Research Article

Psychological Science
2017, Vol. 28(9) 1321–1333

DOI: 10.1177/0956797617708288
www.psychologicalscience.org/PS

Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems

Wouter Kool[1], Samuel J. Gershman[1,2], and Fiery A. Cushman[1]
[1]Department of Psychology, Harvard University, and [2]Center for Brain Science, Harvard University

Abstract
Human behavior is sometimes determined by habit and other times by goal-directed planning. Modern reinforcement-learning theories formalize this distinction as a competition between a computationally cheap but inaccurate *model-free* system that gives rise to habits and a computationally expensive but accurate *model-based* system that implements planning. It is unclear, however, how people choose to allocate control between these systems. Here, we propose that arbitration occurs by comparing each system's task-specific costs and benefits. To investigate this proposal, we conducted two experiments showing that people increase model-based control when it achieves greater accuracy than model-free control, and especially when the rewards of accurate performance are amplified. In contrast, they are insensitive to reward amplification when model-based and model-free control yield equivalent accuracy. This suggests that humans adaptively balance habitual and planned action through on-line cost-benefit analysis.

Keywords
reinforcement learning, decision making, cognitive control, open data, open materials

Received 12/1/16; Revision accepted 4/14/17

Diverse traditions of behavioral research distinguish between two systems for choosing actions: an automatic system that relies on habit and a controlled system that plans toward goals (Dickinson, 1985; Kahneman, 2003; Sloman, 1996). These systems embody different accuracy-demand trade-offs: The habit system has low computational demands but is often less accurate, whereas the planning system achieves greater accuracy with greater computational demands. We ask a simple question: How do people decide from moment to moment which system to use?

To provide a precise answer, we formalized the notions of habit versus planning in a reinforcement-learning (RL) setting. *Model-free* RL leads to choosing actions that have previously led to reward (Thorndike, 1911), a computationally efficient but inflexible strategy. *Model-based* RL achieves flexibility by planning in a causal model of the environment, a comparatively accurate but inefficient strategy. This formalization has facilitated research on each systems' neural basis (Dolan & Dayan, 2013; Gershman, Markman, & Otto, 2014), dependence on executive control (Otto, Gershman, Markman, & Daw, 2013), and contribution to clinical disorders (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016).

Several theories aim to explain how people allocate control between these systems (Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, Lieder, & Goodman, 2015; Keramati, Dezfouli, & Piray, 2011; Pezzulo, Rigoli, & Chersi, 2013), although surprisingly little experimental research has targeted this question directly (but see Lee, Shimojo, & O'Doherty, 2014). In principle, people could choose to employ the more accurate model-based approach whenever cognitive resources are available, relying on habit only when those resources are already occupied (e.g., under cognitive load; Otto et al., 2013). This "first-come-first-served" approach does not, however, attempt to allocate limited computational resources toward tasks offering the maximum benefit. More

Corresponding Author:
Wouter Kool, Harvard University, Department of Psychology, William James Hall, Cambridge, MA 02139
E-mail: wkool@fas.harvard.edu

sophisticated approaches would be sensitive to two task-specific variables: the amount of reward at stake and the size of the model-based advantage in accuracy. Either of these can be incorporated in isolation; people might increase model-based control whenever stakes are high (ignoring whether model-based control is more accurate), or they might increase model-based control when it has the greatest accuracy advantage (ignoring the amount of reward at stake; Daw, Niv, & Dayan, 2005; Rieskamp & Otto, 2006). Finally, people might adaptively integrate both of these pieces of information in order to estimate the comparative reward advantage of model-based control.

We offer a proposal consistent with this final suggestion: that allocation of control is based on the estimated benefits associated with each system in a given task, weighed against the cost of computational demand. Thus, model-based control would be deployed when the combination of high stakes and a robust accuracy advantage are sufficient to offset the cost implied by its reliance on limited cognitive resources. We explore two untested predictions of this proposal: first, that people will increase model-based control when there is a heightened opportunity for reward and, second, that this effect will be eliminated when model-based control cannot reliably outperform model-free control in its accuracy.

Our proposal requires that people assign a cost to model-based control; otherwise, there is nothing to trade off against its potential for a reward advantage over model-free control. Two literatures provide strong circumstantial evidence that people represent such a cost. First, model-based control depends on the capacity for cognitive control. Cognitive load, which decreases the capacity for cognitive control, increases the influence of the model-free system (Otto et al., 2013). Additionally, the model-based system depends on prefrontal structures closely associated with cognitive control (Gläscher, Daw, Dayan, & O'Doherty, 2010; Lee et al., 2014; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). Second, people assign an intrinsic cost to exercising cognitive control (Botvinick, 2007; Botvinick & Braver, 2015; Kool, McGuire, Rosen, & Botvinick, 2010; Kool, Shenhav, & Botvinick, 2017). Thus, people avoid tasks that demand cognitive control unless its cost is offset by task-specific rewards (Kool et al., 2010; Westbrook, Kester, & Braver, 2013). This intrinsic cost would presumably serve as a heuristic representation of the opportunity cost associated with deploying limited cognitive resources (Kool & Botvinick, 2014; Kurzban, Duckworth, Kable, & Myers, 2013). Combining these insights, we posit that people assign an intrinsic cost to model-based planning because of its reliance on cognitive control, which is balanced against a task-specific representation of its potential reward advantage.

Some form of cost-benefit trade-off occurs in several theoretical models of metacontrol (Gershman et al., 2015; Griffiths et al., 2015; Keramati et al., 2011; Payne, Bettman, & Johnson, 1988; Rieskamp & Otto, 2006). Some of these theories formalize competing control systems within the RL framework but are not supported by direct experimental evidence (Gershman et al., 2015; Griffiths et al., 2015; Keramati et al., 2011). Others have generated data consistent with a cost-benefit trade-off (Payne et al., 1988; Rieskamp & Otto, 2006) but do not formalize control systems in terms of dual-systems RL models. Building on this background, our study was motivated by three goals: first, to provide experimental evidence for cost-benefit analyses in metacontrol; second, to accomplish this in a setting amenable to formal analysis in the RL framework; and third, leveraging this formalization, to assess the adequacy of current theoretical proposals to capture the precise form of cost-benefit analysis. In this article, we show how our approach provides new traction in distinguishing among contemporary theories of metacontrol.

