nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0905-y

Humans primarily use model-based inference in the two-stage task

Carolina Feher da Silva ✉ and Todd A. Hare ✉

Distinct model-free and model-based learning processes are thought to drive both typical and dysfunctional behaviours. Data from two-stage decision tasks have seemingly shown that human behaviour 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 have little, if any, simple model-free influence. We also demonstrate that behaviour in the two-stage task may falsely appear to be driven by a combination of simple model-free and model-based learning if purely model-based agents form inaccurate models of the task because of misconceptions. Furthermore, we report evidence that many participants do misconceive 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 behaviour in the two-stage task and connections between reward learning, habit formation and compulsivity.

Investigating the interaction between habitual and goal-directed processes is essential to understand both normal and abnormal behaviour[1–3]. 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 behaviour[4] and may potentially protect against habit formation[6]. Model-based behaviour selects actions by computing their current values on the basis of a model of the environment.

Two-stage tasks have been used frequently to dissociate model-free and model-based influences on behaviour[6–24]. Their critical feature is that participants make choices at the first stage of each trial, then transition probabilistically to a specific second-stage state (Fig. 1a). Each of the two first-stage actions leads to one second-stage state with higher probability (for example, 70%) and to the other with lower probability (for example, 30%). Thus, there are common (high-probability) and rare (low-probability) transitions that depend on the first-stage choice. Model-free and model-based learning generate different predictions about the probability that in the next trial the participant will repeat their previous first-stage choice as a function of the previous transition and outcome. According to traditional logic, simple model-free agents are more likely to repeat a first-stage action that resulted in a reward, regardless of transition, and thus exhibit a positive main effect of reward (Fig. 2a, although see ref. [25]). Conversely, model-based agents first consider which second-stage stimulus is most likely to yield a reward, then select the first-stage action that will most likely lead to it, based on a model (that is, knowledge) of the task's structure[7]. Thus, model-based agents factor in the transition and exhibit a positive reward × transition interaction effect (Fig. 2b). Hybrid agents that combine both strategies exhibit both effects (Fig. 2c).

Past studies using the original two-stage task (Fig. 1a) have always reported that healthy adult humans use a hybrid mixture of model-free and model-based learning (for example, refs. [7,20,21] and Fig. 2d). Moreover, most studies implementing modifications to the two-stage task designed to promote model-based learning[21,22,24] find a reduced but substantial model-free influence on behaviour. Overall, the consensus has been that the influence of simple model-free learning on human behaviour is ubiquitous and robust.

Our current findings call into question how ubiquitous simple model-free learning is. We found that slight changes to the task instructions markedly reduced the apparent evidence for model-free learning in two separate experiments (Fig. 2e,f). We ran a series of simulations of purely model-based agents that used incorrect models of the two-stage task and found that the agents falsely appeared to be hybrid. Lastly, we show that human participants often misrepresent basic features of the two-stage task. This means inaccurate models of the two-stage task could be the true reason for some, or even all, of the past findings of hybrid behaviour rather than competition between model-free and model-based learning algorithms.

Results

Improving the two-stage task instructions decreases the apparent model-free influence on behaviour. We developed two modified versions of the two-stage task—the magic carpet task and the spaceship task—with the goal of clearly explaining all features of the task. We thought that if participants' apparent model-free behaviour was caused by a poor mental representation of the task, our improved instructions would shift them toward the correct model-based behaviour. Conversely, if their apparent model-free behaviour truly resulted from a competition between parallel model-based and model-free systems, our improved instructions would not make any difference.

Specifically, we incorporated a detailed story in the instructions and stimuli (Fig. 1b–d). In the magic carpet task, the participant chose a magic carpet and flew on it to Pink Mountain or Blue Mountain, where genies might give them a gold coin. In the spaceship task, the participant bought a spaceship ticket and flew to Planet Red or Planet Black in search of valuable crystals that grew inside obelisks. Previous studies have already used stories to explain the two-stage task to human participants[20,21] but they did not provide a reason for all the events in the task (see Supplementary Methods). Conversely, our instructions provided a concrete reason for every

Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Zurich, Switzerland. ✉e-mail: carolina.feherdasilva@econ.uzh.ch; todd.hare@econ.uzh.ch

