	0.53
Median	3.99	.75	0.39	0.86	0.50	0.92
75th percentile	5.45	.93	0.80	1.00	0.77	1.00
Experiment 2						
25th percentile	3.42	.61	0.00	0.80	0.53	0.71
Median	4.50	.88	0.64	1.00	0.79	0.98
75th percentile	5.72	1.00	1.00	1.00	0.94	1.00
Experiment 3						
25th percentile	2.43	.07	0.11	0.31	0.06	0.18
Median	3.46	.33	0.42	0.70	0.46	0.65
75th percentile	4.67	.65	0.75	1.00	0.78	1.00
Experiment 4						
25th percentile	2.22	.07	0.00	0.24	0.19	0.35
Median	3.36	.41	0.39	0.81	0.53	0.85
75th percentile	4.63	.67	0.89	1.00	0.81	1.00

Model Fitting Procedure

We used maximum a posteriori estimation with empirical priors, implemented using the *mfit* toolbox (Gershman, 2016) parameters to fit the free parameters in the computational models to observed data, with weak priors for the distributions for the inverse temperature, $\beta \sim$ Gamma (4.82, 0.88) and flat priors for all other parameters (Gershman, 2016). We normalized the reward values to span a range from 0 to 1. We ran the optimization 100 times for each participant with randomly selected initializations for each parameter. The final estimations for the parameters were extracted from the estimation with the maximal log-likelihood. The *t* tests reported here are performed on these subject-wise parameter estimates.[1]

In addition to this, we also used a hierarchical Bayesian modeling procedure, which simultaneously fits the model's parameters for each participant individually, as well as group-level parameters. This method produces a posterior probability distribution over the parameters in our model. For each significant effect from the nonhierarchical analysis, we set up a linear contrast with the means and variances of these distributions and computed the posterior probability that one of the variables (i.e., model parameters) was greater than the other, that is, that the linear contrast was greater than 0 (for more detail on the model-fitting method, see Huys et al., 2011).

Results

Table 1 reports the estimated parameters from our model-fitting procedure. Consistent with prior findings, we found that the weighting parameters indicated a mixture of model-free and model-based decisions at all stages (mean $w = 0.64$). Individual differences in the model-based weighting parameters for all stages significantly predicted the average number of points per trial (w_{low}: $r = .57, p < .001$; $w_{\text{high,top}}$: $r = .20, p < .05$; $w_{\text{high,middle}}$: $r = .52, p < .001$).

Most importantly, we found a significant planning complexity effect: The degree of model-based control was significantly smaller at the start (zeroth stage) of high-effort trials (mean $w_{\text{high,top}} = 0.50$) when compared with the low-effort trials (mean $w_{\text{low}} = 0.70$), $t(97) = 4.46, p < .001, d = 0.45$, posterior probability (*PP*) = .98 (Figure 2A). Model-based control at the start of high-effort trials was also reduced compared with the middle stage of the high-effort trials (mean $w_{\text{high,middle}} = 0.71$),

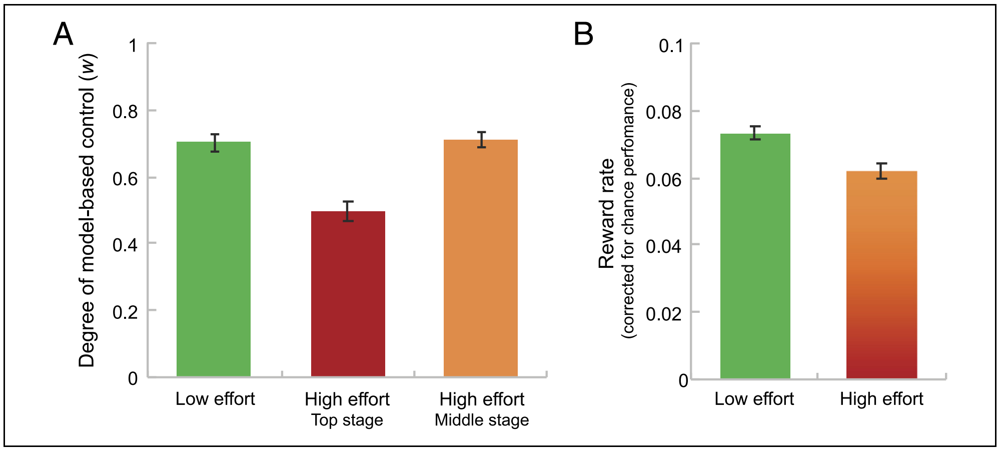

Figure 2. Results of Experiment 1. (A) Degree of model-based control for each of the stages in the experiment. We observed a decrease in model-based control at the start (zeroth stage) of the high-effort trials compared with the start of the low-effort trials and the first stage of the high-effort trials. (B) Average reward rates for low- and high-effort trials, corrected for chance performance. Participants earned significantly less reward on high-effort trials. Error bars indicate within-subject *SEM*. Dashed lines indicate 95% confidence interval.

$t(97) = 4.68$, $p < .001$, $d = 0.47$, $PP = 1.00$, but there was no difference in model-based control between the low-effort trials and the start of the high-effort trials ($t < 1$). In addition, we found that participants earned less reward on high- versus low-effort trials, $t(97) = 2.75$, $p < .01$, $d = 0.28$ (Figure 2B).