Experiment 1

Method

Participants. One hundred one participants (mean age: 32 years, range: 21–61; 47 female, 54 male) were recruited on Amazon Mechanical Turk to participate in the experiment. This sample size was chosen so that we would achieve approximately 90% power to detect a true medium-size effect (Cohen's $d = 0.3$) with a two-tailed alpha of .05. Participants gave informed consent, and the Harvard Committee on the Use of Human Subjects approved the experiment.

Participants were excluded from analysis if they did not respond within the time limit on more than 20% of all trials (more than 40), and we excluded all trials that timed out before a participant made a response (average = 4.4%). After applying these criteria, we included data from 98 participants in subsequent analysis.

Materials and procedure. The experiment was designed to test whether choice behavior shows increased model-based control in the face of increased incentives—that is, when people stand to gain the most from superior accuracy in performance, offsetting the putative subjective cost of executive control. We used a recently developed two-step task (Kool, Cushman, & Gershman, 2016) based, in part, on work by Doll, Duncan, Simon, Shohamy, and Daw (2015; Fig. 1a). In short, this task dissociates model-free from model-based control by exploiting the ability of

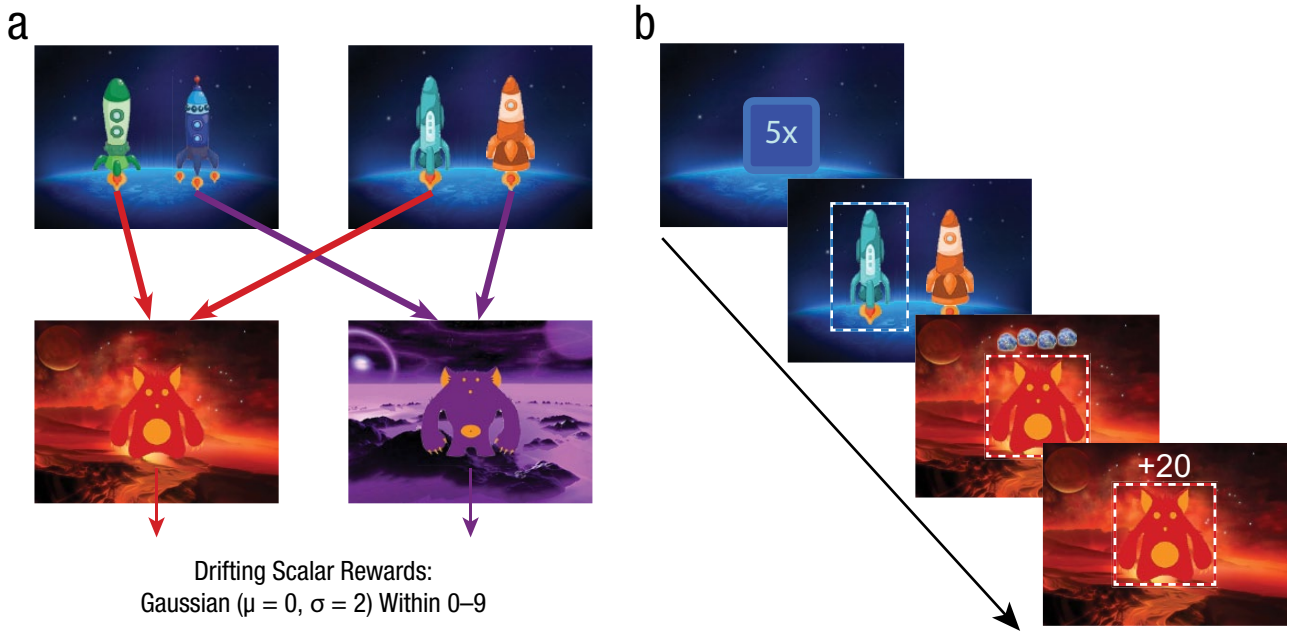

Fig. 1. Experiment 1: design and example trial sequence. There were two possible first-stage states (a), each of which featured a different pair of spaceships that participants had to choose between. Each spaceship was associated with a particular second-stage state (a planet and an alien), which was in turn associated with a scalar reward (between 0 and 9) that changed across the duration of the experiment according to a random Gaussian walk ($\sigma = 2$). At the start of each trial (b), a cue indicated whether the points on that trial would be multiplied by 1 (low stakes) or 5 (high stakes). On the high-stakes trial shown here, the alien gave four pieces of treasure, and so the total points earned was 20. In order to ensure that participants responded at the same rate in all trials, we highlighted the selected spaceship and alien for the predetermined response period.

the model-based system to plan using an explicit model of the task structure, which contrasts with the model-free reliance on the direct experience of action-reward associations.

Each trial started randomly in one of two possible first-stage states, each of which featured a different pair of spaceships that appeared side by side on a blue Earthlike planet background. A given spaceship had an equal probability of appearing on the left or right side of the screen. Participants had to choose between the left- and right-hand spaceships using the "F" and "J" keys within a response deadline of 1,500 ms.

This first-stage choice determined which second-stage state, a purple or a red planet, would then be encountered. Notably, each first-stage state afforded the possibility of transitioning to either planet, with one spaceship always leading to the purple planet and the other always to the red planet. Each second-stage state was associated with a scalar reward. Specifically, on each planet, participants saw 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 in order 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 slowly changed over the course of the experiment according to independent random walks, which encouraged learning throughout the entire session. Note that without this slow change in the scalar rewards, a model-free, trial-and-error strategy would eventually converge with the model-based strategy. The continuous change over time guaranteed that the value representations remained different between systems throughout the experimental session.

One of the alien's reward distributions was initialized randomly within a range of 0 to 4 points, and the other was initialized randomly within a range of 5 to 9 points. They then varied according to a Gaussian random walk ($\sigma = 2$) with reflecting bounds at 0 and 9. A new set of randomly drifting reward distributions was generated for each participant. At the end of the experiment, participants were given 1¢ for every 36 points they earned (i.e., 0.25¢ for a maximal win of 9 points on a given trial).

Both the selected spaceship and alien were highlighted for the remainder of the response period to equate the time demands for the model-based and model-free strategies. This meant that participants were not able to increase their rate of reward over time by responding more rapidly.