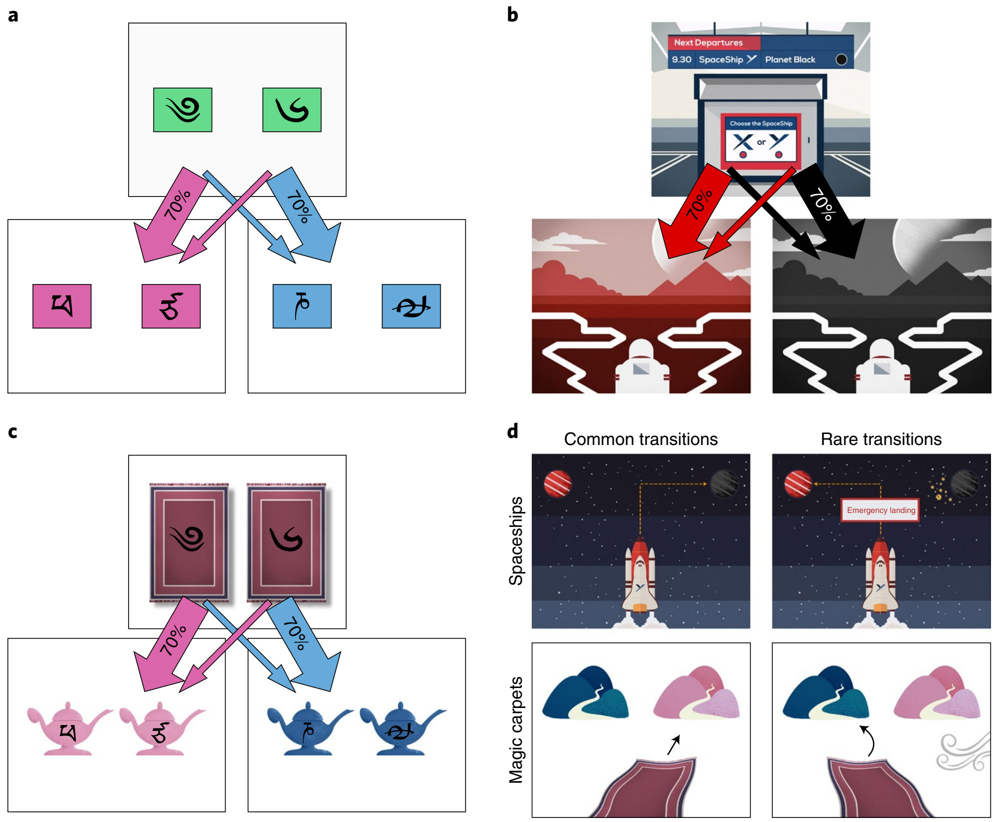


Fig. 1 | The stimuli used in the three versions of the two-stage task. a, The original, abstract version of the task, reproduced from ref. [7]. In each trial, the participant makes choices in two consecutive stages. In the first stage, the participant chooses one of two green boxes, each of which contains 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**. **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 training session, 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. See Extended Data Fig. 1 for the timelines of the spaceship and magic carpet task.

potential event within the task; in particular, different explanations and visual depictions were given for common versus rare transitions (Fig. 1d) during training. Importantly, the magic carpet task used different stimuli for the training trials versus the main task and there were no visual depictions of the transitions during the main task. During the practice trials, we also displayed additional messages, which explained the reason for each event again as it happened and could not be skipped (see Methods).

Both the magic carpet and the spaceship tasks have the same payoff structure and transition probabilities as the original two-stage task[7]. The magic carpet task replicates all features of the original two-stage task, except for the instructions. In contrast, the spaceship task was originally designed to test how humans learn when the first-stage states are defined by a combination of two stimuli. In this task, there were four different first-stage stimuli (flight announcements) that defined the initial state and the transition probabilities associated with the two first-stage actions (Fig. 3a). This feature allowed us to demonstrate that a common reverse inference—significant main effects of reward in a logistic regression analysis indicate model-free learning—is invalid.

Hybrid model fits indicate that behaviour becomes more model-based with comprehensive instructions. To test the impact of our instructions on the apparent levels of model-based and model-free behaviour, we fitted to the data the standard hybrid reinforcement learning model, proposed originally by Daw et al.[7]. To facilitate comparison with previous studies, we fit the model to each participant using maximum likelihood estimation. The hybrid model combines the model-free SARSA(λ) algorithm with model-based learning and explains first-stage choices as a combination of 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 one indicates a purely model-based strategy and, conversely, a model-based weight equal to zero indicates a purely model-free strategy.

