Humans are primarily model-based and not model-free learners in the two-stage task

Carolina Feher da Silva[1] and Todd A. Hare[1]

[1]Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Zurich, Switzerland

July 24, 2019

Abstract

Distinct model-free and model-based learning processes are thought to drive both typical and dysfunctional behaviors. Data from two-stage decision tasks have seemingly shown that human behavior is driven by both processes operating in parallel. However, in this study, we show that more detailed task instructions lead participants to make primarily model-based choices that show little, if any, model-free influence. We also demonstrate that behavior in the two-stage task may falsely appear to be driven by a combination of model-based/model-free learning if purely model-based agents form inaccurate models of the task because of misunderstandings. Furthermore, we found evidence that many participants do misunderstand the task in important ways. Overall, we argue that humans formulate a wide variety of learning models. Consequently, the simple dichotomy of model-free versus model-based learning is inadequate to explain behavior in the two-stage task and connections between reward learning, habit formation, and compulsivity.

Introduction

Investigating the interaction between habitual and goal-directed processes is essential to understand both normal and abnormal behavior [1, 2, 3]. Habits are thought to be learned via *model-free learning* [4], a strategy that operates by strengthening or weakening associations between stimuli and actions, depending on whether the action is followed by a reward or not [5]. Conversely, another strategy known as *model-based learning* generates goal-directed behavior [4], and may even protect against habit formation [6]. In contrast to habits, model-based behavior selects actions by computing the current values of each action based on a model of the environment.

Two-stage learning tasks (Figure 1A) have been used frequently to dissociate model-free and model-based influences on choice behavior [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 6, 20, 21, 22, 23, 24]. Past studies employing the original form of two-stage task (Figure 1A) have always found that healthy adult human participants use a mixture of model-free and model-based learning (e.g. [7, 20, 21]). Moreover, most studies implementing modifications to the two-stage task that were designed to promote model-based over model-free learning [21, 22, 24] find a reduced, but still substantial influence of model-free learning on behavior. Overall, the consensus has been that the influence of model-free learning on behavior is ubiquitous and robust.

Our current findings call into question just how ubiquitous model-free learning is. In an attempt to use a two-stage task to examine features of model-free learning, we found clear evidence that participants misunderstood the task [25]. For example, we observed negative effects of reward that cannot be explained by model-free or model-based learning processes. Inspired by a version of the two-stage decision task that was adapted for use in both children and adults [20, 21], we created task instructions in the form of a story that included causes and effects within a physical system for all task events (Figure 1B-D). This simple change to the task instructions eliminated the apparent evidence for model-free learning in our participants. We obtained the same results when replicating the exact features of the original two-stage task in every way apart from the instructions and how the task's events were framed.

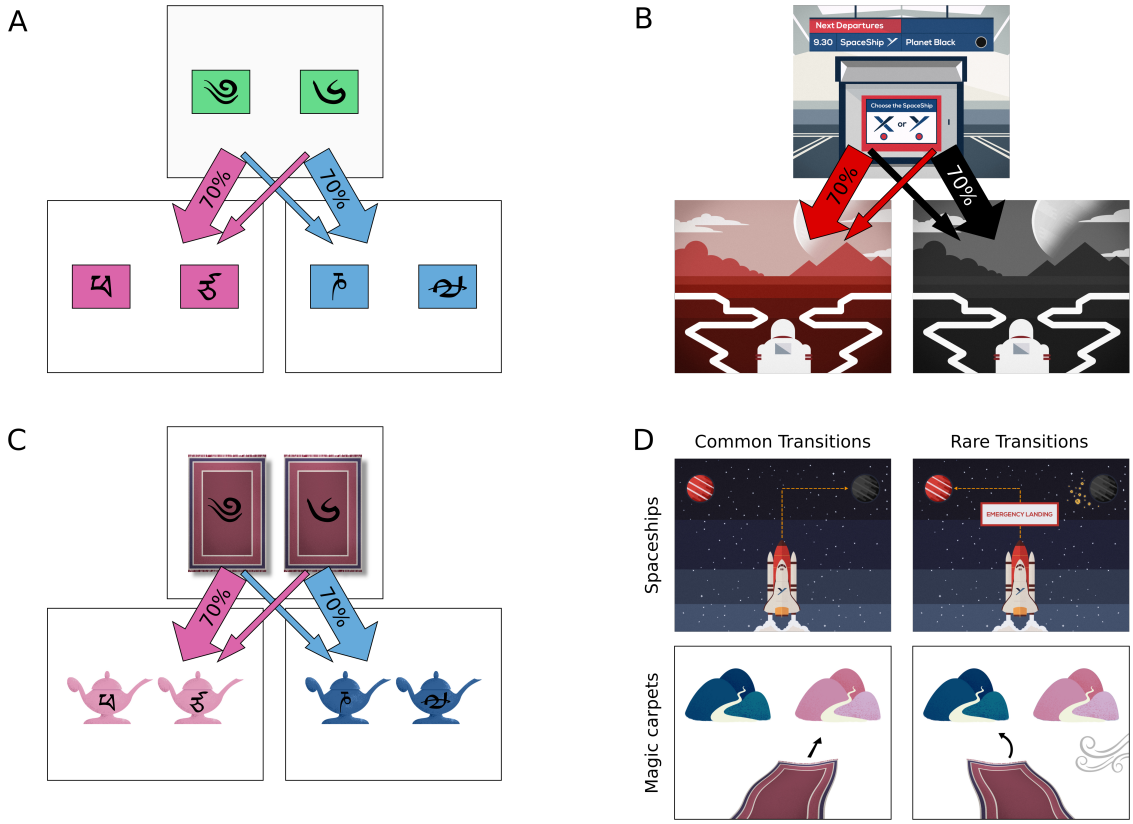

Figure 1: The stimuli used in the three versions of the two-stage task. Panels A-C show the stimuli used for each state in three separate versions of the two-stage task. Panel D shows the stimuli representing common and rare transitions in the spaceship and magic carpet versions of the task. A: The original, abstract version of the task used by Daw et al. [7]. In each trial of the two-stage task, the participant makes choices in two consecutive stages. In the first stage, the participant chooses one of two green boxes, each of which contains a Tibetan character that identifies it. Depending on the chosen box, the participant transitions with different probabilities to a second-stage state, either the pink or the blue state. One green box takes the participant to the pink state with 0.7 probability and to the blue state with 0.3 probability, while the other takes the participant to the blue state with 0.7 probability and to the pink state with 0.3 probability. At the second stage, the participant chooses again between two boxes containing identifying Tibetan characters, which may be pink or blue depending on which state they are in. The participant then receives a reward or not. Each pink or blue box has a different reward probability, which randomly changes during the course of the experiment. The reward and transition properties remain the same in the versions of the two-stage task shown in B and C; only the instructions and visual stimuli differ. B: Spaceship version, which explains the task to participants with a story about a space explorer flying on spaceships and searching for crystals on alien planets. C: Magic carpet version, which explains the task to participants with a story about a musician flying on magic carpets and playing the flute to genies, who live on mountains, inside magic lamps. D: Depiction of common and rare transitions by the magic carpet and spaceship tasks. In the magic carpet task, common transitions are represented by the magic carpet flying directly to a mountain, and rare transitions are represented by the magic carpet being blown by the wind toward the opposite mountain. In the spaceship task, common transitions are represented by the spaceship flying directly to a planet, and rare transitions are represented by the spaceship's path being blocked by an asteroid cloud, which forces the spaceship to land on the other planet. The transitions were shown during each trial in the spaceship task. In order to more closely parallel the original task, transition stimuli were only shown during the initial practice trials for the magic carpet task.

Based on these results, we wondered whether behavior labeled as partially model-free in previous studies could, in fact, be the result of participants performing model-based learning, but using incorrect models of the task to do so. In other words, could misconceptions of the task structure cause participants to falsely appear as if they were influenced by model-free learning? To test this hypothesis, we developed simulations of purely model-based agents that used incorrect models of the two-stage task to make their choices. The results demonstrated that purely model-based learners can appear to be partially model-free if their models of the task are wrong. We also re-analyzed the openly available human choice data from [21] to look for evidence of confusion or incorrect mental models in this large data set. Consistent with our own previous data [25], we found evidence that participants often misunderstood the original two-stage task's features and acted on an incorrect model of the task[1].

Our overall findings show that truly hybrid model-free/model-based learners cannot be reliably distinguished from purely model-based learners that use the wrong model of the task. Critically, they also indicate that when the correct model of the world is easy to conceptualize and readily understood, human behavior is driven primarily by model-based learning.

Results

Model-based learning can be confused with model-free learning

Model-based agents can be confused with model-free agents. The prevalence of such misidentification is in the existing literature is currently unknown, but we present a series of results indicating that it is high. False conclusions about model-free learning can happen when the model used by model-based agents breaks the assumptions that underlie the data analysis methods. We will show that purely model-based agents are misidentified as hybrid model-free/model-based agents in the two-stage task when the data are analyzed by either of the standard methods, logistic regression or reinforcement learning model fitting. We present two examples of incorrect task models that participants could potentially form. We do not suggest that these are the only or even the most probable ways that people may misunderstand the task. These incorrect task models are merely examples to demonstrate our point.

To serve as a reference, we simulated purely model-based agents that use the correct model of the task based on the hybrid reinforcement learning model proposed by Daw et al. [7]. The hybrid reinforcement learning model combines the model-free SARSA(λ) algorithm with model-based learning and explains first-stage choices as a combination of both the model-free and model-based state-dependent action values, weighted by a model-based weight w ($0 \leq w \leq 1$). A model-based weight equal to 1 indicates a purely model-based strategy and, conversely, a model-based weight equal to 0 indicates a purely model-free strategy. The results discussed below were obtained by simulating purely model-based agents ($w = 1$). We also note that consistent with recent work by Sharar et al. [26] we found that even when agents have a w equal to exactly 1, used the correct model of the task structure, and performed 1000 simulated trials, the w parameters recovered by the hybrid reinforcement learning model were not always precisely 1 (see Fig. 2F). This is expected, because parameter recovery is noisy and w cannot be greater than 1, thus any error will be an underestimate of w.

The two alternative, purely model-based learning algorithms we created for simulated agents to use are: the "unlucky symbol" algorithm and the "transition-dependent learning rates" (TDLR) algorithm. The full details of these algorithms and our simulations can be found in the Methods section. Briefly, the unlucky symbol algorithm adds to the purely model-based algorithm the mistaken belief that certain first-stage symbols decrease the reward probability of second-stage choices. We reasoned that it is possible that participants may believe that a certain symbol is lucky or unlucky after experiencing by chance a winning or losing streak after repeatedly choosing that symbol. Thus, when they plan their choices, they will take into account not only the transition probabilities associated to each symbol but also how they believe the symbol affects the reward probabilities of second-stage choices. In the current example, we simulated agents that believe a certain first-stage symbol is unlucky and thus lowers the values of second-stage actions by 50%.

[1] Note that we used these data simply because they were openly available and were from a relatively large sample of people performing the two stage task after receiving, in our view, the best instructions among previously published studies using the two-stage task. We do not expect confusion to be more prevalent in this data set than in any other past work. However, our current results indicate that stringent comprehension checks and double checks should be applied whenever a two-stage task is used.