One potential concern in the model-fitting procedure above is that the weighting parameter was the only parameter that we allowed to vary between effort conditions. This leaves open the possibility that the observed pattern of weights across conditions was caused by differences in the degree of exploration between the decision stages, which were the forced on the weighting parameter by our fitting procedure. To rule out this alternative explanation, we fit a version of the RL model that varied both the weighting parameters and the inverse temperature for all stages of the task. The results from this analysis replicated the planning complexity effect. Model-based control was significantly lower at the start of the high-effort trials (mean $w_{\text{high,top}} = 0.49$) compared with the low-effort trials (mean $w_{\text{low}} = 0.70$), $t(97) = 4.51$, $p < .001$, $d = 0.46$, $PP = .82$. This result suggests that any differences in the degree of exploration between choice stages were not sufficient to explain the planning complexity effect. However, we did find differences in the inverse temperature parameter between the start of the high-effort trials and the low-effort trials, $t(97) = 6.17$, $p < .001$, $d = 0.62$, $PP = .71$, between the top and the middle stage of the high-effort trials, $t(97) = 5.56$, $p < .001$, $d = 0.56$, $PP = .66$, but not between the low-effort trials and the middle stage of the high-effort trials, $t(97) = 1.57$, $p = .12$, $d = 0.16$.

Discussion

We found that people respond to increased planning complexity by reducing model-based control. This finding is consistent with the proposal that people arbitrate between model-free and model-based control through cost–benefit analysis. Under this account, the increased planning demands on the high-effort trials amplified the cost of model-based control and reduced the willingness to engage in effortful planning.

EXPERIMENT 2

Although the findings from Experiment 1 are consistent with the cost–benefit account, it is also possible that the planning complexity effect was caused by reduced ability, rather than reduced willingness, to exert model-based control. For instance, people may have had difficulty recalling the transition structure of the "high-effort" task simply because it involved a greater number of possible transitions.

We adapted the paradigm from Experiment 1 to rule out this concern. In this new task, participants were given ample time for each decision (10 sec instead of 2 sec) and were trained more extensively on the experiment's transition structure. These changes were introduced to minimize the influence of processing limits on the deployment of model-based control. In addition, we tested this alternative hypothesis by embedding "probe" trials in the multistage task. On these trials, participants were instructed that visiting one planet would lead to obtaining a very large reward, whereas the other planets would result in zero reward. We designed these probe trials in such a way that there was always only one correct action that would lead to the probed planet. Therefore, we were able to use performance on these probe trials as a measure of the ability to plan in this task.

We hypothesized that the increased instruction phase and extended response deadline would lead to an increase in model-based control for all choice stages. However, our cost–benefit hypothesis predicts that, despite this increased ability for goal-directed processing, the effort costs of planning in the high-demand condition would still result in a complexity effect, even in participants with perfect performance on the probe trials.

Methods

Participants

One hundred two participants (range = 21–66 years, mean = 37 years, 39 women) on Amazon Mechanical Turk completed the experiment. No participants timed out on more than 20% of all trials. We excluded all trials on which participants timed out (average 0.6%).

Materials and Procedures

This paradigm was similar to Experiment 1, with a few exceptions. First, we extended the response deadline for all choices to 10 sec. The instruction phase of this experiment was also more elaborate. Instead of learning to transition to each planet separately, we interleaved the planets within each effort condition. At the start of each practice trial, participants were cued with a random planet. Participants were required to visit the cued planet successfully 15 times in a row. This phase was completed for the low- and high-effort arms separately. We reasoned that this new instruction phase would lead to a more accurate internal model for goal-directed planning.

We also modified this task to include a subset of 12 probe trials. At the start of each of these probe trials, a display indicated the presence of a large reward (a "diamond") on one particular planet (see Figure 3A). On those trials, visiting the probed trial would lead to earning this very large reward (200 points, compared with a regular maximal point value of 9), whereas visiting any of the other planets would result 0 points. On the next trial, the final-stage rewards would return to being determined by their drifting reward distributions. Our participants were instructed on this feature of the task,