The estimated model-based weights for the participants who performed the spaceship or the magic carpet task were substantially

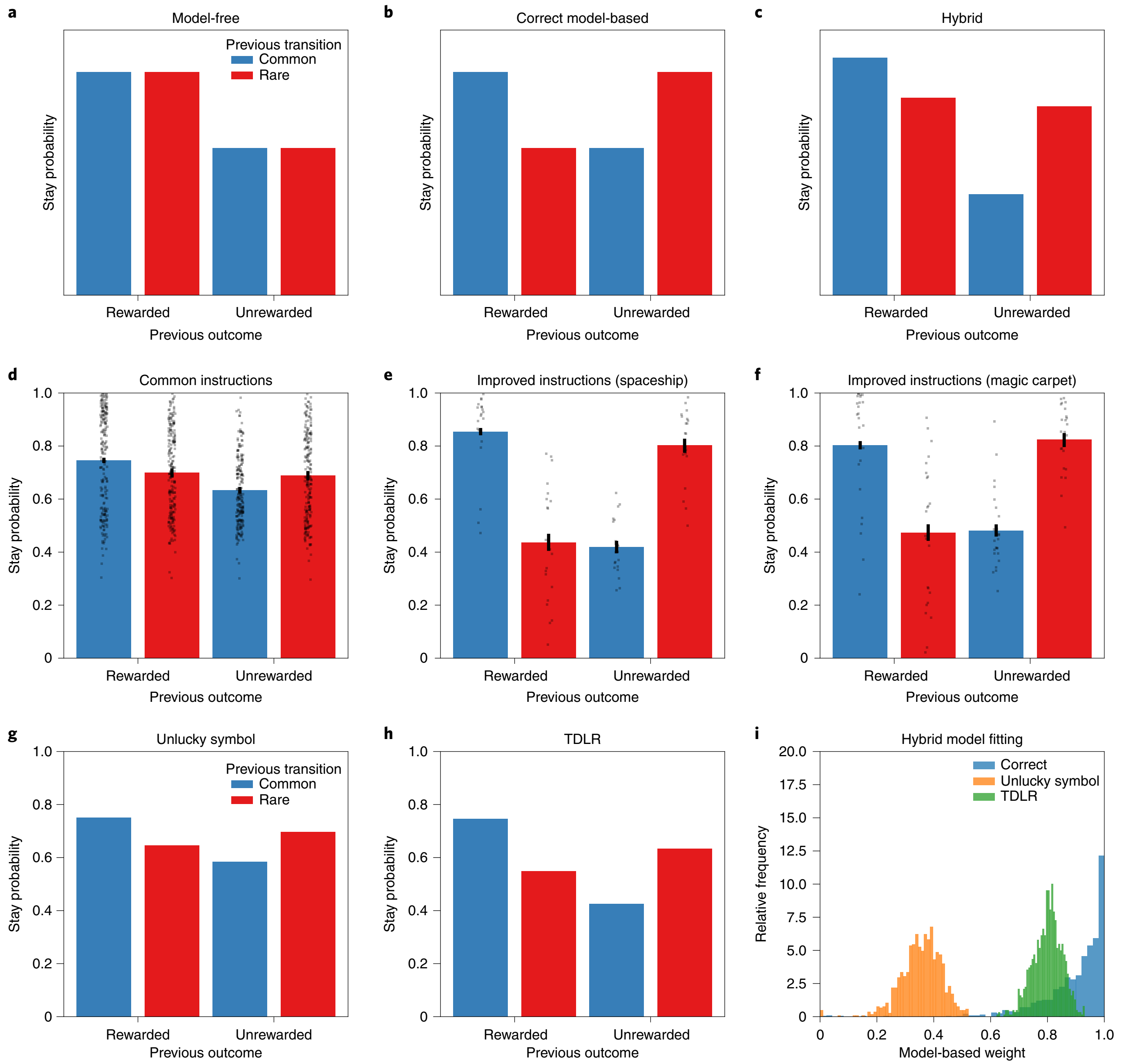


Fig. 2 | Stay probabilities for human participants; stay probabilities and model-based weights for simulated agents. a–c, Idealized stay probabilities of purely model-free (**a**), purely model-based (**b**) and hybrid (**c**) agents as a function of the previous outcome and transition. To generate the hybrid data plotted in this figure, we used the logistic regression coefficient values for data from adult participants in a previous study[20]. **d**, The behaviour of participants ($n=206$) after receiving common instructions[21] shows both a main effect of reward and a reward × transition interaction. Both mean and individual stay probability values were obtained from a hierarchical Bayesian logistic regression model, with the median of the posterior distribution used as a point estimate. **e**,**f**, In contrast, the choices of participants after receiving improved instructions in the new spaceship task ($n=21$) (**e**) and magic carpet task ($n=24$) (**f**) show a much more model-based pattern. **g**–**i**, 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. **g**,**h**, Histograms show the mean stay probabilities for unlucky symbol (**g**) and transition-dependent learning rates (TDLR) (**h**) model-based agents. **i**, The histograms show the fitted model-based weight parameters (w) for simulated agents using the correct (blue; median = 0.94, 95% CI [0.74, 1.00]), unlucky symbol (orange; median = 0.36, 95% CI [0.24, 0.48]) and TDLR (green; median = 0.80, 95% CI [0.70, 0.90]) models of the task. We simulated 1,000 agents of each type. Model-based weights for each agent were estimated by fitting the simulated choice data (1,000 choices per agent) with the original hybrid model by maximum likelihood estimation. Error bars in **d**–**f** represent the 95% HDI. Error bars for the simulated data are not shown because they are very small due to the large number of data points.

higher than the estimated weights obtained in two previous studies[7,21] using common (less comprehensive) task instructions (Fig. 4a). We also fit a Bayesian hierarchical model containing the hybrid algorithm to the data from our magic carpet and spaceship tasks as well as the control condition data from a previous study using common instructions ($n=206$)[21]. Our results indicate that the posterior