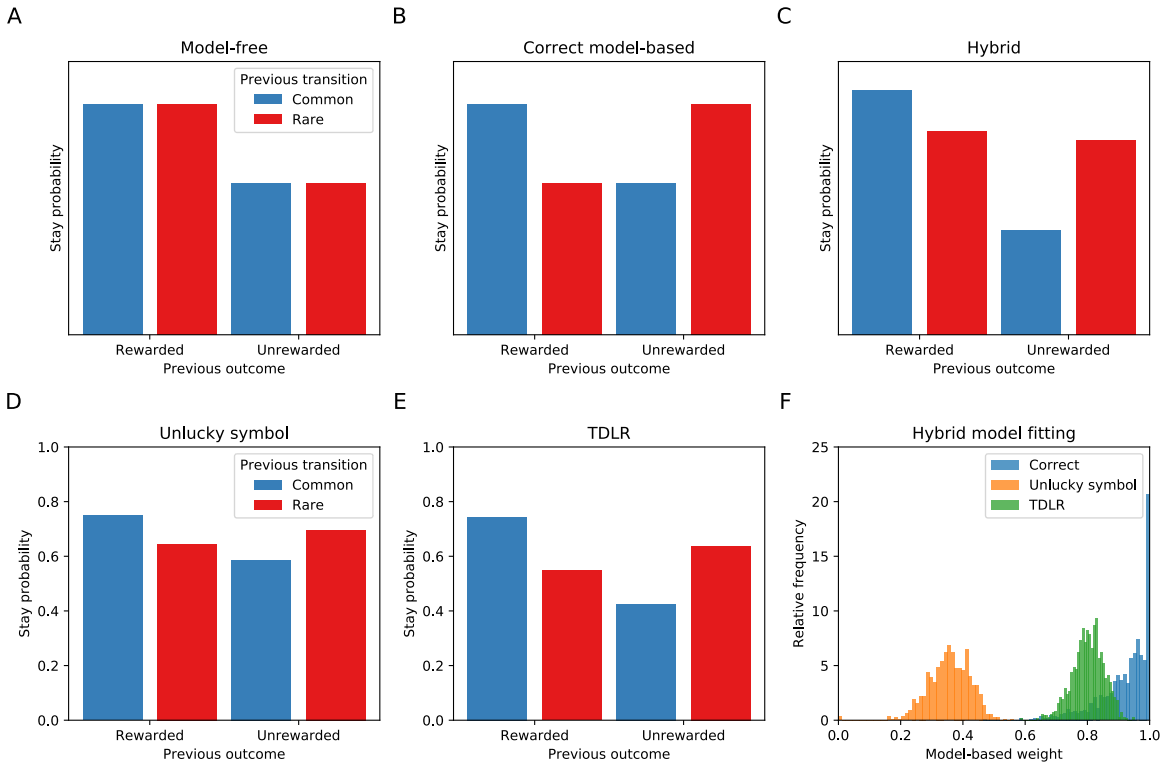

Figure 2: Stay probability and model-based weights for simulated agents. Top row: Idealized stay probabilities of purely model-free (A), purely model-based (B), and hybrid (C) agents as a function of the previous outcome and transition. Model-free and model-based learning predict different behaviors for agents performing the two-stage task. In the two-stage task, model-free learning simply predicts that a first-stage action that resulted in a reward is more likely to be repeated in the next trial. Conversely, model-based learning uses a model (i.e. knowledge or understanding) of the task's structure to determine the first-stage choice that will most likely result in a reward [7]. To this end, it considers which second-stage stimulus it believes to be the best (i.e. most likely to yield a reward at the current point in time). Then it selects the first-stage action that is most likely lead to the best second-stage stimulus. Model-free and model-based learning thus generate different predictions about the stay probability, which is the probability that in the next trial the participant will repeat their previous first-stage choice. Model-free agents exhibit a main effect of reward, while model-based agents exhibit a reward by transition interaction effect, and hybrid agents exhibit both a main effect of reward and a reward by transition interaction effect. To generate the hybrid data plotted in this figure, we used the logistic regression coefficient values for data from adult participants in a previous study [20]. Bottom row: Purely model-based agents that use incorrect models of the two-stage task can look like hybrid agents that use a combination of model-based and model-free learning. Panels D and E show the mean stay probabilities for each type of model-based agent following the four combinations of transition type (Common, Rare) and outcome (Rewarded, Unrewarded). F) The histograms show the fitted model-based weight parameters (w) for simulated agents using the correct (blue), unlucky-symbol (orange), and transition-dependent learning rates (TDLR) (green) models of the task. We simulated 1000 agents of each type and each agent made 1000 choices in the original two-stage task. Model-based weights for each agent were estimated by fitting the simulated choice data with the original hybrid model by maximum likelihood estimation. Error bars for the simulated data are not shown because they are very small due to the large number of data points.

The TDLR algorithm is based on debriefing comments from participants in a pilot study, which suggested that they assign greater importance to outcomes observed after common (i.e. expected) relative to rare transitions. For example, one participant wrote in the debriefing questionnaire that they "imagined a redirection [to the rare second-stage state] with a successful [outcome] as a trap and did not go there again." To formalize this idea, we conceived a simple model-based learning algorithm that had a higher learning rate after a common transition and a lower learning rate after a rare transition. Hence, the learning rates are transition-dependent. Note that although we created this model based on participants' feedback, we are not suggesting it is a model that many or even one participant actually used. We suspect that participants each use a slightly different model and probably even change their mental models of the task over time. In reality, we don't know if and how a given participant misunderstands the task. The TDLR and unlucky symbol algorithms are simply two plausible 'proof of principle' examples to demonstrate our point. We simulated $1,000$ agents of all three purely model-based types (Correct, Unlucky symbol, and TDLR) performing a $1,000$-trial two-stage task. Again, the purpose of these simulations is not to show that real human participants may be employing these specific models when performing a two-stage task. Rather, the proposed algorithms are examples intended to illustrate that when agents do not employ the assumed task model, they may generate patterns of behavior that are mistaken for a model-free influence.

The resulting data were first analyzed by logistic regression of consecutive trial pairs (Figure 2). In a logistic regression analysis of consecutive trial pairs, the stay probability (i.e. probability of repeating the same action) is a function of two variables: reward, indicating whether the previous trial was rewarded or not, and transition, indicating whether the previous trial's transition was common or rare. Model-free learning generates a main effect of reward (Figure 2A), while model-based learning generates a reward by transition interaction (Figure 2B) [7] (although this may not be true for all modifications to the two-stage task [27]). The core finding in most studies that have employed this task is that healthy adult participants behave like hybrid model-free/model-based agents (Figure 2C). Specifically, in the case of a logistic regression analysis on consecutive trial pairs, the results exhibit both a main effect of reward and a reward by transition interaction. Our simulations show that TDLR and unlucky symbol agents, despite being purely model-based, display the same behavioral pattern as healthy adult participants and simulated hybrid agents (Figure 2D and E).

We then analyzed the simulated choice data by fitting them with a hybrid model-based/model-free learning algorithm based on the correct model of the task (i.e. the standard analysis procedure). The resulting distributions of the estimated model-based weights are shown in Figure 2. The median model-based weight estimated for agents using the correct task model was 0.94, and 95% of the agents had an estimated w between 0.74 and 1.00. The estimated w for agents using the other algorithms were, however, significantly lower. The set of model-based agents using the unlucky-symbol algorithm had a median $w = 0.36$, and 95% of the agents had an estimated w between 0.24 and 0.48. The set of agents using the TDLR algorithm had a median $w = 0.80$, and 95% of the agents had an estimated weight between 0.70 and 0.90. Thus, these results demonstrate that analyzing two-stage task choices using a hybrid reinforcement learning algorithm can lead to the misclassification of purely model-based agents as hybrid agents if the agents don't fully understand the task and create an incorrect mental model of how it works.

Human behavior deviates from the hybrid model's assumptions

In order to test if human behavior following typical task instructions violates assumptions inherent in the hybrid model, we re-analyzed the control condition data from a study of 206 participants performing the original two-stage task after receiving a common form of two-stage task instructions that were, in our view, as good or better than all other previous studies [21]. Henceforth, we refer to this as the common instructions data set. First, we note that poor overall hybrid model fits and greater decision noise/exploration were significantly associated with more apparent evidence of model-free behavior. When examining how the overall model fit relates to indications of model-free behavior, we found that maximum likelihood estimates of the model-based weight for each participant correlated with the log-likelihood of the hybrid model's fit. Specifically, we observed a weak but significantly positive correlation between the model-based weight and the log-likelihood of the model fit (Spearman's rho $= 0.19$, $P = 0.005$). Similarly, we found that the soft-max inverse temperature parameters in both the first (Spearman's rho $= 0.24$, $P = 0.0006$) and second-stage (Spearman's rho $= 0.19$, $P = 0.007$) choice functions also correlated with model-based weights. These correlations indicate that more exploratory

decisions, or simply decisions that look noisier according to the hybrid model, are associated with more apparent model-free influence on behavior. This set of results suggests that participants with a lower model-based weight may not necessarily be acting in a more model-free way, but rather, may be acting in a way that deviates from the hybrid model assumptions.

Indeed, the standard hybrid model cannot explain the fact that participants in this experiment behaved differently when first-stage symbols were presented on different sides of the screen in consecutive trials. Note that, in many implementations of the two-stage task, the symbols presented at each stage appear on randomly selected sides. However, the side of the screen a symbol appears on is irrelevant and only the symbol identity matters. Therefore, if participants understand the task and first-stage symbols switch sides between consecutive trials, this should not influence their choices. Nevertheless, Kool et al. anticipated that participants might present a tendency to repeat key presses at the first stage [21] and thus modified the standard hybrid model to add a response stickiness parameter. For comparison with participant data, we simulated hybrid agents with response stickiness that performed 1000 trials of the two-stage task. We then divided consecutive trial pairs from the simulated and participant data sets into two subsets: (1) same sides, if the first-stage choices were presented on the same sides of the screen in both trials, and (2) different sides, if the first-stage choices switched sides from one trial to the next. Each subset was separately analyzed using logistic regressions. The results are presented in Figure 3A and C. Both simulated agents and participants showed a larger intercept in the same sides subset compared to the different sides one. In the simulated data, this effect was caused by response stickiness. However, in contrast to simulated agents, the human participants also showed a larger reward coefficient if the first-stage options remained on the same sides (mean 0.41, 95% highest density interval (HDI) $[0.35, 0.47]$) than if they switched sides (mean 0.14, 95% HDI $[0.08, 0.19]$) with the posterior probability that the reward coefficient is larger in same-side than switched-side trials being greater than 0.9999.

There are several potential explanations for these side-specific results. It could be that model-free learning is sensitive to stimulus locations and/or specific responses (i.e. pressing left or right) and does not generalize across trials if the stimuli and responses change. In other words, it is possible that model-free learning considers that each symbol has a different value depending on where it is presented or on which key the participant has to press to select it. In this case, however, we would expect the reward effect for the different-sides subset to be zero. Another possibility is that when the sides switched, participants were more likely to make a mistake and press the wrong key, based on the previous rather than current trial configuration. To further investigate this later possibility, we fit these data with a hybrid model that included an extra parameter that quantified the probability of making a configuration mistake and pressing the wrong key when first-stage symbols switch sides (see subsection "Fitting of hybrid reinforcement learning models" for details). Note that this configuration mistake parameter is distinct from decision noise or randomness, because it quantifies response probabilities that are specific to cases where the first stage symbols have switched places from one trial to the next, and thus it only decreases effect sizes in the different-sides subset rather than in all subsets (Figure 3B). We found that the estimated median probability of making a configuration mistake when the symbols switched sides was 0.54. However, when looking at individual participants, distinct types of performance were readily apparent. Out of 206 participants, 111 had an estimated probability of making a configuration mistake lower than 0.1 (based on the median of their posterior distributions). In contrast, 51 participants had an estimated probability of making a configuration mistake higher than 0.9. These results suggest that while most participants rarely made configuration mistakes, approximately 25% of them made mistakes more than 9 out of 10 times when the symbols switched sides.