The most important feature of the task was that the choices between spaceships were equivalent between the first-stage states. One spaceship in each pair always led to the red planet and alien, whereas the other always led to the purple planet and alien. Because of this equivalence, this task allowed us to distinguish model-based and model-free strategies, since only the model-based system can transfer experiences learned in one starting state to the other starting state. For a pure model-based strategy, each second-stage outcome should always affect first-stage preferences on the next trial, regardless of whether this trial started with the same or the other pair of spaceships, because a model-based strategy plans toward the second-stage goals. In contrast, a pure model-free strategy does not transfer experiences obtained after one pair of spaceships to the other pair, since a model-free strategy learns only action-reward associations (Doll et al., 2015).

As an example, imagine an agent starting in the left panel in Figure 1a. There, he or she chooses the green rocket ship and transitions to the red planet and receives a very large reward. On the next trial, the agent starts in the right panel in Figure 1a. A model-free algorithm will not have updated the value of either of these rocket ships (the orange or the turquoise)—only the green ship in the left panel will have received a boost in value. In contrast, a model-based algorithm will use its model of the transition structure of the task to assign value to all ships on the basis of their probability of leading to the red planet. Thus, the model-based algorithm alone predicts that the agent will exhibit an increased probability of selecting the turquoise rocket ship in the right panel. The analyses reported here exploited this distinction between decision-making strategies.

In order to test our motivational cost-benefit account of metacontrol, we introduced a "stakes" manipulation into this two-step paradigm (Fig. 1b). Specifically, at the start of each trial, a cue indicated a multiplication factor for the subsequent points earned. On 50% of trials, this cue indicated that all points would be multiplied by 5 (high stakes). For example, earning four pieces of space treasure would increase the overall point total by 20 points, and earning nine pieces would increase the total by 45 points. On the other 50% of trials, the cue indicated that the points would be multiplied by 1, so that each piece of space treasure would be worth only 1 point (low stakes). After the reward on each trial was displayed, participants were also shown how many total points they gained on that trial. The running score was always available in the top-right corner of the screen. We predicted that behavioral performance would show increased contributions of model-based choice on high-stakes trials, since this is the reward-maximizing strategy in this task (see the Simulations section; Kool et al., 2016).

Each participant completed 25 practice trials followed by 200 rewarded trials. Before these practice trials, participants were instructed extensively about the transition structure, the reward distributions of the aliens, and how the stakes manipulation worked. These sections were designed to make sure participants fully understood every task element, introducing one at a time and assessing understanding at several points during the practice session. In all these practice and instructional sessions, there was no time limit for any of the responses.

Dual-systems RL model. In order to estimate the degree of model-based control on high- and low-stakes trials, we employed a dual-systems RL model (see the Supplemental Material available online; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). This model consists of a model-free system and a model-based system that both represent values for the actions at the first-stage state. The model-free system learns state-action values for all first- and second-stage states 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 (i.e., that produce a positive prediction error) and decreases the value of actions that lead to outcomes that are less positive than expected (a negative prediction error). The model-based system, on the other hand, computes the values of available actions on the fly by combining the transition structure of the task with the second-stage model-free values to plan toward goals. In other words, this system plans through its internal model of the experiment to find the expected second-stage outcomes for each first-stage action. Our model included two weighting parameters (w_{low} and w_{high}) that determined the relative contribution of the model-based and model-free system on low- and high-stakes trials, respectively. Model-based control was indexed by weights closer to 1, whereas model-free control was indexed by weights closer to 0.

The model also included an *inverse-temperature* parameter β, which controlled the exploitation-exploration trade-off between two choice options given their difference in value. This parameter dictated choice probability through a logistic function ranging from uniform likelihood across actions (pure exploration, insensitive to the value of actions) toward always picking the action with the highest value, regardless of the difference between options (pure exploitation, never exploring lower value options). In addition to these two choice-related parameters, the model also included a learning-rate parameter α that governed the degree

to which action values were updated after a reward outcome, an eligibility trace parameter λ that controlled the degree to which outcome information at the second stage transferred to the start stage, and "stickiness" parameters π and ρ that captured perseveration on either the response or the stimulus choice.

Simulations. In order to confirm that the model-based strategy was indeed the reward-maximizing strategy on this task, we used RL simulations to estimate the relationship between the degree of model-based control and reward. Specifically, we used the dual-systems RL model to simulate performance on the two-step task, without the stakes manipulation, for agents varying from completely model-free ($w = 0$) to completely model-based ($w = 1$). We recorded the reward rate (the average number of points collected per trial) for each of these allocations between RL strategies. We then estimated the strength of the relationship between model-based control and reward through linear regression. This process was repeated 1,000 times across a range of inverse temperatures (βs) and learning rates (αs), which resulted in a surface of standardized regression coefficients in a space of these RL parameters (for details, see Kool et al., 2016).

The results of this analysis are shown in the surface map in Figure 2a. The key feature of this surface map is that there is a strong relationship between model-based control and reward across a large region of the sample parameter space. This analysis confirms that the model-based strategy is associated with increased reward on this two-step task. Therefore, if model-based planning is costly, it would be rational to shift allocation toward the model-based system when potential incentives are high, since the cost-benefit trade-off is more advantageous under those circumstances.

Results

Computational-model analyses. We used maximum a posteriori estimation with empirical priors on the parameters to fit the free parameters to our participant's choices in this task (Gershman, 2016). Table 1 reports the estimated parameters. Replicating previous work, these results showed that the weighting parameters indicated a mix of model-free and model-based action values (mean w = .65). We also confirmed that the model-based strategy was associated with increased accuracy in this task (Kool et al., 2016), since we found that individual differences in the model-based weighting parameters predicted the reward rate, that is, the average number of points per trial (corrected by the average reward value across each participant's reward distribution), ps < .001 (see Table 2). Most important, we found that the degree of model-based control was significantly greater on high-stakes trials (mean w = .76) compared with low-stakes trials (mean w = .54), $t(97) = 4.67$, $p < .001$, Cohen's $d = 0.47$ (Fig. 3).[1] The Supplemental Material reports additional analyses that replicate these findings using a multilevel logistic-regression model.