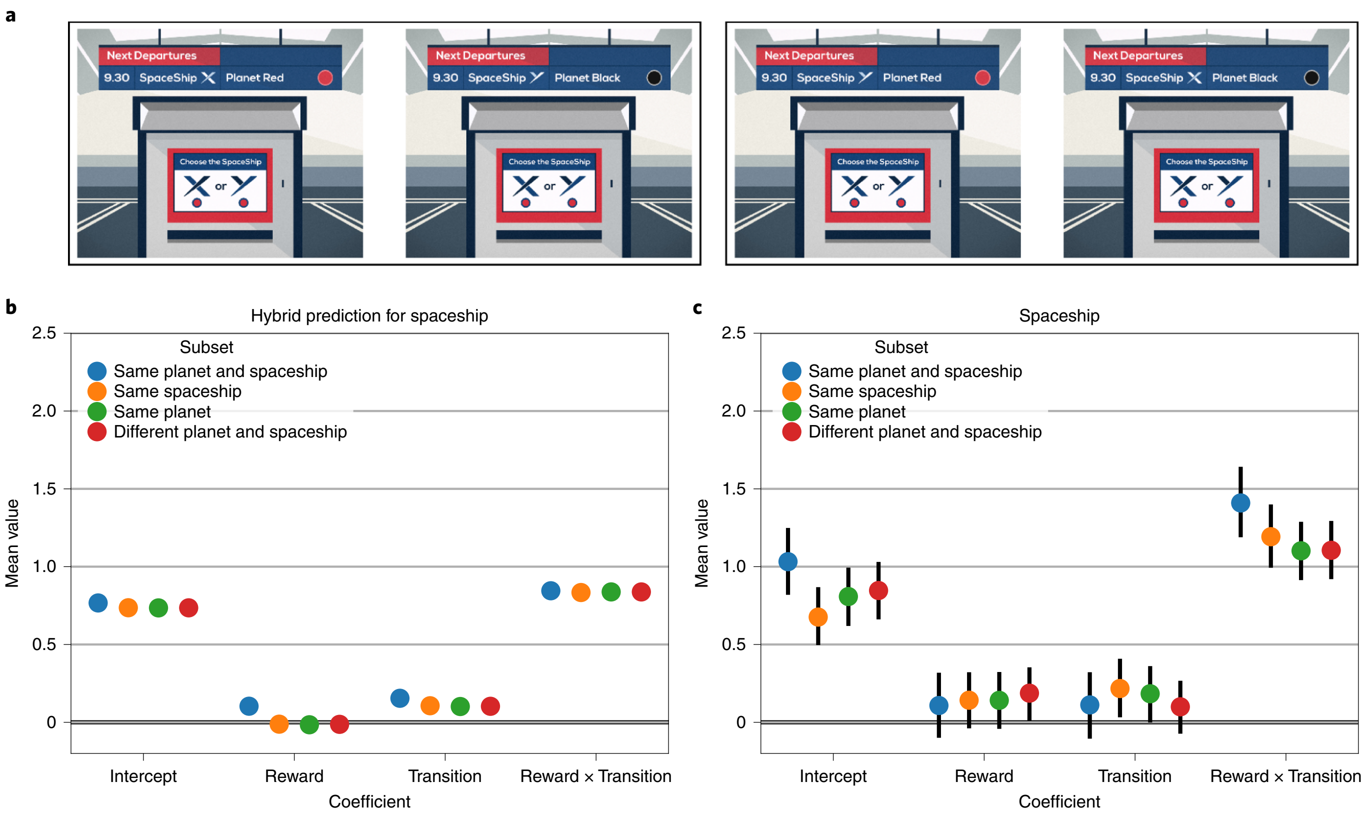


Fig. 3 | Example of a reward main effect that cannot be driven by model-free learning. a, Each trial of the spaceship task began with a flight announcement on an information board above a ticket machine. The name of a spaceship was presented for 2 s, then the name of the planet the spaceship was scheduled to fly to appeared for another 2 s. There were two spaceships (X and Y) and two planets (Red and Black) and hence four possible flight announcements. The participant selected a spaceship to fly on based on the announced flight. Each announcement listed only one flight but still provided complete information about all flights. This is because the spaceships always departed at the same time and flew to different planets: if one would fly to Planet Red, then the other would fly to Planet Black and vice versa. Thus, the two screens in each rectangle of **a** are equivalent. **b**, Results for simulated hybrid agents ($n = 5{,}000$) performing 1,000 trials of the spaceship task under the standard assumption that model-free learning does not generalize between different state representations (that is, different flight announcements). Simulation parameters were the median estimates obtained by fitting the hybrid model to the human spaceship data 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 categories: (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 was announced in the 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 spaceship data ($n = 21$), with consecutive trial pairs divided into the same four categories. In contrast to the simulated hybrid agents, the small reward effect in human behaviour does not differ across categories. Thus, it is inconsistent with a standard model-free learning system. Error bars represent the 95% HDI.

probability that the average weights in the magic carpet and spaceship datasets are greater than the average weight in the common instructions dataset is greater than 0.9999.