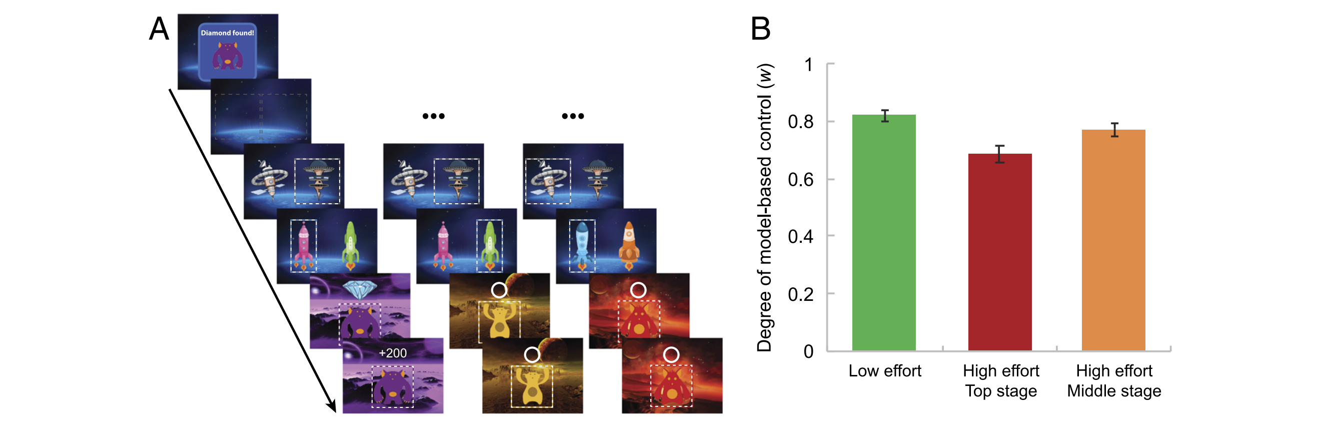

Figure 3. Design and results of Experiment 2. (A) The design was similar to that of Experiment 1, with the exception of the inclusion of a set of "probe trials." The sequences of such probe trials are displayed here. At the start of a probe trial, a display indicated the presence of a large reward (a "diamond") at one final-stage state. If the cued state was visited on that trial, the participant would receive a large reward (leftmost example sequence), but if one of the noncued states was visited, the participants received zero reward (two rightmost sequences). (B) Degree of model-based control for each of the stages. We observed a complexity planning effect, a decrease in model-based control at the start of the high-effort trials compared with the low-effort trials.

and we only used the choices on the regular trials of the dual-system RL model.[2]

The first probe trial was always presented on the 66th trial, and from that moment every 12th trial would involve a probe. Every planet was cued four times, because for each of the two effort conditions there are always two distinct paths toward that planet. For the high-effort trials, there are two space stations that allow a transition toward one particular planet. Therefore, we ensured that, on high-effort probe trials, there was always one correct space station choice at the start of the trial. The order of the probes was determined randomly for each participant. For each effort condition, we calculated probe accuracy as the proportion of trials on which the first decision would lead to the cued planet, excluding trials on which the participants timed out on this first choice.

Results

The estimated parameters for the dual-system RL model are reported in Table 1. We again found that model-free and model-based strategies were mixed in our population (mean $w = 0.76$). Consistent with the hypothesis that the increased response deadline and extensive training on the transition structure would lead to a more accurate internal model, we found that average model-based control was significantly increased in Experiment 2 compared with Experiment 1 for the low-effort trials (mean difference in $w_{\text{low}} = 0.19$), $t(198) = 4.08$, $p < .001$, $d = 0.58$, and for the start of high-effort trials (mean difference in $w_{\text{high,top}} = 0.18$), $t(198) = 3.48$, $p < .001$, $d = 0.49$, but only numerically for the middle stage of the high-effort trials (mean difference in $w_{\text{high,middle}} = 0.07$), $t(198) = 1.42$, $p = .16$, $d = 0.20$.

We again found that individual differences in the model-based weighting parameters correlated with the reward rate across for all stages (w_{low}: $r = .57, p < .001$; $w_{\text{high,top}}$: $r = .38, p < .001$; $w_{\text{high,middle}}$: $r = .57, p < .001$). Participants earned significantly less reward on high- versus low-effort trials, $t(101) = 2.22$, $p < .05$, $d = 0.22$.

Even though our extended instruction phase led to a greater reliance on model-based control, we still observed a complexity effect (Figure 3B): Model-based control was significantly reduced at the start (zeroth stage) of the high-effort trials (mean $w_{\text{high,top}} = 0.69$) compared with the low-effort trials (mean $w_{\text{low}} = 0.82$), $t(101) = 3.19$, $p < .01$, $d = 0.32$, $PP = .73$. Like in Experiment 1, model-based control at the start of high-effort trials was also reduced compared with the middle stage of the high-effort trials (mean $w_{\text{high,middle}} = 0.77$), $t(101) = 2.04$, $p < .05$, $d = 0.20$, $PP = 1.00$, and there was again no difference in model-based control between the low-effort trials and the middle stage of the high-effort trials, $t(101) = 1.50$, $p = .14$, $d = 0.15$.

We now turn to the key question of whether the complexity effect merely reflects an inability to engage in goal-directed processing by examining the performance on the probe trials. Average accuracy scores were high for both low-effort probe trials (mean = 0.92, SD = 0.18) and the high-effort probe trials (mean = 0.88, SD = 0.16). Most importantly, we observed significant complexity effects when we restricted our analysis to participants with 100% accuracy on the high-effort probe trials, $t(58) = 2.49$, $p < .05$, $d = 0.32$, $PP = .98$, the low-effort probe trials, $t(77) = 3.93$, $p < .001$, $d = 0.45$, $PP = .83$, or both, $t(53) = 2.02$, $p < .05$, $d = 0.28$, $PP = .74$. These results are consistent with the cost–benefit framework: Participants with a perfect internal model still withdrew model-based control on regular high-effort trials.