One potential concern in using this model-fitting procedure is that we varied the weighting parameter between conditions only, which forced any behavioral changes induced by the stake manipulation on this parameter. Therefore, we also fitted a version of the RL model that varied all parameters between the high- and low-stakes conditions.[2] These analyses replicated the effect that the weighting parameter was larger when the stakes were high (mean w = .70) compared with when they were low (mean w = .55), $t(97) = 3.15$, p = .002, $d = 0.32$. In addition, they revealed that the inverse temperature was larger on high-stakes trials compared with low-stakes trials, $t(97) = 3.95$, $p < .001$, $d = 0.40$, which indicates that participants were also more likely to pick the high-value action at the first stage when the stakes were high, rather than exploring the alternative choice option. This is a rational pattern of behavior on this task, since the cost of exploration is higher when the amount of potential reward is greater. The other parameters of the model did not show a significant difference between stake-size conditions, ps > .10 (see the Supplemental Material for more detail).

Behavioral performance. Although our model-fitting procedure has the virtue of precision, it has the drawback of being relatively opaque. In order to provide a more intuitive description and statistical test of our data, we analyzed choice probabilities as a function of the previous trial's reward history and the relation between current and previous first-stage state. The basic rationale for this analysis is that for any model-based strategy, the sign of the second-stage prediction error on the previous trial has to influence the likelihood of retaining the same goal on the next trial. For a model-free strategy, this relationship holds only when a reinforced action is presented in the next trial.

In this paradigm, the implicit equivalence between the two first-stage states allows for such a quantification of the goal-directed component (Doll et al., 2015). The model-based strategy constructs action values by planning toward the second-stage model-free values, which allows it to generalize knowledge learned from both starting states. Thus, prediction errors at the second stage equally affect first-stage preferences, regardless of whether this trial starts with the same starting state as the previous trial. In other words, if the outcome of the previous trial is better than expected (a positive prediction error), the model-based system will be more likely to revisit the previous second-stage state,

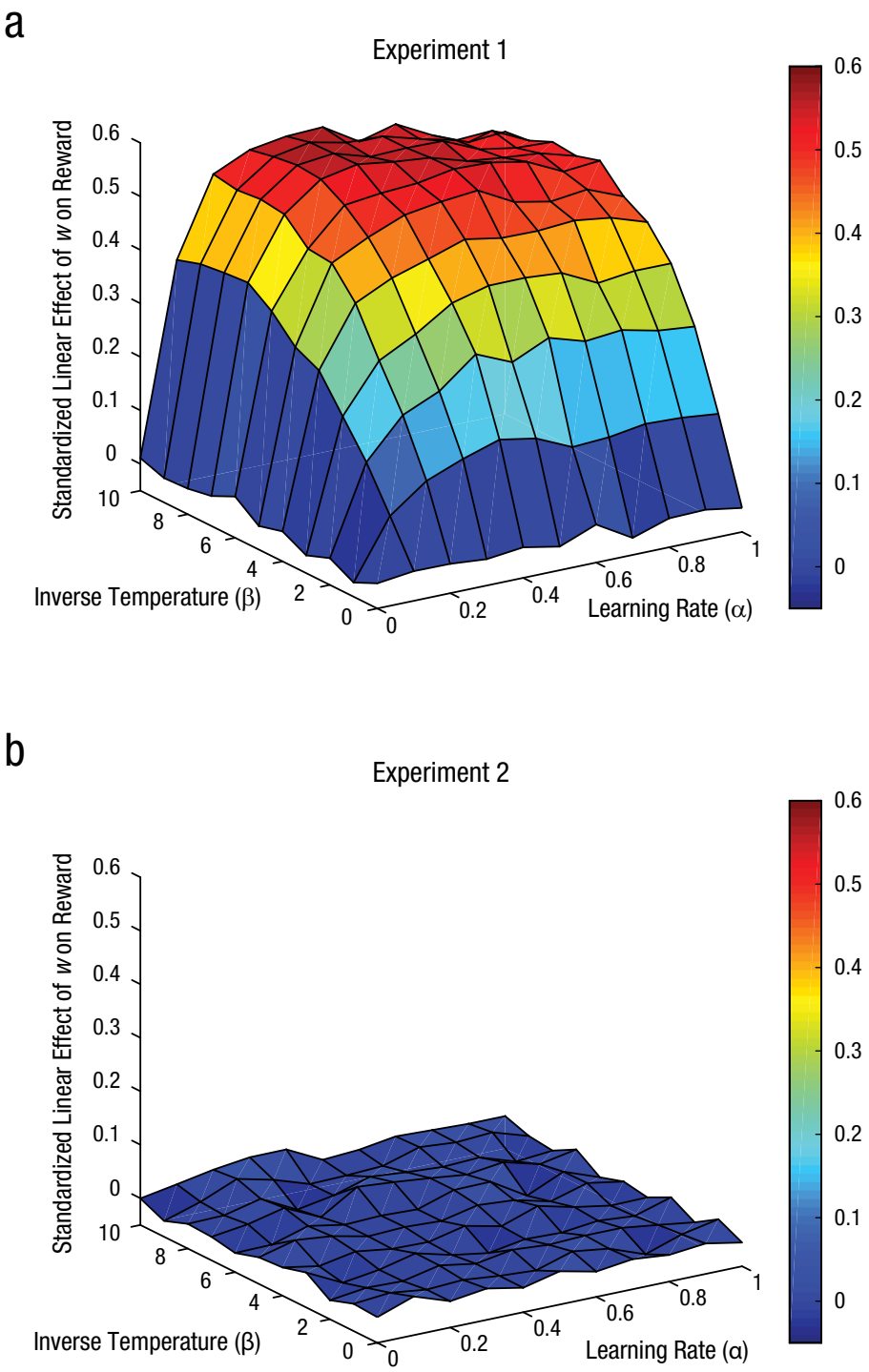

Fig. 2. Surface maps plotting the strength of the control-reward trade-off as a function of learning rate and inverse temperature, separately for (a) Experiment 1 and (b) Experiment 2. The novel version of the two-step task used in Experiment 1 embodies a trade-off between model-based control and reward. For virtually all combinations of learning rate (the degree to which new information is incorporated) and inverse temperature (the randomness of choice), the task shows a strong positive relationship between the degree of model-based control and reward, as measured by linear regression. In the Daw two-step task used in Experiment 2, increased model-based control does not yield increased reward. The plot is uniformly flat around zero across the entire range. See Figures 1 and 5 for specifics of the trial and reward structure of Experiments 1 and 2, respectively.