Next, we compared the fits of the configuration mistake model and the standard hybrid model for each participant data using PSIS-LOO scores (an approximation to leave-one-out cross-validation; see Methods section for details). The goodness of fits for the two models were equivalent on average for the 111 low-configuration-mistake participants (mean score difference: 0.0, standard error: 0.1), As expected, the configuration mistake model fit better for the 51 high-configuration-mistake participants (mean score difference: -34.1, standard error: 6.2). However, it is possible that the high-configuration-mistake participants were not truly making mistakes. As shown in Figure 3B, configuration mistakes decrease all effect sizes in the different sides subset, including the reward by transition interaction effect, but this was not observed in the common instructions data set. Rather, some participants may have instead misunderstood the task—for example, they may have believed that the left/right button

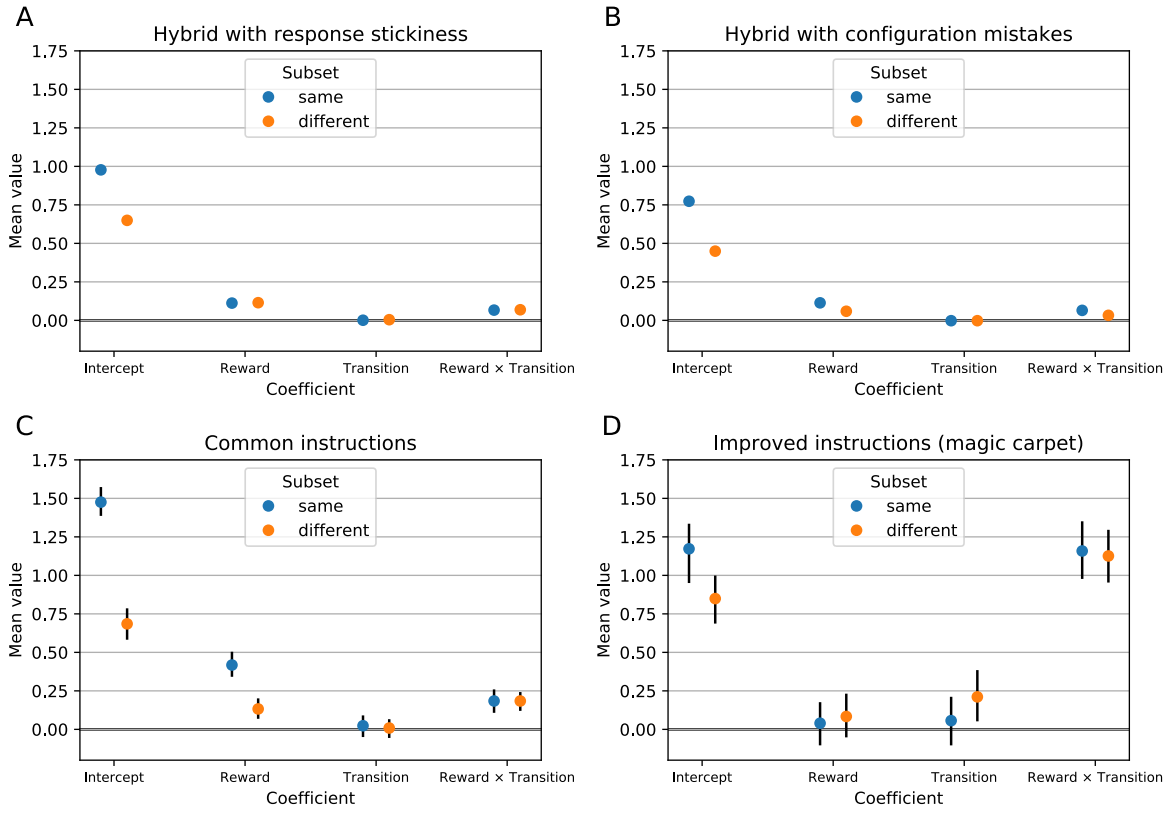

Figure 3: Simulated agents' and real participants' behavior can be influenced by irrelevant changes in stimulus position. All four panels show behavior in the two-stage task on trials in which the position of the first-stage stimuli remains the same (blue) across consecutive trials compared to trials in which the first-stage stimuli are in different (orange) positions from one trial to the next. The coefficients in these graphs are from a logistic regression analysis explaining stay probabilities on the current trial as a function of Reward (rewarded outcome = 1, non-rewarded outcome = 0), Transition (common transition = 1, rare transition = 0), and the interaction between Reward and Transition on the previous trial. In contrast to the typical procedure used in previous studies, which treats all trials as the same, we analyzed the data after dividing the trials into two categories based on whether or not the positions of the first-stage stimuli were the same or different across consecutive trials. A) Results from a simulation of hybrid agents with response stickiness ($N = 5000$) performing 1000 trials of the two-stage task. The median parameter values from the common instructions data set [21] were used in these simulations. B) Results from a simulation of hybrid agents that occasionally made configuration mistakes ($N = 5000$) performing 1000 trials of the two-stage task. The median parameter values from the common instructions data set [21] and a 20% configuration mistake probability were used in these simulations. Error bars are not shown for these simulated results because they are very small due to the large number of data points. C) Results from a re-analysis of the common instructions data set [21] (($N = 206$). This study used story-like instructions, but did not explicitly explain why the stimuli might be on different sides of the screen from trial to trial. The Reward effect significantly changed between trial-type subsets in this data set. D) Results from the magic carpet task ($N = 24$), which provided explicit information about why stimuli may change positions across trials and that these changes were irrelevant for rewards and transitions within the task. There were no significant differences in the regression coefficients between the two subsets of trials on this task. Error bars in panels C and D represent the 95% highest density intervals.

Data set	25th percentile	Median	75th percentile
Daw et al. 2011	0.29	0.39	0.59
Kool et al. 2016	0.00	0.24	0.68
Spaceship	0.61	0.78	0.91
Magic carpet	0.51	0.79	0.85

Table 1: Model-based weight estimates for four data sets: Daw et al. [7] ($N = 17$), Kool et al. [21] (common instructions data set, $N = 206$), the spaceship data set ($N = 21$) and the magic carpet data set ($N = 24$). For the Daw et al. [7] data set, the weight estimates were simply copied from the original article. For the other data sets, we obtained the weight estimates by a maximum likelihood fit of the hybrid reinforcement learning model to the data from each participant.

press was more important than stimulus identity. In any case, these results suggest that *different participants understood the two-stage task in different ways and potentially used different models of the task to make their choices*. Moreover, the large difference in reward effects between same-side and different-side trials indicates that many participants misunderstood basic facts about the task.

Improving the instructions of the two-stage task decreases the apparent influence of model-free learning on behavior

We developed two modified versions of the two-stage task with the goal of clearly explaining all features of the task so participants would be more likely to use the correct model of the task if they were model-based. Specifically, we incorporated a detailed story into the task instructions and stimuli (Figure 1B-D). Previous work has already used stories to explain the two-stage task to human participants [20, 21], but those story-based instructions did not provide a reason for all of the events in the task (see Supplementary Materials for further discussion). We sought to ensure participants' understanding of the task and leave no room for speculation during the experiment. Therefore, we modified the first and second stage stimuli and embedded them within a story that provided a concrete reason for every potential event within the task. We also displayed additional instructions during the practice trials before participants started the task. These instructions explained the reason for each event to the participants again as it happened, in the form of helpful messages that participants could not skip. See the Methods section for descriptions of the magic carpet and spaceship versions of the two-stage task that we developed and links to code for running the tasks.

Hybrid learning model fits indicate that behavior is primarily model-based behavior with comprehensive instructions

We tested the impact of our instructions on the apparent levels of model-based and model-free influences on behavior by fitting the data with the standard hybrid learning model. In order to facilitate comparison with previously reported studies, we fit the model to each participant using maximum likelihood estimation. The estimated model-based weights for the participants who performed the spaceship or the magic carpet task were substantially higher (see Table 1) than the estimated weights found in two previous studies [7, 21] that used less complete task instructions. We also fit the hybrid model to the data from our magic carpet and spaceship tasks as well as the common instructions data using a Bayesian hierarchical model. Our results indicate that the posterior probability that the average weights in the magic carpet and spaceship data sets are greater than the average weight in the common instructions data set is greater than 0.9999.

Standard logistic regression analyses also indicate primarily model-based behavior in the Spaceship and Magic Carpet tasks

We compared the logistic regression results from the common instructions sample [21], to the spaceship and the magic carpet tasks (Figures 3 and 4). First, we used the standard method combining all consecutive trial pairs together regardless of whether or not the stimuli changed sides across trials. Figure 4 shows that when all trials are combined together, the coefficient of the reward by transition

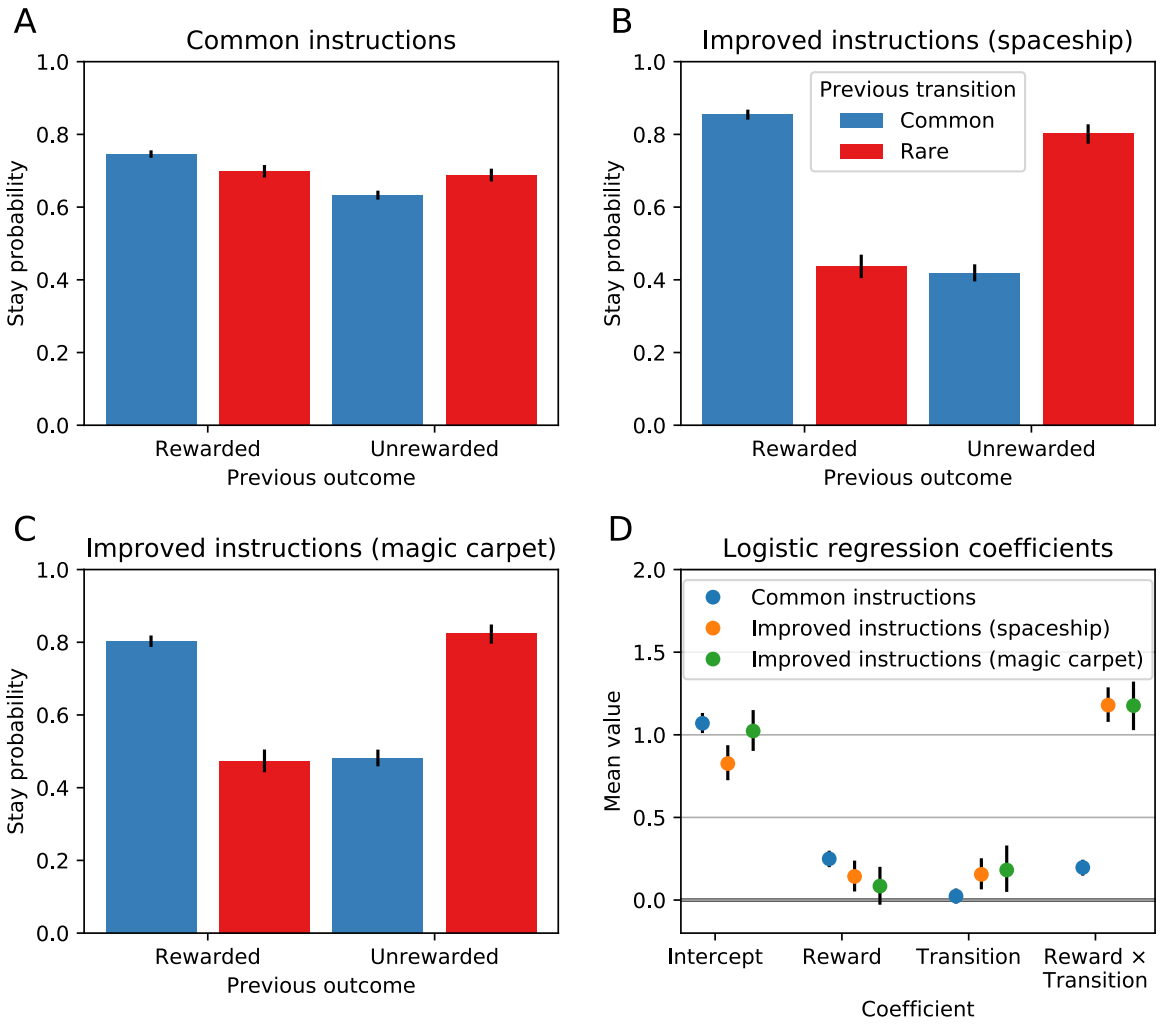

Figure 4: Stay probabilities for three empirical data sets. A) The behavior of participants ($N = 206$) after receiving common instructions [21] shows both a main effect of reward and a reward by transition interaction. In contrast, the choices of participants after receiving improved instructions in the new B) spaceship ($N = 21$) and C) magic carpet ($N = 24$) tasks show a much more model-based pattern. Error bars in panels A to C represent the 95% highest density intervals. D) This plot shows the mean and 95% highest density intervals of all coefficients in the hierarchical logistic regressions on stay probabilities for all three data sets. These logistic regression coefficients used to calculate the stay probabilities shown in A-C. Note that the main effect of reward in the spaceship task is actually inconsistent with a model-free influence on behavior in that task (see Figure 5)

interaction, which indicates model-based control, is 5.9 times larger in the magic carpet (95% HDI $[4.5, 7.3]$) and spaceship (95% HDI $[4.7, 7.2]$) tasks compared to the common instructions results [21].