We found a positive correlation between the planning effect and probe accuracy for the low-effort trials ($r = .24$, $p < .05$), but not for the high-effort trials ($r = -.05$,

p = .60). Individual differences in probe accuracy between effort conditions was significantly correlated with the planning complexity effect (r = .28, p < .01). Note that this correlation validates our approach, as one cannot plan accurately over an inaccurate model.

Discussion

Our new instruction phase and increased response deadline increased participants' ability to employ model-based control, yet we still observed a significant planning complexity effect. Most importantly, we found that participants with perfect accuracy on probe trials showed a reliable complexity effect.

EXPERIMENT 3

The results from the previous experiments are consistent with our cost–benefit hypothesis, but also with two alternatives. First, in those experiments, rewards on low-effort trials were delivered sooner after the first choice than in high-effort trials (where those choices occurred one "stage" earlier). Perhaps participants exhibited less effort because of the reduced reward incentive due to temporal discounting. Second, in both experiments, the degree of model-based control correlated less with average reward rate at Stage 0 then at Stage 1. We have previously shown that people reduce model-based control when it cannot reliably deliver a reward advantage (Kool, Gershman, et al., 2017; Kool et al., 2016), potentially explaining the reduced model-based control observed at Stage 0.

Experiment 3 addresses these alternatives. First, we designed a new task where the correlation between w and reward was equated between stages, as assessed by simulation. Second, we equated the elapsed time between initial choice and reward between the high- and low-effort tasks.

Simulations

We used a generative version of the dual-system RL model to estimate the strength of the relationship between model-based control and reward in our task. Specifically, we simulated performance on 200 trials of either the low- or high-effort arm for agents with RL parameters from the median fits from Experiment 1, but which varied from completely model-free (w = 0) to completely model-based (w = 1). For the high-effort trials, we assumed that w was the same in both stages. For each of these allocations between model-free and model-based control, we recorded the reward rate and calculated the strength of the relationship between w and the reward rate using linear regression. We repeated this process 10,000 times. We found that the reward–control tradeoff was identical between the two conditions (Figure 4A). Next, we adopted the same approach to test for differences in the tradeoff between Stage 0 and Stage 1 decisions of the high-effort task. When we varied w for one stage, we set w for the other stage to 0.5. Mirroring our correlational results, we observed that the control–reward tradeoff was lower for the first stage compared with the second stage (Figure 4B), which is consistent with the idea that the planning effect was driven by differences in the control–reward tradeoff between stages of the task.

Our next goal was to equate the control–reward correlation both across tasks and across stages. We accomplished this by setting it to zero in all cases, guided by an approach we employed in prior research (Kool et al., 2016). Specifically, we (1) changed the reward distribution at the final stage from drifting scalar rewards to reward probabilities and (2) changed the parameters of our Gaussian random walk that determines the change rate of the reward distributions so that they match that of the original two-step paradigm (reflecting bounds at 0.25 and 0.75 and σ = 0.025; Daw et al., 2011). The narrow bounds for this Gaussian walk reduce the distinguishability between final-stage action values, which hurts the effectiveness of the model-based system. Furthermore, the rate of change in this task is slow enough for the model-free system to "catch up" with the model-based system, which is typically better at incorporating sudden changes in its value function.

As we expected, these changes equalized the tradeoff at every choice stage of our task. In simulations, regardless of whether we contrasted the low- and high-effort trials (Figure 4C) or the different stages of the high-effort trials (Figure 4D), we observed no relationship between model-based control and reward for all choice stages.

Methods

Participants

One hundred participants (range = 20–69 years, mean = 35 years, 45 women) were recruited on Amazon Mechanical Turk. Two participants timed out on more than 20% of all trials. We excluded all trials on which participants timed out (average 4.8%).

Materials and Procedures

This paradigm was similar to Experiment 1, but there were a few changes. We adopted the instruction phase from Experiment 2. However, based on feedback from participants in that experiment, we decreased the threshold for passing the transition learning phase from 15 correct trials to 10 correct trials.

Participants were told that the aliens in the final stage of the trial were sometimes in a good part of the mine, where they were more likely to give a space treasure. At other times, the aliens were mining in a bad spot, and they were given treasure. The reward probabilities of the three mines changed slowly according to a Gaussian random walk with reflecting bounds at 0.25 and 0.75 and σ = 0.025. One mine was initialized with a probability of .3, the other with a probability of .5, and the last with a