Table 1. Best-Fitting Parameter Estimates Across Participants in Experiments 1 and 2

Experiment and predictor	β	α	λ	π	ρ	w_{low}	w_{high}
Experiment 1							
25th percentile	0.46	.05	.00	−0.04	−0.34	.00	.63
Median	0.64	.82	.25	0.17	−0.12	.62	.95
75th percentile	2.97	1.00	.85	0.68	0.05	.86	1.00
Experiment 2							
25th percentile	2.10	.02	.39	0.04	−0.02	.00	.00
Median	3.36	.29	.69	0.17	0.05	.43	.44
75th percentile	3.95	.56	1.00	0.37	0.13	.74	.62

independent of which first-stage state is presented. The opposite is true when the previous outcome is worse than expected (a negative prediction error). In that case, the model-based system will reduce the likelihood of revisiting that second-stage state. Thus, the effect of the sign of the previous reward prediction error on the probability that the previous second-stage state is revisited represents the model-based component, since it carries over to the next trial even when the first-stages states are different.

We used this logic to test our cost-benefit hypothesis by analyzing the proportion of trials on which participants revisited the previous second-stage state as a function of the sign of the prediction on the previous trial (positive vs. negative), separately for the high- and low-stakes trials (Fig. 4). The trial-by-trial estimates for the second-stage prediction errors for these analyses were derived from the computational model described above. Consistent with our previous results, these analyses revealed that the model-based choice component (i.e., the effect of the previous trial's prediction error's sign on the probability of revisiting the second-stage state) was significantly higher on high-stakes trials than on low-stakes trials, $t(97) = 3.70$, $p < .001$, $d = 0.37$.

Table 2. Correlation Coefficients Between the Model-Based Weighting Parameter and Reward Rate (Corrected for Differences in Chance Performance) in Experiments 1 and 2

Experiment and parameter	*r*	*p*
Experiment 1		
w_{low}	.54	< .001
w_{high}	.32	.001
Experiment 2		
w_{low}	−.01	.93
w_{high}	−.02	.81

Note: Reward rate was defined as the average number of points per trial.

Discussion

These findings suggest that metacontrol is governed by a cost-benefit analysis. Participants exerted increased model-based control on high-stakes trials compared with low-stakes trials, which indicates increased willingness to engage in effortful planning.

Experiment 2

The results from Experiment 1 are consistent with our cost-benefit hypothesis, but also with a simple decision heuristic that reflexively increases model-based control in any context of enhanced reward opportunity but without assessing the task-specific advantage of either control mechanism, as we proposed. This alternative heuristic model predicted that we should still observe an increase in model-based control on high-stakes trials even when model-based control yields no accuracy advantage.

In prior work (Kool et al., 2016), we found that the original two-step task developed by Daw and colleagues (2011) embodies precisely this detachment of model-based control and performance accuracy. As described in more detail in the Materials and Procedures section, this task comprises only one first-stage state with two choices that lead to the second-stage states with different probabilities: Each choice leads to one state more frequently than the other. Notably, the model-free, but not the model-based, system is affected by these low-probability transitions, since the latter uses the transition structure to discount those associations. Crucially, and in contrast to the two-step task employed in Experiment 1, model-based control in the Daw task does not result in increased reward (see also Akam, Costa, & Dayan, 2015). This difference between the tasks allowed us to discriminate between the two hypotheses about the nature of cost-benefit arbitration. If the stake-size effect is driven by an incentive-based

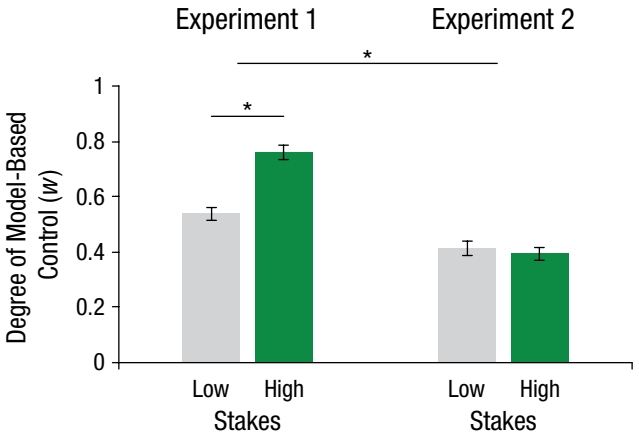

Fig. 3. Degree of model-based control in the low- and high-stakes conditions of Experiments 1 and 2. Error bars indicate ±1 within-subjects *SEM*. An asterisk indicates a significant difference between conditions or experiments ($p < .001$).

heuristic, then high stakes should increase model-based control in both tasks. However, if the brain estimates task-specific values for both systems in order to guide arbitration, then high stakes should not increase model-based control in the Daw task. Here, we tested these contrasting predictions.

Method

Participants. One hundred participants (mean age: 34 years, range: 18–58; 43 female, 57 male) 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 experiment. No participants timed out on more than 20% of all trials. We excluded all trials on which participants timed out (average = 4.1%).

Materials and procedure. The experiment employed the Daw two-step decision making task (Fig. 5a; Daw et al., 2011; Decker, Otto, Daw, & Hartley, 2016). The response buttons and response deadlines were identical to those used in Experiment 1. Participants made an initial choice between a pair of spaceships (green and blue), which led to one of the two second-stage states (red or purple planet). Each spaceship led to one planet more frequently than to the other (70% vs. 30%). On each planet, the participant made a second choice between two aliens that work at different space mines (the second-stage states) and offered them a chance to win a piece of space treasure (in contrast with the scalar reward in Experiment 1). Participants were told that sometimes the aliens were in a good part of the mine, where they were more likely to deliver a piece of space treasure. At other times, the aliens were mining in a bad spot, and they were less likely to deliver space treasure. The reward probabilities of these mines changed slowly over the course of the experiment, which encouraged learning throughout the entire session. One pair of aliens was initialized with probabilities of .25 and .75, and the other pair with probabilities of .40 and .60, after which they changed according to a Gaussian random walk ($\sigma = 0.025$) with reflecting bounds at 0.25 and 0.75 for the remainder of the experiment. A new set of randomly drifting reward distributions was generated for each participant. At the end of the experiment, participants were given 1¢ for every 4 points they earned, so that the maximal reward was worth the same in cents for both experiments (Experiment 1: 9 points; Experiment 2: 1 point; both 0.25¢).