Standard logistic regression analyses also indicate primarily model-based behaviour in the magic carpet and spaceship tasks. We also analysed our results using a logistic regression analysis of consecutive trial pairs. In this analysis, the stay probability (probability of repeating a first-stage action) is a function of two variables: reward, indicating whether the previous trial was rewarded; and transition, indicating whether the previous trial's transition was common or rare. Model-free learning generates a main effect of reward (Fig. 2a), while model-based learning generates a reward × transition interaction (Fig. 2b). The core finding in most studies is that healthy adult participants behave like hybrid agents (Fig. 2c), exhibiting both a main effect of reward and a reward × transition interaction.

For comparison to our data, we reanalysed the common instructions dataset (control condition from ref. [21]). Figure 4b shows that when all trial pairs were combined, the coefficient of the reward × transition interaction that indicates correct model-based control is 5.9 times larger in the magic carpet (95% highest density interval, HDI [4.5, 7.3]) and spaceship (95% HDI [4.7, 7.2]) tasks compared to the common instructions data[21]. Moreover, the reward effect, which is generally considered evidence of model-free control, is higher in the common instructions results (95% HDI [0.20, 0.30]) compared to the magic carpet results (95% HDI [−0.03, 0.20]) with 0.99 probability (Bayes factor 145) and the spaceship results (95% HDI [0.05, 0.24]) with 0.98 probability (Bayes factor 51).

Spaceship task data reveal misleading evidence of model-free influence. Although choices in spaceship tasks showed large reward × transition interaction effects, there was a small but significant main effect of reward on choices (Fig. 4b). This may at first suggest that our enhanced instructions decreased but did not eliminate the influence of model-free learning on these participants. We took advantage of specific properties of the spaceship task to further investigate this reward effect. We found that it is misleading as evidence of model-free influence because it contradicts one of the basic properties of model-free learning.