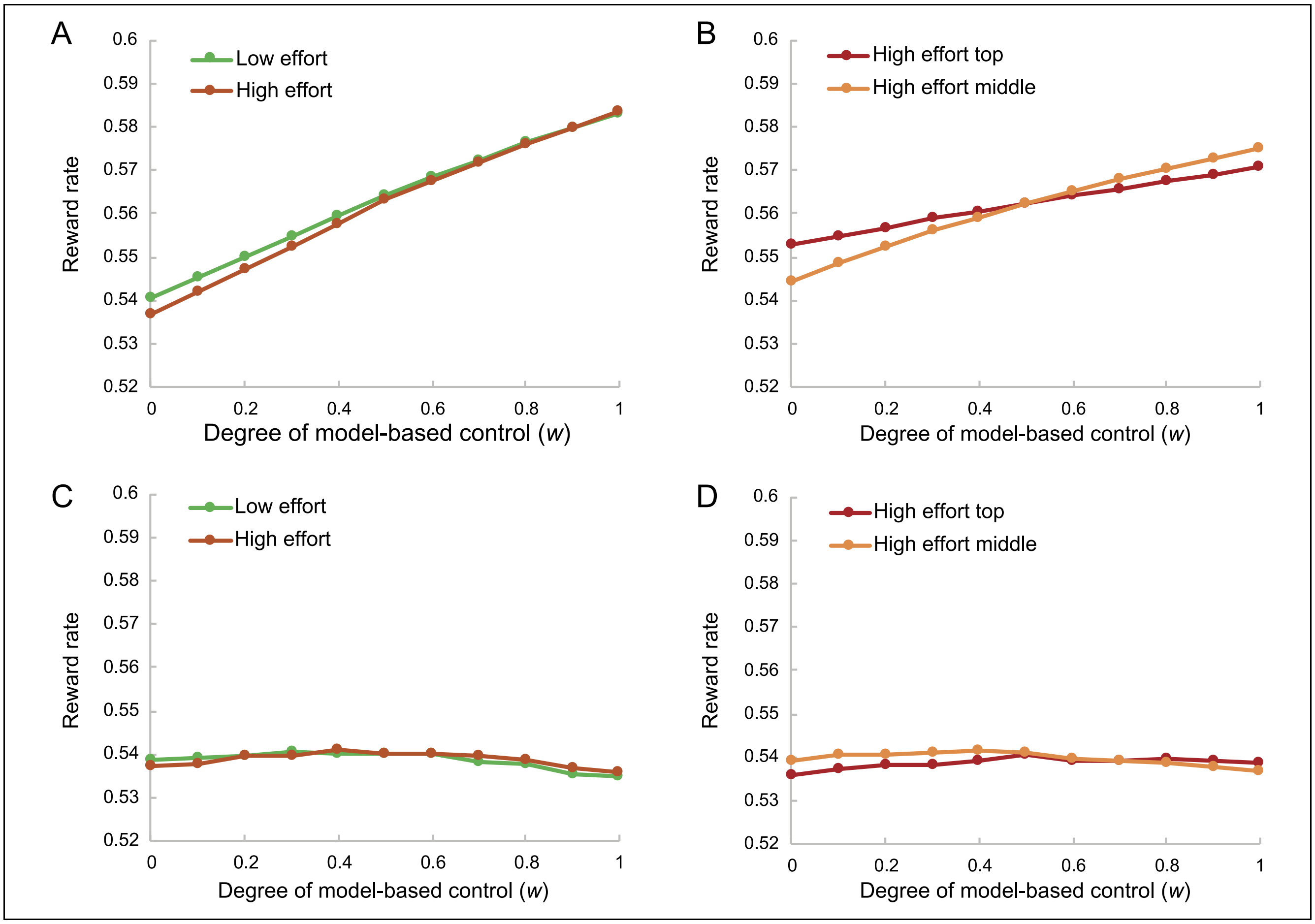

Figure 4. Results of simulations of the control–reward tradeoff in Experiment 1 and Experiment 3. (A) The low- and high-effort conditions of the multistage paradigm in Experiment 1 show a strong and similar relationship between model-based control and average reward. (B) Within the high-effort trials of the multistage paradigm in Experiment 1, the relationship between model-based control and average reward was stronger for the second choice stage compared with the start of the trial. (C) The multistage paradigm of Experiment 3 included stochastic instead of scalar rewards and a different random Gaussian walk governing their drift. These changes eliminated the control–reward tradeoff in both the low- and high-effort trials. (D) The revised design of our multistage paradigm similarly eliminated the relationship between model-based control and average reward in both choice stages of the high-effort trials, resulting in an equal control–reward tradeoff between them.

probability of .7. At the end of the experiment, participants received 9¢ for every 10 points they earned, so that the maximal reward on each trial was worth the amount in cents compared with the previous experiments (Experiments 1 and 2: 9 points; Experiment 3: 1 point; both 0.9¢).

We also equated the time between the start of the trial and the reward outcome between effort conditions. As in Experiment 1, the spaceship and alien were highlighted on screen for the remainder of the response deadline. Moreover, on low-effort trials (which has fewer choices), the selected action was highlighted for an additional 1 sec to equate the time until the reward outcome with the high-effort trials (high-effort trials 3 × 2 sec = 6 sec, low-effort trials: 2 × 3 sec = 6 sec).

Results

The estimated parameters from the dual-system RL model are reported in Table 1. Model-free and model-based strategies both influenced choice behavior (mean $w = 0.55$). Consistent with our simulations, we found that individual differences in the model-based weighting parameters did not predict the participant's reward rate for all stages (w_{low}: $r = .16$, $p = .11$; $w_{\text{high,top}}$: $r = -.03$, $p = .77$; $w_{\text{high,middle}}$: $r = .17$, $p = .09$). There was no difference in reward between effort conditions ($t < 1$).