In this task, a pure model-free strategy facilitates learning about the spaceships' action values in an associative manner, which increases the probability of choosing a spaceship if it previously led to reward, independently of the type of transition that preceded this reward. Choice under a model-based strategy, however, takes into account the type of transition that was preceded by the reward, since the values of the first-stage actions are computed in a prospective fashion on the basis of the transition structure and the learned value of each of the second-stage aliens.

As an example, imagine an agent picking the green spaceship, which makes a rare transition to the purple planet, and then receiving a reward. What spaceship should the agent choose on the following trial? A pure model-free agent would be more likely to repeat the

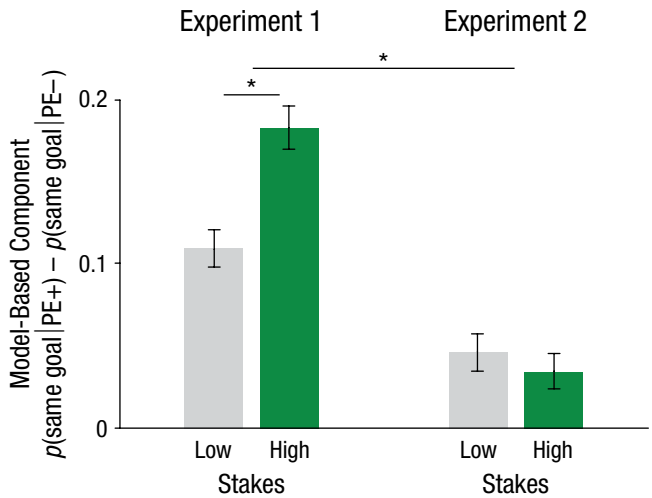

Fig. 4. Model-based choice component for the low- and high-stakes conditions of Experiments 1 and 2. We calculated the model-based choice component as the increase in probability of choosing the same goal as on the previous trial after positive prediction errors (PEs) than after negative PEs. Error bars indicate ±1 within-subjects *SEM*. An asterisk indicates a significant difference between conditions or experiments ($p < .001$).

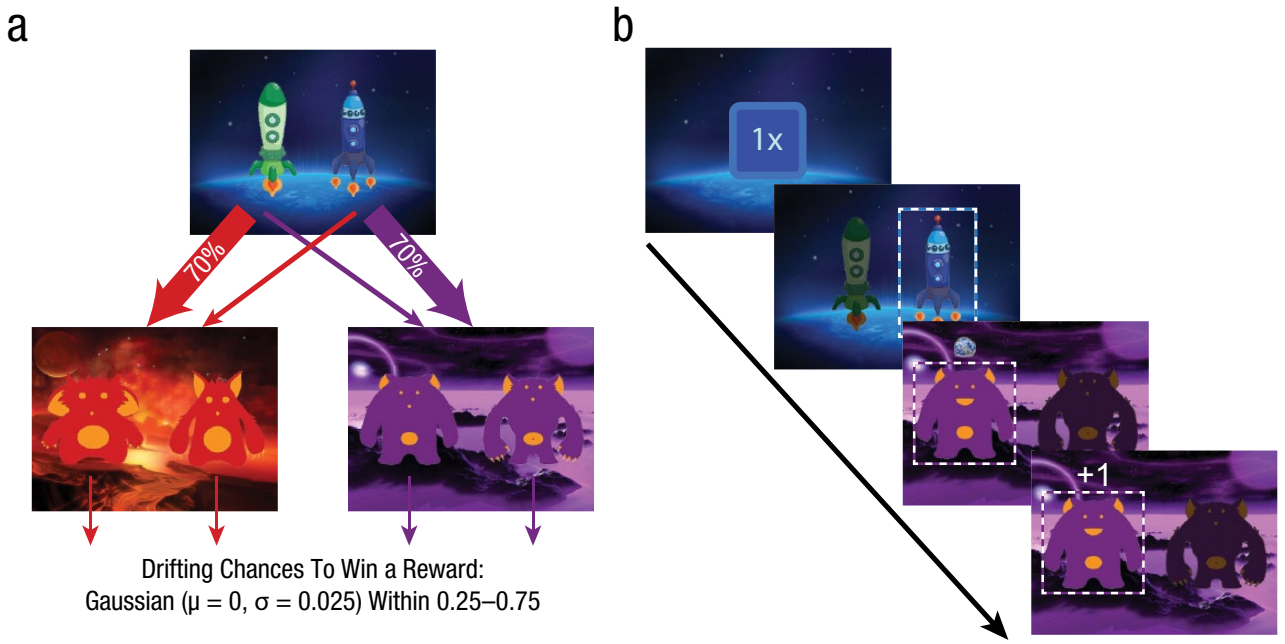

Fig. 5. Experiment 2: design and example trial sequence. Participants made a first-stage choice between a pair of spaceships (a). Each first-stage choice had a high probability of transitioning to one of two second-stage states and a low probability of transitioning to the other. Each second-stage choice was associated with a probability of winning a reward (between 0.25 and 0.75) that changed across the duration of the experiment according to a random Gaussian walk ($\sigma = 0.025$). At the start of each trial (b), a cue indicated whether 1 point (low stakes) or 5 points (high stakes) could be won. On the low-stakes trial shown here, the alien gave one piece of treasure, and so the total number of points earned was 1. In order to ensure that participants responded at the same rate in all trials, we highlighted the selected spaceship and alien for the predetermined response period.

previous trial's choice (green), since this is the choice that led to the positive reward outcome. However, this choice would be less likely to lead to the purple planet where reward was received, given the transition structure of the task. In light of this, a model-based agent would be more likely to switch spaceship choices (to blue) on the next trial, thus maximizing the expectation of reward given the agent's model of the environment.

We implemented the same stakes manipulation used in Experiment 1 (Fig. 5b). At the start of each trial, a cue indicated how many points could be won on a trial. On 50% of trials, this cue indicated that space treasure would be worth only 1 point; on the other 50%, it indicated that space treasure was worth 5 points. After participants observed the reward outcome on the trial, they were also given an explicit computation of the number of points gained on that trial multiplied by the trial's stake. The running score was always available in the top-right corner of the screen.