No side-specific effects and fewer configuration mistakes with enhanced magic carpet instructions

In contrast to the common instructions sample [21], participants who performed the magic carpet task and were given more comprehensive instructions did not show differences between same-side and different-side trial sets in the influence of reward or reward by transition interactions (Fig. 3C). Again, trials were divided into same sides category if the first-stage choices were presented on the same sides of the screen in both trials, and different sides category, if the first-stage choices switched sides from one trial to the next. Recall that we found that first-stage stimulus location significantly influenced behavior in the common instructions sample, with same-side trials showing a significantly larger reward effect than different-side trials (Fig. 3B). The more comprehensive instructions eliminated this difference between trial types, which are, in fact, equivalent under the correct model of the task.

Accordingly, the more comprehensive instructions almost completely eliminated configuration mistakes. We fit the configuration mistake model to the magic carpet data, and the results were that 23 of 24 participants had a mistake probability smaller than 0.05 and the remaining participant had a 97% mistake probability. Thus, we conclude that the enhanced magic carpet instructions vastly increased the probability that participants would act correctly when equivalent choices switched sides compared to the common instructions sample [21].

Spaceship task data reveal misleading evidence for model-free influence

Although choices in both the magic carpet and spaceship tasks showed large reward by transition interaction effects, there is a small, but significant main effect of reward on choices in the spaceship task (lower right panel of Figure 4). This may at first suggest that our enhanced instructions decreased but did not eliminate the influence of model-free learning on these participants' behavior. After all, learning based on the correct model of the task does not introduce a main effect of reward on choices. We took advantage of specific properties of the spaceship task to investigate this reward effect in greater detail. We found that it is misleading as evidence for model-free influence because it contradicts one of the basic properties of model-free learning.

Within the spaceship task there are pairs of first-stage stimuli that indicate the same initial state, but do so by presenting different information. In other words, the stimuli are different, but the initial state is the same. These pairs are shown in Figure 5A. This feature of the spaceship task allows us to subdivide trials into four categories in order to examine the reward effect as a function of both stimuli and required responses. We divided the spaceship trial pairs into four subsets based on the information that was announced on the flight board above the ticket machine in the current and previous trials. The information on the flight boards are the stimuli that determine which first-stage action will most probably lead to a given second-stage state (see Figure 5A). The four trial categories were: (1) same spaceship and planet, if the same flight, with the same spaceship and planet, was announced on the board in both trials, (2) same spaceship, if the announced flights in both trials had the same spaceship, but the spaceship was departing for different planets on each trial, (3) same planet, if the announced flights in both trials had the same planet destination, but different spaceships were scheduled to fly there, and (4) different spaceship and planet, if the flight announced in the previous trial had a different spaceship and planet destination than the flight announced in the current trial. We analyzed the data from each trial pair category using the standard logistic regression approach.

If the reward effect was driven by model free learning, then we would expect that it was positive only when the stimulus-response pairing is identical across trials (i.e. category (1), same spaceship and planet). Model-free learning is frequently assumed to be unable to generalize between distinct state representations [15, 16, 21]. Thus it should have no influence on choices in categories 2-4 (Figure 5B). The analysis results, however, are contrary to the expectation from a model-free driven reward effect (Figure 5C). The observed reward effect had a similar magnitude for all four categories: 0.12 (95% HDI $[-0.10, 0.32]$) for the same spaceship and planet, 0.14 (95% HDI $[-0.03, 0.32]$) for the same spaceship, 0.14 (95% HDI $[-0.04, 0.33]$) for the same planet, and 0.19 (95% HDI $[0.02, 0.36]$) for a different spaceship and planet. The corresponding posterior probabilities that the reward effect is greater than zero in each of the four categories are 87%, 94%, 93%, and 98%, respectively. Analyzing trials from

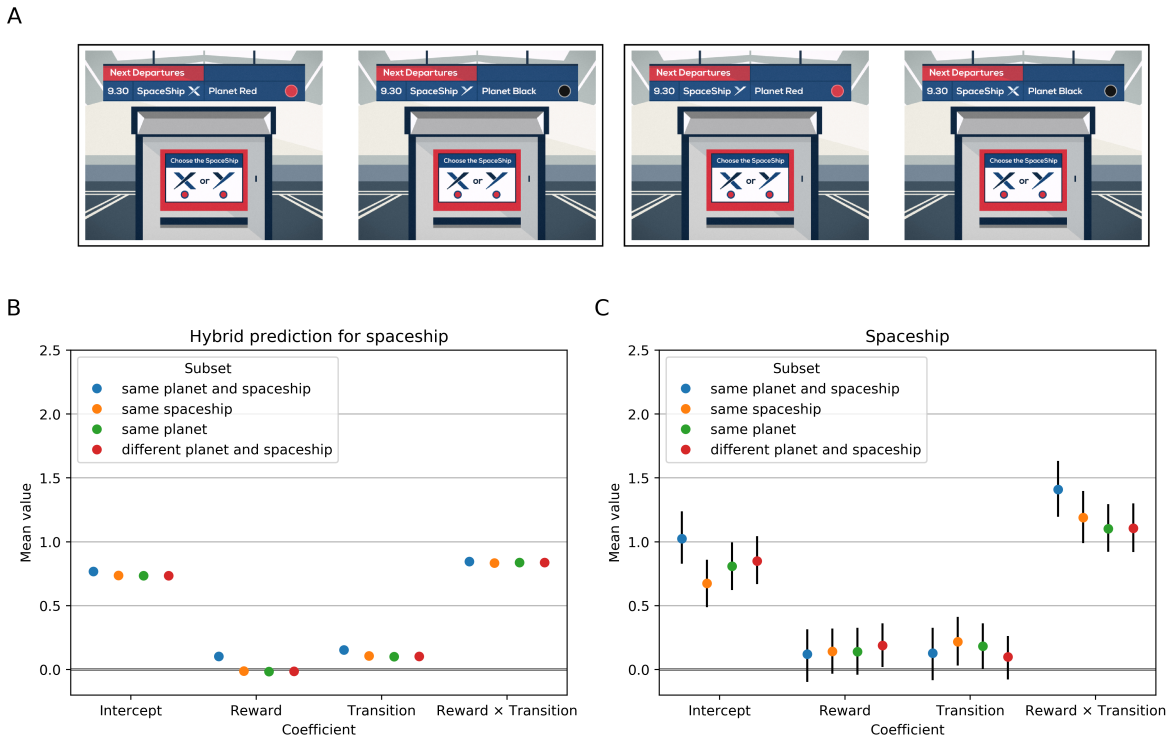

Figure 5: An example of a reward main effect that cannot be driven by model-free learning. A) The spaceship task contains the four possible flight announcements shown here. Each trial of the spaceship task began with a flight announcement on an information board above a ticket machine. The name of a spaceship was presented first for 2 seconds, then the name of the planet the spaceship was scheduled to fly to appeared to the right of the spaceship's name, which remained on the screen, for another 2 seconds. There are four possible flight announcements, because there are two spaceships (X and Y) and two planets (Red and Black). The participant must select a spaceship to fly on based on the announced flight. Each announcement listed only one spaceship's destination, but still provided complete information about the intended destination of all spaceships. This is because there is a constant rule governing the spaceships' flights—they always depart at the same time and fly to different planets. Thus, if one spaceship is listed as flying to Planet Red, then the other ship will fly to Planet Black and vice versa. This means that the two screens in the left rectangle of panel A are equivalent and depict the same initial state. The two screens in the right rectangle of panel A are also equivalent to one another. B) This plot shows the results for simulated hybrid agents ($N = 5000$) performing 1000 trials of the spaceship task under the standard assumption that model-free learning does not generalize between different state representations (i.e. different flight announcements). The hybrid model parameters used for this simulation were the median parameters obtained by fitting the hybrid model to the human spaceship data set by maximum likelihood estimation. The points on the plot represent coefficients from a logistic regression analysis on the simulated choices, with consecutive trial pairs divided into four subsets: (1, blue) same planet and spaceship, (2, orange) same spaceship, (3, green) same planet, and (4, red) different planet and spaceship. This division was made with regard to which flight is announced in current trial compared to the preceding trial. Error bars are not shown because they are very small due to the large number of data points. C) Logistic regression results for the human data from the spaceship task ($N = 21$), with consecutive trial pairs divided into four subsets: (1, blue) same planet and spaceship, (2, orange) same spaceship, (3, green) same planet, and (4, red) different planet and spaceship. In contrast to the simulated hybrid agents, the small reward effect in human behavior does not differ across changes in the initial state. Thus, the driving factor behind this reward effect is inconsistent with a standard model-free learning system. Error bars represent the 95% highest density intervals.

categories 2 and 3 together, which corresponded to trial pairs where a "stay" choice required that participants choose a different response (spaceship), the posterior probability that the reward effect is greater than zero is 98%. Thus, the reward effect observed in the spaceship task choices is roughly equal in every trial category, including those where either the stimuli presented to participants or the response required to implement the same choice as before are different between the two trials. Note that generalization between different flight announcements implies knowledge about the task structure, because the flight announcements define the transitions between first- and second-stage states. This strongly suggests that the observed reward effects are not model-free, but result instead from model-based behavior, except that the models participants use are not identical to the one assumed by the standard analyses. (See in the Supplementary Material further evidence of deviations from the correct model in the magic carpet and spaceship tasks.)

Discussion

We show that simple changes to the task instructions and practice trials led healthy adult humans to behave in a model-based manner during a commonly used two-stage decision task. This is in contrast to the majority of the literature on two-stage decision tasks, which suggest that decisions in these tasks are driven by a combination of model-based and model-free learning. However, we also show that if purely model-based agents use mental models that differ from the true model, which they may well do if they misunderstood the task, then the analysis can falsely indicate an influence of model-free learning. In other words, agents that are purely model-based can be mistakenly classified as hybrid model-free/model-based learners if they do not fully understand the environment (i.e. task). Therefore, our work here, together with other recent reports on the limitations of hybrid model fits [28, 26], indicates the need to carefully reconsider aspects of both the empirical tools and theoretical assumptions that are currently pervasive in the study of reward learning and decision making.

Behavior does not necessarily fall into the dichotomous modes of simple model-free and correct model-based learning assumed in the design and analysis of two-stage decision tasks [30, 31, 32, 33, 34, 35, 36, 37]. Instead, agents can form and act on a multitude of strategies from which the simple win-stay, lose-switch model-free strategy and the correct model of the task are only two possibilities (Figure 6). It is known that this multidimensional strategy space includes more complex model-free algorithms that can mimic model-based actions in some cases [31, 33]. Here, we show that it also includes model-based strategies that can appear to be partially model-free because they are inconsistent with the actual state of the environment or task rules (i.e. they are incorrect models). This mimicry of behavioral patterns between the infinite variations of model-free and model-based algorithms an agent might employ adds considerable uncertainty to any attempt to classify reward learning strategies based on behavior in a two-stage choice task.

Drawing the conclusion that behavior is a hybrid mix of two exemplar algorithms (e.g. simple model-free and correct model-based) is especially tenuous. In such a case, the observed pattern of choices does not match the predictions of either of the two algorithms. Arguably the most plausible reason that observed behavior does not match either of the two candidate algorithms is that participants are not using either of those algorithms. However, it is also possible that behavior does not strictly follow either algorithm because both reward learning algorithms are operating in parallel and jointly influence decisions. Indeed, to date, the most common conclusion has been that there is a joint influence of model-free and model-based learning in two-stage decision tasks. The popularity of this conclusion may stem from strong prior beliefs in the existence of dual systems for learning. Regardless of the reason for its popularity, this conclusion relies on the assumption that participants have an accurate mental model of the task that they could use to guide their behavior. We argue that this assumption of understanding may not hold in many cases.