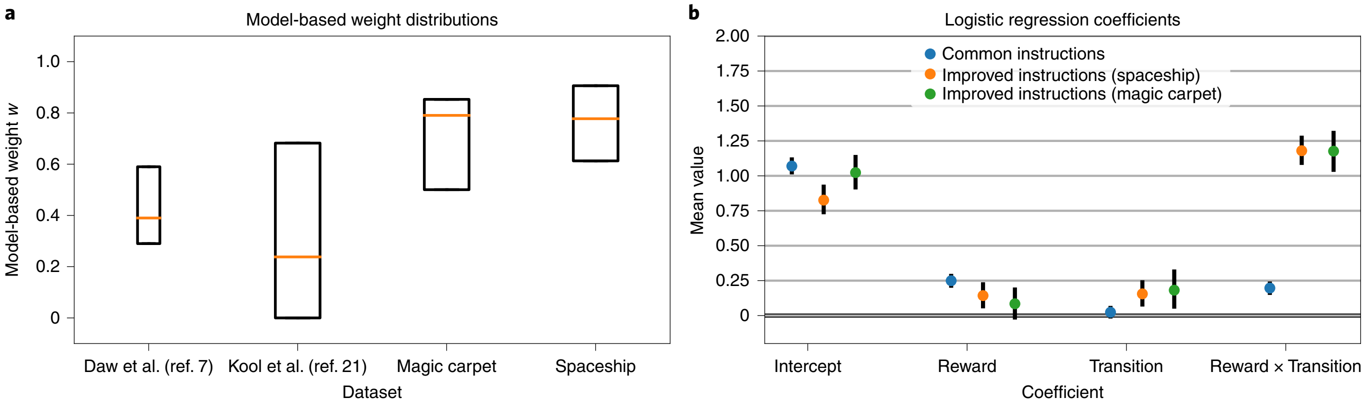


Fig. 4 | Model-based weights and logistic regression coefficients for different empirical datasets. a, Summary statistics (25%, 50% and 75% percentiles) of the estimated model-based weights for four datasets: Daw et al.[7] ($n=17$), Kool et al.[21] (common instructions dataset, $n=206$), the spaceship dataset ($n=21$) and the magic carpet dataset ($n=24$). For the Daw et al.[7] dataset, the weight estimates were simply copied from the original article. For the other datasets, we obtained the weight estimates by a maximum likelihood fit of the hybrid reinforcement learning model to the data from each participant. **b**, This plot shows the mean and 95% HDI of all coefficients in the hierarchical logistic regressions on stay probabilities for the common instructions[21], magic carpet and spaceship datasets. These logistic regression coefficients were used to calculate the stay probabilities shown in Fig. 2d–f. The improved instruction datasets have both lower reward and higher reward × transition interaction effects compared to the common instructions dataset. Note that the main effect of reward in the spaceship task is actually inconsistent with a model-free influence on behaviour in that task (see Fig. 3). There are significant effects of transition on stay probabilities in the improved instruction datasets as well. This coefficient indicates that the probability of repeating the same first-stage action increases after a common transition and decreases after a rare transition to the second stage. Transition effects are consistent with standard model-free, model-based or hybrid behaviour under certain combinations of the learning and choice parameters in those algorithms (see Supplementary Results).

Within the spaceship task there are pairs of first-stage stimuli that indicate the same initial state by presenting different information (Fig. 3a). This allows us to subdivide consecutive trial pairs into four categories on the basis of the information announced on the flight board. This information determines which first-stage action will most likely lead to a given second-stage state (Fig. 3a and Table 1). We analysed the data from each category using the standard logistic regression.

If the reward effect was driven by model-free learning, it should be positive only when the stimulus–response pairing is identical across trials (category 1, same spaceship and planet). This is because model-free learning is assumed to be unable to generalize between distinct state representations[15,16,21] and thus should have no influence on choices in categories 2–4 (Fig. 3b). The results, however, are contrary to this expectation (Fig. 3c): The observed reward effect had a similar magnitude in all four categories, including those where either the stimuli or the response required to implement the same choice are different between two consecutive 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. These results strongly suggest that the observed reward effects are not model-free. Instead, they are probably model-based, except that the models participants used were not completely correct, as assumed by the analysis.