Participants again completed 25 practice trials followed by 200 rewarded trials. Before these practice trials, participants were instructed extensively on the task, as in Experiment 1. We again employed the dual-systems RL model to capture the degree of model-based control on high- and low-stakes trials. This model was identical to the one described in Experiment 1, but with the number of states and actions changed to accommodate the new task structure.

Simulations. We first used RL simulations to show that the model-based strategy was not reward advantageous in this task, a basic premise of our experimental design. As before, we ran RL simulation to estimate the strength of the relationship between model-based control and reward on this task, across a wide range of inverse temperatures and learning rates (see Kool et al., 2016, for details), by recording the reward rate (i.e., the proportion of trials with a positive outcome) of agents varying from completely model free ($w = 0$) to completely model based ($w = 1$). We then used linear regression to calculate the strength of the relationship between reward and control for each combination of inverse temperature and learning rate.

The results are shown Figure 2b. Notably, the regression coefficients of the resulting surface map are uniformly close to zero. This indicates that nowhere across the sample range of RL parameters was there a positive relationship between model-based control and reward. This confirms that the model-based strategy was not reward advantageous in the Daw two-step task. Therefore, a cost-benefit account would not predict an

increase in control in response to high stakes. However, an incentive-heuristic account, which does not rely on the explicit representation of costs and rewards, would still predict an increase in model-based control on the high-stakes trials.

The absence of this trade-off is produced by the interaction of several factors (for a detailed description, see Kool et al., 2016). First, the model-based strategy is weakened in this task, because the first-stage choices carry relatively decreased importance as a result of the rare transitions, the existence of a second-stage choice, and the low distinguishability between second-stage reward distributions. Furthermore, it is much harder for the model-based system to establish accurate representations of the second-stage probabilistic reward outcomes, compared with the scalar (point value) outcomes used in the novel two-step task. To see this, note that one needs multiple reward observations from the same alien in this experiment, integrating across outcomes, whereas in Experiment 1 a single observation theoretically provides full information about that alien's value.

Results

Computational-model analyses. We fitted the dual-systems RL model to the participants' choices. This model was largely similar to the model used in Experiment 1, except that the number of states and actions were changed to reflect the structure of the novel paradigm. Most important, the model again included weighting parameters that determined the contributions of the model-based and model-free system on the high- and low-stakes trials separately. The Supplemental Material reports additional analyses that replicate these findings using a multilevel logistic-regression model.

The estimated parameters for Experiment 2 are reported in Table 1. As before, we found that the weighting parameters suggested that both model-based and model-free strategies were mixed in the population (mean $w = .40$). Also, consistent with our previous findings and RL simulations, these results showed that individual differences in the model-based weighting parameters did not predict the reward rate (i.e., the proportion of trials with a positive outcome), ps > .80 (see Table 2). Most important, we observed no difference in model-based control in the high-stakes trials (mean $w = .39$) compared with the low-stakes trials (mean $w = .41$), $t(99) = -0.38$, $p = .71$, $d = -0.04$ (Fig. 3), in contrast with the increase in model-based control in Experiment 1.

As before, we estimated a model that varied all parameters between stake-size conditions. This model replicated the finding that model-based control did not differ on high-stakes trials (mean $w = .36$) compared with low-stakes trials (mean $w = .38$), $t(99) = -0.48$, $p = .63$, $d = -0.05$. However, we found that the inverse temperature, controlling the exploration-exploitation trade-off, significantly increased in high-stakes compared with low-stakes trials, $t(99) = 4.00$, $p < .001$, $d = 0.40$. This result suggests that participants showed more exploiting behavior under high-stakes than low-stakes trials, independently of their use of RL strategy. As noted before, this was a rational decision in the current task. Furthermore, this result rules out the potential concern that participants were simply not paying attention to the stake-size cue selectively for this task. Rather, the lack of a stake-size effect on model-based control was the result of an analysis that took into account the costs and benefits of either strategy. The other parameters of the model did not show a significant difference between stake-size conditions, ps > .50 (see the Supplemental Material for more details).

Behavioral performance. In addition to performing the computational-modeling analysis, we again analyzed choice probabilities by computing a metric of model-based influence. As before, the rationale was that for the model-based strategy, the outcome on the previous trial had to influence the likelihood of retaining the identical goal state on the next trial.

Here, the estimation of the model-based influence on choice followed a slightly different logic than in Experiment 1. The model-based strategy used the second-stage model-free values and the experiment's transition structure to compute the expected values of the first-stage actions. Therefore, if a positive outcome was obtained at the second stage, the model-based system would increase the likelihood of planning toward that goal on the next trial. After a rare transition, this meant that the system would decrease the probability of repeating the previous choice after a reward, because this achieves a higher likelihood of getting to the previously rewarded second-stage state. After a rare transition and a loss, the model-based strategy was more likely to stick with the original action, since this decreases the likelihood of getting to the unrewarded state.

We tested our cost-benefit analysis by analyzing the proportion of trials on which participants chose the first-stage action that would most likely lead to the previous second-stage state as a function of the outcome on the previous trial, separately for the high- and low-stakes trials (Fig. 4). We found that this model-based component was not affected by the stakes manipulation, $t(99) = -0.72$, $p = .470$, $d = -0.07$, consistent with the computational-modeling results.

Comparison between experiments

We directly compared behavior between the two experiments in order to test whether the different task structures yielded reliable differences. First, we found that the increase in model-based control induced by increased incentives was significantly larger in Experiment 1 than in Experiment 2, both for the weighting parameter estimated from the computational model (Fig. 3), $t(196) = 3.56$, $p < .001$, $d = 0.51$, as well as for the more direct behavioral estimation of the model-based component (Fig. 4), $t(196) = 3.29$, $p = .001$, $d = 0.47$. Second, confirming earlier results (Kool et al., 2016), we found a reliable shift in model-based control between experiments, with the average weighting parameter significantly higher in Experiment 1 (mean $w = .65$) than in Experiment 2 (mean $w = .40$), $t(196) = 6.21$, $p < .001$, $d = 0.88$. Third, consistent with our previous findings (Kool et al., 2016), multiple regression analyses confirmed that the relationships between the weighting parameter and average reward rate were significantly stronger in the novel paradigm than in the original paradigm for the low-stakes trials, $t(194) = 3.97$, $p < .001$, as well as the high-stakes trials, $t(194) = 2.41$, $p < .05$.