In line with the possibility that apparently hybrid behavior is, instead, driven by misconceptions of the task, we found drastic shifts toward (correct) model-based behavior in healthy adult humans when we made seemingly small changes to the task instructions. It comes as no surprise that people do not always understand and/or follow instructions well. However, the extent to which people apparently misconstrued how the original two-stage task works and the impact of these misunderstandings on our ability to make accurate inferences about reward learning was unexpected for us. In both our local samples and when re-analyzing open data shared by other groups, we found strong indications that people misunderstood and/or incorrectly modeled the two-stage task in important ways. Among

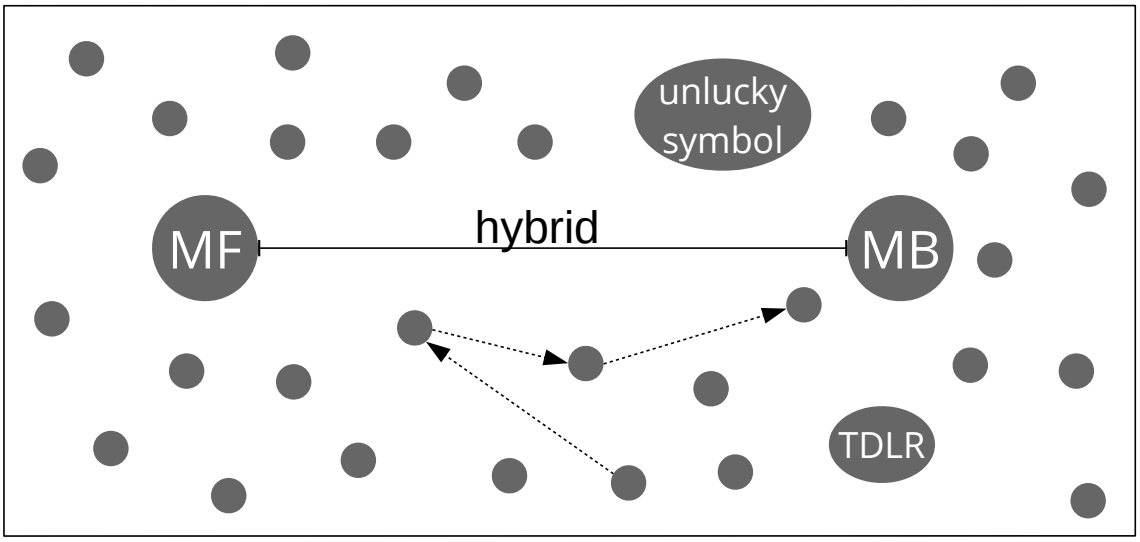

Figure 6: A simplified two-dimensional diagram representing the strategy space in the two-stage task. The space of all strategies agents can employ to perform the two-stage task contains the model-free (MF) and correct model-based (MB) strategies, as well as intermediate hybrid strategies on the line between them. However, it also contains other purely model-based strategies, such as the transition-dependent learning rates (TDLR) and unlucky symbol strategies, among countless others. If these incorrect model-based strategies are projected onto the line segment between model-free and correct model-based behavior, they will yield the false conclusion that the participants are using a hybrid MF/MB strategy. This projection of behavior onto the one-dimensional space along the line between MF and MB is essentially what the standard analyses of the two-stage task do. However, in reality, a human participant who misunderstands the two-stage task may use any of the strategies that we depict here in two dimensions only for simplicity. In reality, the potential strategies exist in a multidimensional space. Importantly, if people misunderstand the task in different ways, then they will use different incorrect models that correspond to their current understanding of the task. Moreover, people may well change strategies over the course of the experiment (represented by the dashed arrows) if they determine that their understanding of the task was wrong in some way. Indeed, there is evidence that extensive training may lead participants to learn and utilize the correct model of the two-stage task over time [29]. Unless we can ensure a complete understanding of the correct model-based strategy a priori, the vast space of possible incorrect models, heterogeneity across participants, and potential for participants to change models over time make accurate identification of a hybrid mixture of model-free and model-based learning a daunting, if not impossible, task.

these indications are the location effects detected in the common instructions data set [21]. The participants in this data set were recruited on Amazon Mechanical Turk, but similar behavior has been recently reported by Shahar et al. for participants who performed the two-stage task in a laboratory [38]. Shahar et al. proposed that location effects are due to model-free credit assignment to outcome-irrelevant task features, such as stimulus locations and specific key responses [38]. However, this proposal is inconsistent with the results of the magic carpet and spaceship tasks, which show a reduction in location effects under our improved instructions. Moreover, the spaceship task's results, if interpreted according to the model-free/model-based framework, would lead to the conclusion that model-free learning is not only insensitive to outcome-irrelevant features, but also able to generalize between distinct outcome-relevant features (i.e. flight announcements). Clearly, the effects of symbol location and identity on behavior vary across tasks depending on the instruction. This fact argues against such effects being model-free and suggests instead that they are likely due to model-based confusion about the relevance of different task features. The reduction in location effects in our tasks also argues against another alternative explanation that our instructions, rather than alleviating confusion, emphasized the task's transition structure, and thus simply encouraged participants to be more model-based. While it remains possible that our instructions also emphasize and/or encourage model-based behavior, that is not the whole story. The nearly complete elimination of location effects and side switching mistakes is direct evidence that our instructions for the spaceship and magic carpet tasks cleared up participants' confusion and incorrect modeling of the task.

As discussed above, our simulations of TDLR and unlucky symbol agents provide a concrete demonstration that choices based on an incorrect model of the task can falsely indicate an influence of model-free learning. The main effect of reward in our spaceship task is empirical evidence that healthy adult humans may indeed employ one such misleading incorrect model. Overall, behavior on the spaceship task was almost completely (correct) model-based. However, there was a small main effect of reward when computing a logistic regression analysis on consecutive trials with both the same or different initial states. The main effect of reward is assumed to be evidence of model-free learning [7], but model-free learning is also assumed to be unable to generalize between distinct but equivalent state representations [15, 16, 21, 31]. If the small reward main effect in the spaceship task was driven by model-free learning, then model-free learning would have to able to generalize not only between distinct states but also between distinct actions. Therefore, the observed reward effect was more likely generated by model-based learning acting on a model of the task that does not precisely match the true underlying task model instead of being a model-free learning effect. We emphasize again that we don't think the TDLR or unlucky symbol models are the exact models that participants used. They are simply two plausible examples from the infinite set of incorrect models participants could have formed and acted on. The fact that there is an unlimited set of incorrect models and that some models used by human participants seem to mimic a hybrid model-based plus model-free agent is the core obstacle to drawing accurate conclusions from behavior in the two-stage task.

The critical issue is that participants' use of incorrect mental models in the original two-stage task not only produces noisy behavior, but can lead to false indications of model-free behavior. As noted above, we have demonstrated that simulated choices from purely model-based agents acting on incorrect mental models of the task appear to be driven by a combination of model-based and model-free influences (i.e. hybrid) when the data are analyzed using any of the currently employed methods (logistic regressions, hybrid model fits). We also found that in human choice data there is a small, but significant association between the ability of the hybrid reinforcement learning model to explain a participant's choice pattern (i.e. the log-likelihood of the model) and the value of the model-based weight parameter. Specifically, when the hybrid model doesn't explain human choices well, and/or estimates that a participant's choices are noisy it is biased toward indicating that there is more model-free influence on behavior. These findings are consistent with previous reports that a measure of understanding, the number of attempts required to pass a four-question comprehension quiz, is related to model-based weight parameters [6]. In that study, worse, or at least slower achievement of, understanding was associated with more model-free influence as well. Overall, the data show that in the absence of sufficient understanding of the task, which would allow an agent to use the correct mental model, the standard analysis methods for two-stage task choices overestimate model-free influences on behavior. However, we should not forget that despite the limitations in our ability to accurately measure its influence on two-stage task choices, humans and other animals may use model-free learning algorithms in some cases.

Model-free reinforcement learning can readily explain the results of many animal electrophysiology and human neuroimaging studies. In particular, a highly influential finding in neuroscience is that dopamine neurons respond to rewards in a manner consistent with signalling the reward prediction errors that are fundamental to temporal difference learning algorithms, a form of model-free reward learning [39, 40, 41]. However, in the studies showing this response, there was no higher-order structure to the task—often, there was no instrumental task at all. Thus, it was not possible to use more complex or model-based learning algorithms in those cases. Conversely, when the task is more complicated, it has been found that the dopamine prediction error signal reflects knowledge of task structure and is therefore not a model-free signal (for example, [42, 43]). A recent optogenetic study showed that dopamine signals are both necessary and sufficient for model-based learning in rats [44], and consistent with this finding, neuroimaging studies in humans found that BOLD signals in striatal and prefrontal regions that receive strong dopaminergic innervation correlate with model-based prediction error signals [7, 45]. Moreover, although there is evidence that anatomically distinct striatal systems mediate goal-directed and habitual actions [46], to date there is no evidence for anatomically separate representations of model-free and model-based learning algorithms.

Model-free learning algorithms are generally assumed to be the computational analogs of habits, but they are not necessarily the same thing [1]. Initial theoretical work proposed the model-based versus model-free distinction to formalize the dual-system distinction between goal-directed and habitual control [4]. However, model-free learning has never been empirically shown to equate with or even exclusively lead to habitual behavior. Indeed, it is generally assumed that goal-directed actions can be based on model-free learning too. Consider a participant who is purely goal-directed but does not understand how the state transitions work in a two-stage task. This participant may resort to employing a simple win-stay, lose-shift strategy, which is model-free, but his behavior will not be habitual.

Apparently model-free participants behave in ways inconsistent with the habitual tendencies that model-free learning is supposed to index. A study by Konovalov and Krajbich [47] combined eye-tracking with two-stage task choices to examine fixation patterns as a function of estimated learning type. In addition to those authors' primary conclusions, we think this work highlights inequalities between seemingly model-free behavior, and what one would expect from a habitual agent. Their analysis strategy divided participants into model-free and model-based learners, based on a median ($w = 0.3$) split of the model-based weight parameter estimated from the hybrid reward learning algorithm. They reported that when the first-stage symbols were presented, model-based learners tended to look at most once at each symbol, as if they had decided prior to trial onset which symbol they were going to choose. In contrast, learners classified as model-free tended to make more fixations back and forth between first-stage symbols, and their choices were more closely related to fixation duration than those of the model-based group. We interpret this pattern of fixation and choices as suggesting that model-free participants made goal-directed comparisons when the first-stage symbols were presented, rather than habitual responses. This is because similar patterns of back and forth head movements, presumably analogous to fixations, are seen when rats are initially learning to navigate a maze [48]. Furthermore, the rats' head movements are accompanied hippocampal representations of reward locations in the direction the animal is facing. Such behavior is seen as evidence that the animals are deliberating over choices in a goal-directed fashion. Humans also make more fixations per trial as trial difficulty increases in goal-directed choice paradigms [49]. Notably, these patterns of head movements and hippocampal place cell signaling actually cease once animals have extensive training on the maze and act in an automated or habitual fashion at each decision point [48]. Thus, supposedly model-free human participants' fixation patterns during the two-stage task suggest that they are acting in a goal-directed rather than a habit-like fashion.

In contrast to habits, model-free behavior decreases with extensive training on the two-stage task. In general, the frequency and strength of habitual actions increase with experience in a given task or environment. However, Economides et al. showed that the estimated amount of model-free influence in human participants decreases over three days of training on the two-stage task [29]. They also found that, after two days of training, human behavior remains primarily model-based in the face of interference from a second task (the Stroop task) performed in parallel. Both of these results raise questions about the relative effortfulness of model-based versus model-free learning in the two-stage task. After all, although it is apparently hard to explain, the transition model behind the two-stage task is rather easy to follow once it is understood. Rats also show primarily model-based behavior

after receiving extensive training on the two-stage task [23]. In fact, the rats showed little or no influence of model-free learning. Moreover, Miller et al. [23] also showed that inactivation of the dorsal hippocampus or orbitofrontal cortex impairs model-based planning, but does not increase the influence of model-free learning. Instead, any influence of model-free learning in the rats remained negligible. As a whole, these results are difficult to reconcile with the idea of an ongoing competition or arbitration between model-based and model-free control over behavior.