Table 1 | Reward main effects in different trial pair categories for the spaceship task

Category	Reward effect magnitude	Probability > 0 (Bayes factor)	Announcements included
1	0.12 [−0.10, 0.32]	85% (5.6)	The same spaceship and planet
2	0.14 [−0.03, 0.32]	94% (14.9)	The same spaceship but different planets
3	0.14 [−0.04, 0.33]	93% (13.9)	Different spaceships but the same planet
4	0.19 [+0.02, 0.36]	98% (62.9)	Different spaceships and different planets

Pairs of consecutive trials were divided into four categories depending on the flight announcement, and each subset was separately analysed by logistic regression. The categories are numbered in the order (left to right) that they are shown in Fig. 3. Column two lists the group-level means for the reward main effect by category. The values in square brackets denote the range of the 95% HDI. Column three lists the posterior probability that the main effect of reward is greater than zero in each category and its equivalent Bayes factor. Column four lists the (dis)similarities in the flight announcement screens across consecutive trial pairs that are used to define categories 1–4. For categories 2 and 3 combined, which corresponded to 'stay' choices requiring the selection of a different spaceship, the posterior probability that the reward effect is greater than zero is 98% (Bayes factor 46.0).

The logistic regression model is better than the hybrid model at explaining first-stage choices in all tasks. We found that the hybrid reinforcement learning model[7] does not describe humans' first-stage choices well in any of the three datasets. We compared how well three models—the simple logistic regression model of stay probabilities in consecutive trial pairs, the hybrid model and the correct model-based model—fit human participants' first-stage choices. We compared the models based on how well they explained first-stage choices only because the logistic regression model is only fit to first-stage choices, and the model-free and model-based algorithms only generate different behaviour at the first stage. Both Akaike information criterion (AIC) and Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO) metrics strongly indicated that the simple logistic regression model best explained participants choices (Table 2). In fact, PSIS-LOO scores favoured the logistic regression model by >4 s.e.m.

The logistic regression model describes stay probabilities as constant throughout the experiment; it is not a learning model. This suggests that participants' behaviour in the two-stage task is not well described as a combination of the correct model-based and model-free learning algorithms that constitute the hybrid model. Notably, the logistic regression model revealed little evidence of simple model-free learning with our improved instructions. Overall, our results suggest that participants in the magic carpet and

Table 2 | Overall AIC and PSIS-LOO scores for three models fitted to human participant data

Condition	AIC			PSIS-LOO		
	Logistic regression	Model-based	Hybrid	Logistic regression	Model-based	Hybrid
Common instructions	27,073	27,441	27,246	26,725.7 ± 140.4	27,685.0 ± 152.8	27,290.8 ± 158.6
Magic carpet	4,682	5,113	5,063	4,665.7 ± 67.8	5,032.5 ± 63.2	4,921.8 ± 63.5
Spaceship	5,118	5,684	5,849	5,109.9 ± 73.3	6,415.0 ± 85.7	5,617.2 ± 66.0

Participants belonged to either the common instructions ($n = 206$), magic carpet ($n = 24$) or spaceship ($n = 21$) condition. We fit participants' choice data using a logistic regression model of consecutive trial pairs, the correct model-based reinforcement learning model and the hybrid reinforcement learning model. AIC calculation failed for one participant in the common instructions condition and so this participant was excluded from the AIC analysis. The difference between PSIS-LOO scores for the logistic regression and the hybrid model were −565.0 ± 104.6, −256 ± 60.2 and −507.2 ± 63 for the common instructions, magic carpet and spaceship datasets, respectively. Errors given for the PSIS-LOO scores correspond to the s.e.m. For both the AIC and PSIS-LOO metrics, smaller values indicate better model fits.

spaceship tasks used a slightly incorrect model-based strategy. (See also Supplementary Results.)

Model-based learning can be confused with model-free learning. Our experiments with improved instructions demonstrate that participants' understanding of the task plays a large role in how model-based or model-free they appear to be. Here, we present simulated data to demonstrate that purely model-based agents using incorrect task models may be misclassified as hybrid when the data are analysed by either of the standard methods: logistic regression or reinforcement learning model fitting. We present two examples of incorrect task models. We do not suggest that they are the only or even the most probable ways that people may misconceive of the task. Instead, they are merely examples to demonstrate our points.