Despite our task modifications, we again replicated the complexity effect (Figure 5A): Model-based control was significantly reduced at the start of the high-effort trials (mean $w_{\text{high,top}} = 0.47$) compared with the low-effort trials (mean $w_{\text{low}} = 0.61$), $t(97) = 2.80$, $p < .01$, $d = 0.28$, $PP = 1.00$. We found that model-based control was marginally reduced when comparing the middle stage of the high-effort trials and the low-effort trials (mean $w_{\text{high,middle}} = 0.58$), $t(97) = 1.95$, $p = .05$, $d = 0.20$, $PP = .81$. There was no significant difference in model-based control between the low-effort trials and the middle stage of the high-effort trials ($t < 1$).

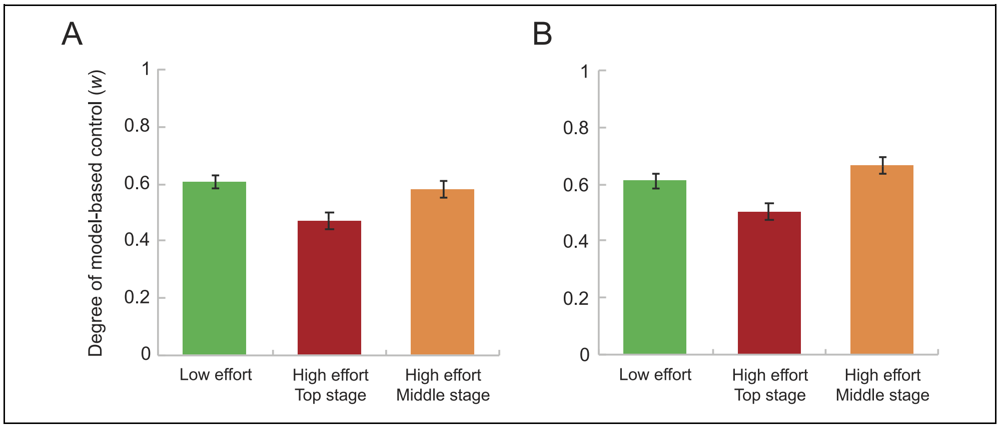

Figure 5. Degree of model-based control for each all the choice stages in Experiments 3 (A) and 4 (B).

Discussion

In this experiment, we observed a significant planning complexity effect, even though we controlled for the timing between the start of trial and the reward outcome, even though we equated the control–reward tradeoff between all choice stages, and even though participants again were extensively trained on the experiment's transition structure. These results suggest that the planning complexity effect is best explained by a cost–benefit account.

EXPERIMENT 4

In this study, we included the probe trials from Experiment 2 in the design of Experiment 3. We did this to replicate the results from Experiments 2 and 3. In addition, by comparing the degree of model-based control in Experiment 2 to the current experiment, we were able to test the effect of the reward–control tradeoff on metacontrol, because they matched on all other features (e.g., the response deadline). We predicted that the absence of the model-based reward advantage in this study would reduce its influence across all stages.

Methods

Participants

One hundred two participants (range = 21–59 years, mean = 34 years, 37 women) were recruited on Amazon Mechanical Turk to participate in the experiment. No participants timed out on more than 20% of all trials. We excluded all trials on which participants timed out (average 0.01%).

Procedure and Analysis

Experiment 4 involved the adapted version of the multistage paradigm used in Experiment 3, but with the probe trials as implemented in Experiment 2. The response deadline for each choice was 10 sec. The analysis of the performance on the probe trials was identical to Experiment 2.[3]

Results

The estimated parameters from the dual-system RL model are reported in Table 1. As before, choice reflected a mixture of model-free and model-based control (mean $w = 0.59$).

We found that individual differences in the model-based weighting parameters did not predict the participant's reward rate for all stages (w_{low}: $r = .14$, $p = .16$; $w_{\text{high,top}}$: $r = .03$, $p = .73$; $w_{\text{high,middle}}$: $r = .15$, $p = .15$), as predicted by the simulations of Experiment 3. There was no difference in earned reward between effort conditions ($t < 1$).

We also replicated the planning complexity effect (Figure 5B). The degree of model-based control was significantly lower at the start of the high-effort trials (mean $w_{\text{high,top}} = 0.50$) compared with the low-effort trials (mean $w_{\text{low}} = 0.61$), $t(101) = 2.17$, $p < .05$, $d = 0.22$, $PP = 1.00$. Model-based control was also reduced when comparing the middle stage of the high-effort trials and the low-effort trials (mean $w_{\text{high,middle}} = 0.67$), $t(101) = 3.04$, $p < .01$, $d = 0.30$, $PP = .92$. There was no difference in model-based control between the low-effort trials and the middle stage of the high-effort trials, $t(101) = 1.20$, $p = .23$, $d = 0.12$.

Probe accuracy was once again high for the low-effort trials (mean = 0.89, $SD = 0.20$) and the high-effort trials (mean = 0.81, $SD = 0.20$). Most importantly, we found a significant complexity effect for participants with 100% accuracy on the high-effort probe trials, $t(41) = 3.39$, $p < .01$, $d = 0.52$, $PP = .79$, or the low-effort probe trials, $t(67) = 3.96$, $p < .001$, $d = 0.48$, $PP = .87$, or both, $t(35) = 3.04$, $p < .01$, $d = 0.51$, $PP = .81$.