Humans have been reported to arbitrate between model-based and model-free strategies based on both their relative accuracy and effort. We know from several lines of evidence that both humans and other animals are sensitive to and generally seek to minimize both physical and mental effort if possible [50]. Model-based learning is often thought to require more effort than model-free learning. A well-known aspect of the original two-stage task [7] is that model-based learning does not lead to greater accuracy or monetary payoffs compared to model-free learning [31, 21]. Thus, it has been hypothesized that an aversion to mental effort coupled with lack of monetary benefits from model-based learning may lead participants to use a partially model-free strategy on the original two-stage task [31, 21]. Previous studies have tested this hypothesis by modifying the original two-stage task so that model-based learning strategies do achieve significantly greater accuracy and more rewards [21, 22, 24]. They found that participants appeared more model-based behavior when it paid off to use a model-based strategy. The conclusion in those studies was that participants will employ model-based learning if it is advantageous in a cost-benefit analysis between effort and money.

Our results and those from studies with extensive training [51, 23] cannot be explained by cost-benefit or accuracy trade-offs between model-free and model-based learning. The magic carpet and spaceship tasks led to almost completely model-based behavior, but had the same equivalency in terms of profit for model-based and model-free learning as in the original two-stage task [7]. The objective effort in the magic carpet task was also equivalent to the original two-stage task. Although an interesting possibility that merits further study is that giving concrete causes for rare transitions also reduced the subjective effort of forming or using the correct mental model of the task. Similarly, the profitability of model-based learning does not change with experience. If anything, more experience with the task should allow the agent to learn that the model-based strategy is no better than the model-free if both are being computed and evaluated in parallel for a long period of time. Therefore, these two sets of results cannot be explained by an increase in model-based accuracy and profits compared to model-free learning.

Seemingly model-free behavior may be reduced in all three sets of experiments through better understanding of the task. Clearly, improved instructions and more experience can give participants a better understanding of the correct task model. Most, if not all, of the modified two-stage tasks also have the potential to facilitate understanding as well as making model-based learning more profitable. This is because in addition to generating higher profits, the differential payoffs also provide clearer feedback to participants about the correctness of their mental models. If both correct and incorrect models lead to the same average payoffs, participants may be slow to realize their models are incorrect. Conversely, if the correct model provides a clear payoff advantage over other models, participants will be guided to quickly change their mental models through feedback from the task. Of course, increased understanding and changes in the cost-benefit ratio may jointly drive the increases in (correct) model-based behavior in modified two-stage tasks. Additional data are needed to carefully tease apart these two potential reasons for increased model-based behavior in many newer versions of the two-stage task.

Two-stage tasks have also been used to test for links between compulsive behavior and model-free learning in healthy and clinical populations. Compulsive symptoms have been found to correlate with apparent model-free behavior in the two-stage task [6, 14, 52]. Given our current results, however, the conclusion that model-free learning and compulsive behaviors are linked should be drawn with caution. We have shown that it is not clear what exactly is being measured by the two-stage task in healthy young adult humans. The same questions should be extended to other populations, including those with obsessive compulsive disorder (OCD). In the case of OCD, is the two-stage task picking up alterations in how distinct habitual (indexed by model-free learning) and goal-directed (model-based learning) systems interact to control behavior? Or are differences in two-stage choices driven by the ability to understand a task, create and maintain accurate mental models of it, and use these models to make decisions? It is certainly possible that OCD patients and other individuals with sub-clinical compulsive symptoms do indeed use more model-free learning during the two-stage task. However, a plausible alternative explanation for the correlations with model-free indexes in the two-stage task is

that compulsive individuals form and continue to follow inaccurate models of the world. Patients with OCD have been found to be impaired in causal reasoning between actions and outcomes [53], deficits which would likely impair the ability to form accurate task models. In fact, individuals with OCD have been reported to be impaired in multiple measures of cognitive flexibility, but the fundamental reasons for these impairments remain unclear [54, 55]. Our current findings do not change the fact that behavior on the two-stage task is correlated with compulsive symptoms, but they do indicate a need to continue investigating the underlying cause for these correlations.

Reward learning is one of the most central and widely studied processes in the neural and behavioral sciences. Given that learning and decision processes are key drivers of behavior, it is important for researchers to have and use tools that can ascertain their mechanistic properties, their potential neural substrates, and how their dysfunction might lead to various forms of sub-optimal behavior or psychopathology. The specification of algorithmic formulae for model-free versus model-based learning has advanced the study of reward learning in many important conceptual and practical ways. However, as Nathaniel Daw recently noted, "such clean dichotomies are bound to be oversimplified. In formalizing them, the [model-based]-versus-[model-free] distinction has also offered a firmer foundation for what will ultimately be, in a way, its own undoing: getting beyond the binary" [30]. We believe that our current results are another strong indication that the time to move beyond oversimplified binary frameworks is now.

Methods

All the code used to perform the simulations, run the magic carpet and the spaceship tasks, and analyze the results, as well as the data obtained from human participants, are available at `https://github.com/carolfs/muddled_models`

Simulations of model-based agents

The model-based agents described in the Results section were simulated and their decisions analyzed by reinforcement learning model fitting. 1 000 agents of each type (original hybrid, unlucky symbol, and transition-dependent learning rates) performed a two-stage task with 1 000 trials, and the raw data were used to plot the stay probabilities depending on the previous trial's outcome and reward. The hybrid reinforcement learning model proposed by Daw et al. [7] was fitted to the data from each agent by maximum likelihood estimation. To this end, the optimization algorithm LBFGS, available in the PyStan library [56], was run 10 times with random seeds and for 5000 iterations to obtain the best set of model parameters for each agent. The three types of model-based agents we simulated are described below.

The hybrid algorithm

Daw et al. [7] proposed a hybrid reinforcement learning model, combining the model-free SARSA(λ) algorithm with model-based learning, to analyze the results of the two-stage task.

Initially, at time $t = 1$, the model-free values $Q_1^{MF}(s, a)$ of each action a that can be performed at each state s are set to zero. At the end of each trial t, the model-free values of the chosen actions are updated. For the chosen second-stage action a_2 performed at second-stage state s_2 (the pink or blue states in Fig. 1A), the model-free value is updated depending on the reward prediction error, defined as $\delta_t^2 = r_t - Q_t^{MF}(s_2, a_2)$, the difference between the chosen action's current value and the received reward r_t. The update is performed as

$$Q_{t+1}^{MF}(s_2, a_2) = Q_t^{MF}(s_2, a_2) + \alpha_2[r_t - Q_t^{MF}(s_2, a_2)], \tag{1}$$

where α_2 is the second-stage learning rate ($0 \leq \alpha_2 \leq 1$). For the chosen first-stage action a_1 performed at the first-stage state s_1, the value is updated depending on the reward prediction error at the first and second stages, as follows:

$$Q_{t+1}^{MF}(s_1, a_1) = Q_t^{MF}(s_1, a_1) + \alpha_1[Q_t^{MF}(s_2, a_2) - Q_t^{MF}(s_1, a_1)] + \alpha_1\lambda[r_t - Q_t^{MF}(s_2, a_2)], \tag{2}$$

where α_1 is the first-stage learning rate ($0 \leq \alpha_1 \leq 1$), $\delta_t^1 = Q_t^{MF}(s_2, a_2) - Q_t^{MF}(s_1, a_1)$ is the reward prediction error at the first stage, and λ is the so-called eligibility parameter ($0 \leq \lambda \leq 1$), which modulates the effect of the second-stage reward prediction error on the values of first-stage actions.

The model-based value $Q_t^{MB}(s_2, a_2)$ of each action a_2 performed at second-stage state s_2 is the same as the corresponding model-free value, i.e., $Q_t^{MB}(s_2, a_2) = Q_t^{MF}(s_2, a_2)$. The model-based value of each first-stage action a_1 is calculated at the time of decision making from the values of second-stage actions as follows:

$$Q_t^{MB}(s_1, a_1) = \sum_{s_2 \in \mathcal{S}} P(s_2|s_1, a_1) \max_{a_2 \in \mathcal{A}} Q_t^{MB}(s_2, a_2), \tag{3}$$

where $P(s_2|s_1, a_1)$ is the probability of transitioning to second-stage state s_2 by performing action a_1 at first-stage state s_1, $\mathcal{S} = \{pink, blue\}$ is the set of second-stage states, and $\mathcal{A}$ is the set containing the actions available at that state.

The agent makes first-stage choices based on both the model-free and the model-based state-action pairs, weighted by a model-based weight w ($0 \leq w \leq 1$), according to a soft-max distribution:

$$P_t(s_1, a_1) = \frac{e^{\beta_1[wQ_t^{MB}(s_1,a_1)+(1-w)Q_t^{MF}(s_1,a_1)+p \cdot rep_t(a_1)]}}{\sum_{a' \in \mathcal{A}} e^{\beta_1[wQ_t^{MB}(s_1,a')+(1-w)Q_t^{MF}(s_1,a')+p \cdot rep_t(a')]}}, \tag{4}$$

where β_1 is the first-stage's inverse temperature parameter, which determines the exploration-exploitation trade-off at this stage, p is a perseveration parameter that models a propensity for repeating the previous trial's first-stage action in the next trial, and $rep_t(a') = 1$ if the agent performed the first-stage action a' in the previous trial, and zero otherwise. Kool et al. [21] have added an additional parameter to the hybrid model—the response stickiness ρ—and the above equation becomes

$$P_t(s_1, a_1) = \frac{e^{\beta_1[wQ_t^{MB}(s_1,a_1)+(1-w)Q_t^{MF}(s_1,a_1)+p \cdot rep_t(a_1)+\rho \cdot resp_t(a_1)]}}{\sum_{a' \in \mathcal{A}} e^{\beta_1[wQ_t^{MB}(s_1,a')+(1-w)Q_t^{MF}(s_1,a')+p \cdot rep_t(a')+\rho \cdot resp_t(a')]}}, \tag{5}$$

where the variable $resp_t(a')$ is equal to 1 if a' is the first-stage action performed by pressing the same key as in the previous trial, and zero otherwise.

Choices at the second stage are simpler, as the model-free and model-based values of second-stage actions are the same and there is no assumed tendency to repeat the previous action or key press. Second-stage choice probabilities are given as follows:

$$P_t(s_2, a_2) = \frac{e^{\beta_2 Q_t(s_2,a_2)}}{\sum_{a' \in \mathcal{A}} e^{\beta_2 Q_t(s_2,a')}}. \tag{6}$$

We propose two alternative algorithms below to demonstrate that model-based agents may be mistakenly classified as hybrid agents. These algorithms are based on the algorithm by Daw et al. [7] detailed above, except that the inverse temperature parameter is the same for both stages (for simplicity because these models are only intended as demonstrations), the perseveration parameter p is equal to 0 (again, for simplicity), and the model-based weight w is equal to 1, indicating a purely model-based strategy.

The unlucky-symbol algorithm

We simulated an agent that believes a certain first-stage symbol is unlucky and lowers the values of second-stage actions by 50%. This model-based algorithm has three parameters: $0 \leq \alpha \leq 1$, the learning rate, $\beta > 0$, an inverse temperature parameter for both stages (for simplicity), and $0 < \eta < 1$, a reduction of second-stage action values caused by choosing the unlucky symbol. The value of each first-stage action a_1 is calculated from the values of second-stage actions as follows:

$$Q_t(s_1, a_1) = \sum_{s_2 \in \mathcal{S}} P(s_2|s_1, a_1) \max_{a_2 \in \mathcal{A}} Q_t(s_2, a_2), \tag{7}$$

The probability of choosing a first-stage action is given by:

$$P_t(s_1, a_1) = \frac{e^{unlucky(a_1)\beta Q_t(s_1,a_1)}}{\sum_{a' \in \mathcal{A}} e^{unlucky(a')\beta Q_t(s_1,a')}}, \tag{8}$$

where $unlucky(a) = \eta$ if the agent thinks action a is unlucky and $unlucky(a) = 1$ otherwise. Second-stage value updates and second-stage choices are made as described above for the original hybrid model. The probability of choosing a second-stage action is given by

$$P_t(s_2, a_2) = \frac{e^{unlucky(a_1)\beta Q(s_2,a_2)}}{\sum_{a' \in \mathcal{A}} e^{unlucky(a_1)\beta Q_t(s_2,a_2)}}. \tag{9}$$

Learning of second-stage action values occurs as in the original hybrid model.