Discussion

In contrast with Experiment 1, participants did not increase model-based control when this strategy was not associated with superior performance. This finding is consistent with a cost-benefit analysis but not with an incentive heuristic in which high stakes always trigger increased model-based control.

If model-based control did not yield increased reward, then why did we still observe a mixture of model-based and model-free strategies in Experiment 2? As we have previously noted (Kool et al., 2016), one possibility is that behavior on this task reflects a prior belief that model-based control is associated with increased reward in the real world (where this is presumably valid). Furthermore, the extensive training on the transition structure may have induced an assumption that it should be employed during task performance.

Of course, the two-step tasks used in Experiments 1 and 2 differed in several respects. We have shown that each of these is necessary to yield a robust accuracy advantage for model-based control (Kool et al., 2016). Although there is a strong a priori basis for predicting that the relationship between stake size and metacontrol depends on our intended manipulation of the model-based accuracy advantage for each task, it may instead be moderated incidentally by some other, correlated difference between the tasks. This is a potential topic for further study.

General Discussion

We found that arbitration between model-based and model-free control is sensitive to the task-specific costs and benefits associated with each system. Participants relied more on model-based control on trials with larger incentives (Experiment 1), but only when this yielded more accurate performance (Experiment 2). This implies that participants estimated the expected reward for each system and then weighed this against the increased costs of model-based control.

Several contemporary models of metacontrol are broadly consistent with the idea of calculating the costs and benefits of each system but differ in their details. For instance, some models posit that control is eventually allocated to whichever system yields greater accuracy (Daw et al., 2005) or reward (Rieskamp & Otto, 2006). This cannot explain our stake-size effect on metacontrol, because the relative accuracy of two strategies will not change under a monotonic scaling of reward. Rather, our data favor models that balance accuracy against the cost of cognitive control (Gershman et al., 2015; Griffiths et al., 2015), such as increased decision time (Keramati et al., 2011) or limited cognitive resources (Kurzban et al., 2013).

How might this cost be assigned? In principle, it could be computed from a model of opportunity costs. For real-world tasks, however, this is likely to be prohibitively demanding. One way around this problem is to assign model-based control an intrinsic subjective cost. Consistent with the observation that model-based control relies on cognitive control (Otto et al., 2013), there have been several prior proposals that cognitive control involves an intrinsic cost (Botvinick & Braver, 2015). For example, Shenhav, Botvinick, and Cohen (2013) propose that the brain selects actions on the basis of the "expected value of control," the expected rewarded associated with exerting cognitive control discounted by the cost of this exertion. Similarly, motivational-intensity theory (Brehm, Wright, Solomon, Silka, & Greenberg, 1983) proposes that effort is invested only when success on the task is both possible and worthwhile, and the effort is justified. Our results suggest that a similar process governs the deployment of model-based control, as formalized within the RL framework.

Our results also demonstrate that the costs of model-based control are weighed against its accuracy benefits in a task-specific manner: High stakes failed to increase model-based control on a task in which it provided no advantage. Several current models of metacontrol cannot accommodate this finding, because they do not explicitly compare the rewards obtained by both systems. Rather, they depend on heuristic approximations

of model-based advantage: for instance, by assuming perfect model-based accuracy (Keramati et al., 2011), calculating errors in the state transitions predicted by the model-based system (Lee et al., 2014), or assuming a model-based advantage when uncertainty in model-free value estimates is high (Keramati et al., 2011; Lee et al., 2014; Pezzulo et al., 2013). Our data instead favor a mechanism that explicitly compares the rewards obtained by model-based with model-free control.

Combining these insights, our results favor a model of metacontrol that would first learn task-dependent reward history of different control mechanisms and then integrate these rewards with their unique cognitive-control costs to guide controller allocation. We provide the first direct support for this class of model by explicitly comparing two tasks that (a) both distinguish model-based and model-free control, (b) vary in the benefits of model-based control (Akam et al., 2015; Kool et al., 2016), and (c) both include trial-by-trial variation in the magnitude of rewards at stake. These findings invite a computational formalization, as well as neuroimaging work to establish the locus of metacontrol in the brain.

Action Editor

Ralph Adolphs served as action editor for this article.

Author Contributions

S. J. Gershman and F. A. Cushman contributed equally to this article. All of the authors contributed to the study design. Testing, data collection, and data analysis were performed by W. Kool. All authors wrote the manuscript and approved the final version for submission.

Acknowledgments

We thank Catherine Hartley for sharing stimuli and the Moral Psychology Research Laboratory and Computational Cognitive Neuroscience Laboratory for advice and assistance.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by Grant No. N00014-14-1-0800 from the Office of Naval Research, by the Center for Brains, Minds and Machines (which is funded by National Science Foundation Science and Technology Center Award CCF-1231216), and by the Pershing Square Fund for Research on the Foundations of Human Behavior.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617708288

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/yg82m/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617708288. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. We replicated this result in an experiment (N = 93) that was identical to Experiment 1, except with negative reward (reward range = −4 to +5). We again found that the degree of model-based control was higher in the high-stakes trials (mean w = .62) than in the low-stakes trials (mean w = .51), $t(93) = 2.81$, $p = .006$, $d = 0.29$.
2. These results should be interpreted with some caution because the inverse-temperature parameter (exploration vs. exploitation) interacts multiplicatively with the weighting parameter (model-based vs. model-free) to determine choice probabilities. This potentially creates a nonidentifiability issue: Different combinations of parameter values can result in the same likelihood (Gershman, 2016).

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology*, *11*(12), Article e1004648. doi:10.1371/journal.pcbi.1004648

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*, 356–366.

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

Brehm, J. W., Wright, R. A., Solomon, S., Silka, L., & Greenberg, J. (1983). Perceived difficulty, energization, and the magnitude of goal valence. *Journal of Experimental Social Psychology*, *19*, 21–48.

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.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*, 848–858.

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

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.

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.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*, 182–194.

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*, Article e11305. doi:10.7554/eLife.11305.001

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

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.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), Article e1002055. doi:10.1371/journal.pcbi.1002055

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*, 131–141.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology*, *12*(8), Article e1005090. doi:10.1371/journal.pcbi.1005090

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, England: John Wiley & Sons.

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

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.

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*, 534–552.

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*, Article 92. doi:10.3389/fpsyg.2013.00092

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

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, NY: 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*, *8*(7), Article e68210. doi:10.1371/journal.pone.0068210