Consistent with the idea that the brain is sensitive to the reward–control tradeoff, we found that model-based control was significantly decreased in Experiment 4 compared with Experiment 2 for the low-effort trials (mean difference in $w_{\text{low}} = 0.21$), $t(202) = 4.12$, $p < .001$, $d = 0.57$, the start of the high-effort trials (mean difference in $w_{\text{high,top}} = 0.18$), $t(202) = 4.00$, $p < .001$, $d = 0.56$, and the middle stage of the high-effort trials (mean difference in $w_{\text{high,middle}} = 0.11$), $t(202) = 2.08$, $p < .05$, $d = 0.29$. This result suggests that people adapt their control allocation according to the efficiency of the model-based system. However, we found no difference in the size of the complexity effect between experiments ($t < 1$). This result is consistent with the cost–benefit account, because the subjective effort cost of model-based control should be tied to the planning demands and not its reward efficiency.

GENERAL DISCUSSION

Humans use diverse decision-making mechanisms, and these embody distinct tradeoffs between accuracy and computational demand. The RL framework is widely used to model this tradeoff. Yet, it remains poorly understood how the brain decides from moment to moment which system to use—whether to favor the accuracy of model-based planning or instead the reduced cognitive demands of model-free habits. We address one aspect of this computation, asking whether people are sensitive to task-specific variability in the complexity of planning required. Specifically, we found that the influence of the deliberative, model-based system was reduced when its exertion required planning over a more complex internal model whereas the influence of the simpler habitual or model-free system became relatively stronger. The second experiment replicated this finding and demonstrated a planning complexity even in participants with perfect knowledge of the task, suggesting that it was at least partly driven by a disinclination, rather than an inability, to exert model-based control. The third and fourth experiment also replicated the planning complexity effect and ruled out that our findings were due to increased temporal discounting when planning over a deeper internal model or to a reduced reward advantage of model-based control on the first step of the multistage transition structure.

There is growing interest in the possibility that people allocate control between model-based and model-free decision-making strategies by performing some variety of cost–benefit analysis. On this view, the planning complexity of a task may participate in setting the task-specific cost of mental effort. Within such a cost–benefit framework, our findings suggest that the brain estimates the expected reward of using each system but that it discounts the estimate for the model-based control by its increased effort costs. Indeed, in Experiments 1 and 2, where model-based control had a reward advantage, participants earned less reward on trials with high planning demands, indicating that they gave up some monetary reward to relinquish model-based control.

The present results naturally complement the prior finding that people increase model-based control on trials with larger incentives, but only when this strategy yielded more accurate performance (Kool, Gershman, et al., 2017). Together, these studies lend complementary support to the cost–benefit hypothesis, because they show that the arbitration between model-free and model-based control can be altered by manipulating either side of the tradeoff: the rewards of accuracy and the costs of computation.

Our findings leave open several possible mechanisms for how planning costs are computed by the brain. Notably, this estimation of planning costs recapitulates the unavoidable tradeoff between accuracy and computational demand. At one extreme, people could use simulation of the planning process to gain an on-the-fly estimate of its effort costs (see Pezzulo, Rigoli, & Chersi, 2013, for a similar account of the estimation of model-based values). However, such a process would itself impose substantial cognitive demands. In other words, "model-based" computation of expected effort costs can be self-defeating or even introduces the specter of infinite regress (Boureau et al., 2015), because a key goal of metacontrol is to precisely to minimize unwarranted effort costs.

At the other extreme, the metacontrol process could operate with a cached (i.e., model-free) estimate of each system's costs and benefits, thus avoiding the computational cost of deriving these values via online planning (Gershman, Horvitz, & Tenenbaum, 2015). There are several ways that the cached value might be derived. One possibility is that experienced rewards directly reinforce strategy selection—a "meta" level of model-free learning (Braem, 2017). Another possibility is that the brain uses a heuristic approach in estimating the effort costs, without assessing the task-specific cognitive demands. For example, Dunn, Lutes, and Risko (2016) have argued that the perception of effort can be explained by participants' reliance on cues that are shaped by intuitive theories rather than experienced effort costs. Under any of these scenarios, our results provide initial footing for a computational formulation of arbitration processes.

Our study does not afford direct inferences about neural function, but some indirect inferences are warranted. Research on cognitive costs and on the exertion of model-based control suggests a key role for the dorsolateral pFC (dlPFC), a region of the brain that has long been known to be critical for the exertion of cognitive control (Miller & Cohen, 2001). For example, Smittenaar and colleagues (2013) disrupted activity in dlPFC while participants were performing a two-step task and observed a decrease in model-based control (see also Lee, Shimojo, & O'Doherty, 2014; Gläscher et al., 2010). Meanwhile, McGuire and Botvinick (2010) showed that increased activity in dlPFC during task switching predicted increased subjective effort costs. It is interesting to consider whether, in this study, the enhanced cost of planning in high-effort trials was estimated by the required increase in dlPFC activity as participants mentally navigated the transition structure, resulting in an enhanced aversion to exert model-based control on high-effort trials.