The transition-dependent learning rates (TDLR) algorithm

This is a simple model-based learning algorithm that has a higher learning rate after a common transition and a lower learning rate after a rare transition; hence, the learning rates are transition-dependent. This model-based TDLR algorithm has three parameters: α_c, the higher learning rate for outcomes observed after common transitions ($0 \leq \alpha_c \leq 1$), α_r, the lower learning rate for outcomes observed after rare transitions ($0 \leq \alpha_r < \alpha_c$), and $\beta > 0$, an inverse temperature parameter that determines the exploration-exploitation trade-off. In each trial t, based on the trial's observed outcome ($r_t = 1$ if the trial was rewarded, $r_t = 0$ otherwise), the algorithm updates the estimated value $Q_t(s_2, a_2)$ of the chosen second-stage action a_2 performed at second-stage state s_2 (pink or blue). This update occurs according to the following equation:

$$Q_{t+1}(s_2, a_2) = Q_t(s_2, a_2) + \alpha[r_t - Q_t(s_2, a_2)], \tag{10}$$

where $\alpha = \alpha_c$ if the transition was common and $\alpha = \alpha_r$ if the transition was rare. The value of each first-stage action a_1 is calculated from the values of second-stage actions as follows:

$$Q_t(s_1, a_1) = \sum_{s_2 \in \mathcal{S}} P(s_2|s_1, a_1) \max_{a_2 \in \mathcal{A}} Q_t(s_2, a_2), \tag{11}$$

where $P(s_2|s_1, a_1)$ is the probability of transitioning to second-stage state s_2 by performing action a_1 at first-stage s_1, $\mathcal{S}$ is the set of second-stage states, and $\mathcal{A}$ is the set of all second-stage actions. Choices made at first- or second-stage states are probabilistic with a soft-max distribution:

$$P_t(s, a) = \frac{e^{\beta Q_t(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\beta Q_t(s,a')}}. \tag{12}$$

When this model was fitted to human participant data, five parameters were used instead: α_c, the learning rate for outcomes observed after common transitions, α_r, the learning rate for outcomes observed after rare transitions, β_1, the first-stage's inverse temperature parameter, β_2, the second-stage's inverse temperature parameter, and p, the perseveration parameter.

Simulation parameters

We simulated 1 000 purely model-based agents performing the two-stage task using each of the algorithms described above: (1) the original hybrid algorithm using a model-based weight $w = 1$ and $\alpha_1 = \alpha_2 = 0.5$, (2) the unlucky-symbol algorithm with $\alpha = 0.5$ and $\eta = 0.5$, and (3) the TDLR algorithm with $\alpha_c = 0.8$ and $\alpha_r = 0.2$. For all agents, the β parameters had a value of 5.

Analysis of the common instructions data

In [21], 206 participants recruited via Amazon Mechanical Turk performed the two-stage task for 125 trials. See [21] for further details. The behavioral data were downloaded from the first author's Github repository (https://github.com/wkool/tradeoffs) and reanalyzed by logistic regression and reinforcement learning model fitting, as described below.

Logistic regression of consecutive trials

This analysis was applied to all the analyzed behavioral data sets. Consecutive trial pairs were divided into subsets, depending on the presentation of first-stage stimuli. The results for each subset were then

separately analyzed, using a hierarchical logistic regression model whose parameters were estimated through Bayesian computational methods. The predicted variable was p_{stay}, the stay probability for a given trial, and the predictors were x_r, which indicated whether a reward was received or not in the previous trial (+1 if the previous trial was rewarded, −1 otherwise), x_t, which indicated whether the transition in the previous trial was common or rare (+1 if it was common, −1 if it was rare), the interaction between the two. Thus, for each condition, an intercept β_0 for each participant and three fixed coefficients were determined, as shown in the following equation:

$$p_{\text{stay}} = \frac{1}{1 + \exp[-(\beta_0 + \beta_r x_r + \beta_t x_t + \beta_{r\times t} x_r x_t)]}. \tag{13}$$

The distribution of y was Bernoulli(p_{stay}). The distribution of the $\vec{\beta}$ vectors was $\mathcal{N}(\vec{\mu}, \vec{\sigma^2})$. The parameters of the $\vec{\beta}$ distribution were given vague prior distributions based on preliminary analyses—the $\vec{\mu}$ vectors' components were given a $\mathcal{N}(\mu = 0, \sigma^2 = 25)$ prior, and the $\vec{\sigma^2}$ vector's components were given a Cauchy(0, 1) prior. Other vague prior distributions for the model parameters were tested and the results did not change significantly.

To obtain parameter estimates from the model's posterior distribution, we coded the model into the Stan modeling language [57, 58] and used the PyStan Python package [56] to obtain 60 000 samples of the joint posterior distribution from four chains of length 30 000 (warmup 15 000). Convergence of the chains was indicated by $\hat{R} \approx 1.0$ for all parameters.

Fitting of hybrid reinforcement learning models

The hybrid reinforcement learning model proposed by Daw et al. [7] was fitted to all data sets (common instructions, magic carpet, and spaceship). To that end, we used a Bayesian hierarchical model, which allowed us to pool data from all participants to improve individual parameter estimates. For the analysis of the spaceship data, four distinct first-stage states were assumed, corresponding to the four possible flight announcements (Figure 5A).

The parameters of the hybrid model for the ith participant were α_1^i, α_2^i, λ^i, β_1^i, β_2^i, w^i, and p^i. Vectors $(\text{logit}(\alpha_1^i), \text{logit}(\alpha_2^i), \text{logit}(\lambda^i), \log(\beta_1^i), \log(\beta_2^i), \text{logit}(w^i), p^i)$, obtained for each participant, were given a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. These transformation of the parameters were used because the original values were constrained to an interval and the transformed ones were not, which the normal distribution requires. The model's hyperparameters were given weakly informative prior distributions. Each component of $\boldsymbol{\mu}$ was given a normal prior distribution with mean 0 and variance 5, and $\boldsymbol{\Sigma}$ was decomposed into a diagonal matrix $\boldsymbol{\tau}$, whose diagonal components were given a Cauchy prior distribution with mean 0 and variance 1, and a correlation matrix $\boldsymbol{\Omega}$, which was given an LKJ prior [59] with shape $\nu = 2$ [58]. This model was coded in the Stan modelling language [58, 57] and fitted to each data set using the PyStan interface [56] to obtain a chain of 40 000 iterations (warmup: 20 000) for the common instructions data set and 80 000 iterations (warmup: 40 000) for the magic carpet and spaceship data sets. Convergence was indicated by $\hat{R} \leq 1.1$ for all parameters.

The same procedure above was performed to fit a hybrid model with a mistake probability to the common instructions and magic carpet data sets. An additional parameter was added to the original hybrid reinforcement learning model: ρ^i, the probability of making a mistake and making the wrong choice when the first-stage symbols switched sides from one trial to the next. Precisely, in trials with the first-stage symbols on different sides compared with the previous trials, let $P_t^h(s_1, a_1)$ be the probability, according to the standard hybrid model, that the participant selected action a_1 at the first-stage s_i in trial t. The same probability according the hybrid model with a mistake probability was given by

$$P_t(s_1, a_1) = (1 - \rho)P_t^h(s_1, a_1) + \rho(1 - P_t^h(s_1, a_1)). \tag{14}$$

This model also assumes that the participant realized their mistake after making one and that action values were updated correctly. Data from each participant were described by a vector $(\text{logit}(\alpha_1^i), \text{logit}(\alpha_2^i), \text{logit}(\lambda^i), \log(\beta_1^i),$

The computed log-likelihoods obtained for each participant and model at each iteration were used to calculate the PSIS-LOO score (an approximation of leave-one-out cross-validation) of each model for each participant. To this end, the loo and compare functions of the loo R package were employed [60].

The standard hybrid model was fit to all data sets by maximum likelihood estimation. The model was coded in the Stan modelling language [58, 57] and fitted 1000 times (for robustness) to each participant's choices using LBFGS algorithm implemented by Stan through the PyStan interface [56].

The magic carpet and spaceship tasks

24 healthy participants participated in the magic carpet experiment and 21 in the spaceship experiment. In both cases, participants were recruited from the University of Zurich's Registration Center for Study Participants. The inclusion criterion was speaking English, and no participants were excluded from the analysis. The sample sizes were based on our previous pilot studies using the two-stage task [25]. The experiment was conducted in accordance with the Zurich Cantonal Ethics Commission's norms for conducting research with human participants, and all participants gave written informed consent.

Participants first read the instructions for the practice trials and completed a short quiz on these instructions. For the magic carpet and spaceship tasks, 50 and 20 practice trials were performed respectively. Next, participants read the instructions for the main task, which was then performed. For the magic carpet task, they performed 201 trials and for the spaceship task, 250 trials. This number of trials excludes slow trials. Trials were divided into three blocks of roughly equal length. For every rewarded trial in the magic carpet or spaceship task, participants were paid CHF 0.37 or CHF 0.29 respectively. The total payment was displayed on the screen and the participants were asked to fill in a short questionnaire. For the magic carpet task, the questionnaire contained the following questions:

1. For each first-stage symbol, "What was the meaning of the symbol below?"
2. "How difficult was the game?" with possible responses being "very easy," "easy," "average," "difficult," and "very difficult."
3. "Please describe in detail the strategy you used to make your choices."

For the spaceship task, participants were only asked about their strategy. The questionnaire data are available in our Github repository along with all the code and the remaining participant data.

Magic carpet task description

Our magic carpet version of the two-stage task was framed as follows. Participants were told that they would be playing the role of a musician living in a fantasy land. The musician played the flute for gold coins to an audience of genies, who lived inside magic lamps on Pink Mountain and Blue Mountain. Two genies lived on each mountain. Participants were told that the symbol written on each genie's lamp (a Tibetan character, see Figure 1C) was the genie's name in the local language. When the participants were on a mountain, they could pick up a lamp and rub it. If the genie was in the mood for music, he would come out of his lamp, listen to a song, and give the musician a gold coin. Each genie's interest in music could change with time. The participants were told that the lamps on each mountain might switch sides between visits to a mountain, because every time they picked up a lamp to rub it, they might put it down later in a different place.

To go to the mountains, the participant chose one of two magic carpets (Figure 1C). They had purchased the carpets from a magician, who enchanted each of them to fly to a different mountain. The symbols (Tibetan characters) written on the carpets meant "Blue Mountain" and "Pink Mountain" in the local language. A carpet would generally fly to the mountain whose name was written on it, but on rare occasions a strong wind blowing from that mountain would make flying there too dangerous because the wind might blow the musician off the carpet. In this case, the carpet would be forced to land instead on the other mountain. The participants were also told that the carpets might switch sides from one trial to the next, because as they took their two carpets out of the cupboard, they might put them down and unroll them in different sides of the room. The participants first did 50 "tutorial flights," during which they were told the meaning of each symbol on the carpets, i.e., they knew which transition was common and which was rare. Also, during the tutorial flights, the participants saw a transition screen (Figure 1D), which showed the carpet heading straight toward a mountain (common transition) or being blown by the wind in the direction of the other mountain (rare transition). During the task, however, they were told their magic carpets had been upgraded to be entirely self-driving. Rather then drive the carpet, the musician would instead take a nap aboard it and would only wake

up when the carpet arrived on a mountain. During this period a black screen was displayed. Thus, participants would have to figure out the meaning of each symbol on the carpets for themselves. The screens and the time intervals were designed to match the original abstract task [7], except for the black "nap" screen displayed during the transition, which added one extra second to every trial.

Spaceship task description

We also designed a second task, which we call the spaceship task, that differed from the original task reported in Daw et al. [7] in terms of how the first stage options were represented. Specifically, there were four configurations for the first-stage screen rather than two. These configurations were represented as four different flight announcements displayed on a spaceport information board (Figure 5A). The design of task was based on the common assumption that model-free learning is unable to generalize between distinct state representations [15, 16, 21]. It has also been argued that reversals of the transition matrix should increase the efficacy of model-based compared to model-free learning [31]. This type of reversal could happen between each trial in the spaceship task depending on which flight was announced. Thus, there are two reasons to expect that participants completing the spaceship task may be more model-based compared to the magic carpet task.

The spaceship task instructions stated that the participant would play the role of a space explorer searching for crystals on alien planets. These crystals possessed healing power and could be later sold in the intergalactic market for profit. The crystals can be found inside obelisks that were left on the surfaces of planets Red and Black by an ancient alien civilization. The obelisks grew crystals like oysters grow pearls, and the crystals grew at different speeds depending on the radiation levels at the obelisk's location. There were two obelisks on each planet, the left and the right obelisk, and they did not switch sides from trial to trial. To go to planet Red or Black, the participant would use the left or the right arrow key to buy a ticket on a ticket machine that would reserve them a seat on spaceship X or Y. The buttons to buy tickets on spaceships X and Y were always the same. A board above the ticket machine announced the next flight, for example, "Spaceship Y will fly to planet Black." Participants were told that the two spaceships were always scheduled to fly to different planets, that is, if spaceship Y was scheduled to fly to planet Black, that meant spaceship X was scheduled to fly to planet Red. Thus, if the announcement board displayed "Spaceship Y" and "Planet Black," but they wanted to go to planet Red, they should book a seat on spaceship X.

After buying the ticket, the participant observed the spaceship flying to its destination. The participant was able to see that the spaceship would usually reach the planet it was scheduled to fly to, but in about one flight out of three the spaceship's path to the target planet would be blocked by an asteroid cloud that appeared unpredictably, and the spaceship would be forced to do an emergency landing on the other planet. The precise transition probabilities were 0.7 for the common transition and 0.3 for the rare transition (Figure 1B). This transition screen was displayed during both the practice trials and the task trials (Figure 1D), and it explained to the participants why the spaceship would commonly land on the intended planet but in rare cases land on the other instead.

Thus, other than the four flight announcements, the spaceship task differed from the original two-stage task in that (1) the first-stage choices were labelled X and Y and were always displayed on the same sides, (2) for each choice, the participants were told which transition was common and which was rare, as well as the transition probabilities, (3) the participants saw a transition screen that showed if a trial's transition was common or rare, (4) the second-stage options were identified by their fixed position (left or right), and (5) the time intervals for display of each screen were different. Many of these changes should facilitate model-based learning by making the task easier to understand.

Acknowledgements

We would like to thank Giuseppe M. Parente for the wonderful illustrations used in the spaceship and magic carpet tasks, Susanna Gobbi, Gaia Lombardi, and Micah Edelson for many helpful discussions and ideas, participants at the NYU Neuroeconomics Colloquium for useful feedback, and Peter Dayan, Stephan Nebe, Arkady Konovalov, and Ian Krajbich for helpful comments on an early draft of this manuscript. Note that our acknowledgment of their feedback does not imply that these individuals fully agree with our conclusions or opinions in this paper. We would also like to acknowledge Wouter Kool, Fiery Cushman, and Sam Gershman for making the data from their 2016 paper openly avail-

able at https://github.com/wkool/tradeoffs. This work was supported by the CAPES Foundation (https://www.capes.gov.br, grant number 88881.119317/2016-01) and by EU FP7 grant 607310.

Author Contributions

C.F.S. and T.A.H. designed the tasks and novel computational models. C.F.S. programmed the tasks, collected the data, and performed the analyses with input from T.A.H. C.F.S. and T.A.H. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

References

[1] Ahmet O Ceceli and Elizabeth Tricomi. Habits and goals: a motivational perspective on action control. *Current Opinion in Behavioral Sciences*, 20:110–116, apr 2018.

[2] A David Redish, Steve Jensen, and Adam Johnson. Addiction as vulnerabilities in the decision process. *Behavioral and Brain Sciences*, 31(4):461–487, 2008.

[3] Antonio Rangel, Colin Camerer, and P Read Montague. A framework for studying the neurobiology of value-based decision making. *Nature reviews. Neuroscience*, 9(7):545–56, jul 2008.

[4] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, dec 2005.

[5] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, first edition, 1998.

[6] Claire M. Gillan, A. Ross Otto, Elizabeth A. Phelps, and Nathaniel D. Daw. Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3):523–536, sep 2015.

[7] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215, mar 2011.

[8] Klaus Wunderlich, Peter Smittenaar, and Raymond J. Dolan. Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, 75(3):418–424, aug 2012.

[9] Ben Eppinger, Maik Walter, Hauke R. Heekeren, and Shu-Chen Li. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, 7, 2013.

[10] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, 110(52):20941–20946, dec 2013.

[11] A. Ross Otto, Samuel J. Gershman, Arthur B. Markman, and Nathaniel D. Daw. The Curse of Planning. *Psychological Science*, 24(5):751–761, may 2013.

[12] Peter Smittenaar, Thomas H.B. FitzGerald, Vincenzo Romei, Nicholas D. Wright, and Raymond J. Dolan. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. *Neuron*, 80(4):914–919, nov 2013.

[13] Miriam Sebold, Lorenz Deserno, Stefan Nebe, Daniel J. Schad, Maria Garbusow, Claudia Hägele, Jürgen Keller, Elisabeth Jünger, Norbert Kathmann, Michael Smolka, Michael A. Rapp, Florian Schlagenhauf, Andreas Heinz, and Quentin J.M. Huys. Model-Based and Model-Free Decisions in Alcohol Dependence. *Neuropsychobiology*, 70(2):122–131, 2014.

[14] V Voon, K Derbyshire, C Rück, M A Irvine, Y Worbe, J Enander, L R N Schreiber, C Gillan, N A Fineberg, B J Sahakian, T W Robbins, N A Harrison, J Wood, N D Daw, P Dayan, J E Grant, and E T Bullmore. Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*, 20(3):345–352, mar 2015.

[15] Bradley B. Doll, Daphna Shohamy, and Nathaniel D. Daw. Multiple memory systems as substrates for multiple decision systems. *Neurobiology of Learning and Memory*, 117:4–13, jan 2015.

[16] Bradley B Doll, Katherine D Duncan, Dylan A Simon, Daphna Shohamy, and Nathaniel D Daw. Model-based choices involve prospective neural activity. *Nature Neuroscience*, 18(5):767–772, mar 2015.

[17] Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822, nov 2015.

[18] A. Ross Otto, Anya Skatova, Seth Madlon-Kay, and Nathaniel D. Daw. Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27(2):319–333, feb 2015.

[19] Lorenz Deserno, Quentin J. M. Huys, Rebecca Boehme, Ralph Buchert, Hans-Jochen Heinze, Anthony A. Grace, Raymond J. Dolan, Andreas Heinz, and Florian Schlagenhauf. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, 112(5):1595–1600, feb 2015.

[20] J. H. Decker, A. R. Otto, N. D. Daw, and C. A. Hartley. From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychological Science*, 27(6):848–858, jun 2016.

[21] Wouter Kool, Fiery A. Cushman, and Samuel J. Gershman. When Does Model-Based Control Pay Off? *PLOS Computational Biology*, 12(8):e1005090, aug 2016.

[22] Wouter Kool, Samuel J. Gershman, and Fiery A. Cushman. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, page 095679761770828, jul 2017.

[23] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276, jul 2017.

[24] Wouter Kool, Samuel J. Gershman, and Fiery A. Cushman. Planning Complexity Registers as a Cost in Metacontrol. *Journal of Cognitive Neuroscience*, 30(10):1391–1404, oct 2018.

[25] Carolina Feher da Silva, Yuan-Wei Yao, and Todd A Hare. Can model-free reinforcement learning operate over information stored in working-memory? *bioRxiv*, 2018.

[26] Nitzan Shahar, Tobias U. Hauser, Michael Moutoussis, Rani Moran, Mehdi Keramati, Nspn Consortium, and Raymond J. Dolan. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, 15(2):e1006803, February 2019.

[27] Carolina Feher da Silva and Todd A. Hare. 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(4):e0195328, apr 2018.

[28] Asako Toyama, Kentaro Katahira, and Hideki Ohira. Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *Journal of Mathematical Psychology*, 91:88–102, August 2019.

[29] Marcos Economides, Zeb Kurth-Nelson, Annika Lübbert, Marc Guitart-Masip, and Raymond J. Dolan. Model-Based Reasoning in Humans Becomes Automatic with Training. *PLOS Computational Biology*, 11(9):e1004463, September 2015.

[30] Nathaniel D. Daw. Are we of two minds? *Nature Neuroscience*, 21(11):1497, November 2018.

[31] Thomas Akam, Rui Costa, and Peter Dayan. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLOS Computational Biology*, 11(12):e1004648, dec 2015.

[32] Kevin J. Miller, Amitai Shenhav, and Elliot A. Ludvig. Habits without values. *Psychological Review*, 126(2):292–311, mar 2019.

[33] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, sep 2017.

[34] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

[35] Peter Dayan and Kent C. Berridge. Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, 14(2):473–492, jun 2014.

[36] Peter Dayan and Yael Niv. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2):185–96, apr 2008.

[37] Angela Radulescu, Yael Niv, and Ian Ballard. Holistic Reinforcement Learning: The Role of Structure and Attention. *Trends in Cognitive Sciences*, 23(4):278–292, April 2019.

[38] Nitzan Shahar, Rani Moran, Tobias U. Hauser, Rogier A. Kievit, Daniel McNamee, Michael Moutoussis, , and Raymond J. Dolan. Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, 2019.

[39] W. Schultz, P. Dayan, and P. R. Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, mar 1997.

[40] Hannah M. Bayer and Paul W. Glimcher. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, 47(1):129–141, July 2005.

[41] Andrew Caplin and Mark Dean. Axiomatic methods, dopamine and reward prediction error. *Current Opinion in Neurobiology*, 18(2):197 – 202, 2008. Cognitive neuroscience.

[42] Ethan S. Bromberg-Martin, Masayuki Matsumoto, Simon Hong, and Okihide Hikosaka. A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. *Journal of Neurophysiology*, 104(2):1068–1076, aug 2010.

[43] Brian F Sadacca, Joshua L Jones, and Geoffrey Schoenbaum. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife*, 5, mar 2016.

[44] Melissa J Sharpe, Chun Yun Chang, Melissa A Liu, Hannah M Batchelor, Lauren E Mueller, Joshua L Jones, Yael Niv, and Geoffrey Schoenbaum. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, 20(5):735–742, apr 2017.

[45] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6):1075–1081, dec 2012.

[46] Bernard W. Balleine and John P. O'Doherty. Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, 35(1):48–69, January 2010.

[47] Arkady Konovalov and Ian Krajbich. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nature Communications*, 7:12438, aug 2016.

[48] A. David Redish. Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3):147–159, mar 2016.

[49] Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, oct 2010.

[50] Amitai Shenhav, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, 40(1):99–124, 2017.

[51] Marcos Economides, Zeb Kurth-Nelson, Annika Lübbert, Marc Guitart-Masip, and Raymond J. Dolan. Model-Based Reasoning in Humans Becomes Automatic with Training. *PLOS Computational Biology*, 11(9):e1004463, sep 2015.

[52] Claire M Gillan, Michal Kosinski, Robert Whelan, Elizabeth A Phelps, and Nathaniel D Daw. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, 5, mar 2016.

[53] Claire M. Gillan, Martina Papmeyer, Sharon Morein-Zamir, Barbara J. Sahakian, Naomi A. Fineberg, Trevor W. Robbins, and Sanne de Wit. Disruption in the Balance Between Goal-Directed Behavior and Habit Learning in Obsessive-Compulsive Disorder. *American Journal of Psychiatry*, 168(7):718–726, July 2011.

[54] Patricia Gruner and Christopher Pittenger. Cognitive inflexibility in Obsessive-Compulsive Disorder. *Neuroscience*, 345:243–255, March 2017.

[55] Hannah R Snyder, Roselinde H Kaiser, Stacie L Warren, and Wendy Heller. Obsessive-compulsive disorder is associated with broad impairments in executive function: A meta-analysis. *Clinical Psychological Science*, 3(2):301–330, 2015.

[56] Stan Development Team. PyStan: the Python interface to Stan, 2017.

[57] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

[58] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.16.0, 2017.

[59] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, oct 2009.

[60] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, aug 2016.

[61] Kevin J Miller, Carlos D Brody, and Matthew M Botvinick. Identifying Model-Based and Model-Free Patterns in Behavior on Multi-Step Tasks. *bioRxiv*, page 14, 2016.

[62] Stefano Palminteri, Valentin Wyart, and Etienne Koechlin. The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, 21(6):425–433, 2017.