Finally, the current paradigm may have applications for the understanding of clinical disorders resulting from disrupted metacontrol. Recent findings have shown that subclinical measures of individual differences in psychopathology are predicted by individual differences in the model-based control index by the two-step task (Gillan et al., 2016). From a cost–benefit framework, reduced model-based control can be attributed either to an increase in the subjective cost of planning or else to a reduction in the subjective value of its associated reward. We have shown that psychopathology does not reduce the effect of increased reward incentives on increased model-based control (Patzelt et al., submitted for publication). This suggests that, in some clinical disorders, such as obsessive-compulsive disorder, the shift in meta-control primarily reflects an enhanced cost of model-based control. The current paradigm may aid in further developing this hypothesis; for instance, it predicts that psychopathology should moderate the effect of the complexity effect, because this manipulation is directly related to participants' sensitivity to the cost of planning.

Our results show that the allocation of model-based control to a sequential task scales with the intensity of planning demands. This comports with the proposal that arbitration between habit and planning is a form of cost–benefit decision-making.

Acknowledgments

We thank the Moral Psychology Research Laboratory and Computational Cognitive Neuroscience Laboratory for advice and assistance. This research was supported by grant N00014-14-1-0800 from the Office of Naval Research and by the Center for Brains, Minds and Machines, funded by NSF STC Award CCF-1231216, and by the Pershing Square Fund for Research on the Foundations of Human Behavior.

Open practices

All tasks and data are available on the Open Science Framework (https://osf.io/793yw/).

Reprint requests should be sent to Wouter Kool, Department of Psychology, William James Hall, Harvard University, Cambridge, MA 02139, or via e-mail: wkool@fas.harvard.edu.

Notes

1. In addition to model-fitting techniques, researchers using the two-step task often assess performance in a more "direct" manner by inspecting the probability that the action (or terminal state) of the previous trial is repeated as a function of the main effect of the outcome of that trial (positive or negative) and its interaction with some feature of the transition structure (whether the same actions are available or the type of transition of the previous trial). We do not perform these analyses for two reasons. First, the reward structure of our task does not allow us to unambiguously identify model-free control as a main effect of the previous outcome (for a detailed review, see Feher da Silva & Hare, 2018; Kool et al., 2016). Second, our use of continuous rewards prevents the classification of trials as positive or negative based on just the points earned on that trial. This is because a high number of points can still yield a negative prediction error (if an even higher number of points was expected) and a low number of points can still result in a positive prediction error (if an even lower number of points was expected). Because the estimation of prediction errors requires us to specify a computational model of choice, any advantage of a "direct" measure of control is mitigated.

2. Based on recommendations by a reviewer, we fit a model that incorporated the reward outcomes on the probe trials into the model-free action values, but this model (average Bayesian Information Criterion [BIC] = 478) showed a significantly reduced fit to the data compared with the model used in the main text (average BIC = 379; exceedance probability ≈ 1.00).

3. As in Experiment 2, we also fit a model that incorporated the reward outcomes on the probe trials into the model-free action values. This model (average BIC = 467) again showed a worse fit to the data compared with the model used in the main text (average BIC = 417; exceedance probability ≈ 1.00).

REFERENCES

Botvinick, M. M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology, 66,* 83–113.

Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience, 9,* 16–27.

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences, 19,* 700–710.

Braem, S. (2017). Conditioning task switching behavior. *Cognition, 166,* 272–276.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69,* 1204–1215.

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

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 308,* 67–78.

Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: Neural and behavioral evidence. *PLoS One, 7,* e51637.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80,* 312–325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18,* 767–772.

Dunn, T. L., Lutes, D. J. C., & Risko, E. F. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 1372–1387.

Feher da Silva, C., & Hare, T. A. (2018). A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. *PLoS One, 13,* e0195328.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71,* 1–6.

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

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife, 5,* e11305.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience, 15,* 523–536.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66,* 585–595.

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., et al. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7,* e1002028.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

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

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology, 12,* e1005090.

Kool, W., Cushman, F. A., & Gershman, S. J. (in press). Competition and cooperation between multiple reinforcement learning systems. In R. W. Morris, A. M. Bornstein, & A. Shenhav (Eds.), *Understanding goal-directed decision making: Computations and circuits*. Amsterdam: Elsevier.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost–benefit arbitration between multiple reinforcement-learning systems. *Psychological Science, 28,* 1321–1333.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139,* 665–682.

Kool, W., Shenhav, A., & Botvinick, M. (2017). Cognitive control as cost–benefit decision making. In T. Egner (Ed.), *Wiley handbook of cognitive control* (pp. 167–189). Chichester, United Kingdom: Wiley.

Kurzban, R., Duckworth, A. L., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36,* 661–726.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81,* 687–699.

McGuire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 7922–7926.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24,* 751–761.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E., & Daw, N. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences, U.S.A., 110,* 20941–20946.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience, 27,* 319–333.

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (submitted for publication). Model-based control across the psychopathology spectrum: Impaired, but responsive to incentives.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology, 4,* 92.

Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. Cambridge, United Kingdom: Cambridge University.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., et al. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience, 40,* 99–124.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80,* 914–919.

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

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: Macmillan.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS One, 22,* e68210.

Wunderlich, K., Smittenaar, P., & Dolan, R. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75,* 418–424.