As a reference, we simulated correct model-based agents, based on the hybrid model proposed by Daw et al.[7] with $w = 1$. Consistent with recent work by Sharar et al.[26], even when agents have a w equal to exactly one, used the correct model of the task and performed 1,000 trials, the recovered w parameters were not always precisely one (Fig. 2i). This is expected, because parameter recovery is noisy and, in the standard specification of the hybrid model, w cannot be greater than one, thus any error was an underestimate of w.

The two alternative, purely model-based learning algorithms are: the 'unlucky symbol' and the 'transition-dependent learning rates' (TDLR) algorithm (see Methods for full details). Briefly, the unlucky-symbol algorithm has the mistaken belief that certain first-stage symbols decrease the reward probability of second-stage choices. In the current example, we simulated agents that believe a certain first-stage symbol is unlucky and lowers the values of second-stage actions by 50%. The TDLR algorithm is a model-based learning algorithm that has a higher learning rate after common than rare transitions. Note that both algorithms contain the correct state transition structure. This suggests that checking if participants understood the transitions is not enough to show that they properly modelled the task; there may be other elements of their models that are incorrect.

Our simulations show that TDLR and unlucky-symbol agents, despite being purely model-based, display the same behavioural pattern as healthy adults given the common task instructions and simulated hybrid agents (Fig. 2). We analysed the simulated data with the hybrid algorithm containing the correct model of the task (that is, the standard analysis). The resulting distributions of the estimated model-based weights are shown in Fig. 2i. They indicate a model-free influence on the behaviour of purely model-based TDLR and unlucky-symbol agents. Moreover, when fitting the correct model-based and hybrid models to the data and performing model comparisons, we found lower AIC scores (indicating a better fit; difference = −5,897) for the purely model-based algorithm only when agents used the correct model. For the TDLR and unlucky-symbol agents, the correct model-based algorithm yielded higher AIC scores (that is, a worse fit) than the hybrid model (AIC differences = 26,858 and 198,932, respectively) even though these agents did not include any model-free influence. Together, these results demonstrate that analysing two-stage task choices using a hybrid algorithm can lead to the misclassification of purely model-based agents as hybrid if the agents have an incorrect model of the task.

Human behaviour deviates from the hybrid model's assumptions. We tested if human behaviour following common instructions[21] also violates assumptions of the hybrid model. First, we note that poor overall hybrid model fits were significantly associated with more apparent evidence of model-free behaviour. There was a significant positive correlation between the model-based weight and the log-likelihood of the model fit (Spearman's rho = 0.19, $n = 206$, %95 CI [0.06, 0.33], $P = 0.005$). There is no inherent reason for such correlation if participants are using a hybrid mixture of simple model-free and correct model-based learning, assuming all other factors (for example, exploration) are equal. Upon further analysis, we found that this correlation was driven by participants with an estimated model-based weight smaller than 0.1 (about 43% of the sample). This suggests that human participants that behaved in a way that substantially deviated from the hybrid model's assumptions were assigned very low model-based weights to maximize the correspondence between hybrid model predictions and their behaviour (although this was still a poor match). One potential reason for this bias is that the model-free portion of the hybrid model has two unique free parameters, giving it flexibility to match different behaviours, while the model-based portion has none. Therefore, the model-free portion of the hybrid algorithm may be better at fitting behaviour that deviates from the correct model.

Second, the standard hybrid model cannot explain why participants in this experiment behaved differently when first-stage symbols were presented on different sides of the screen in consecutive trials. In many implementations of the two-stage task, the symbols presented at each stage appear on random sides. The sides are irrelevant and only the symbol identity matters. If participants understand the task, changes in first-stage symbols locations should not influence their choices. Nevertheless, past studies have anticipated that participants might have a tendency to repeat key presses at the first stage regardless of symbol locations and thus modified the standard hybrid model to add a response stickiness parameter[21]. For comparison with participant data, we simulated hybrid agents with response stickiness. We then divided consecutive trial pairs from the simulated and participant datasets into two subsets: (1) same sides, if the first-stage choices were presented on the same sides in both trials and (2) different sides, if the first-stage choices switched sides from one trial to the next. We analysed each subset separately using logistic regressions (Fig. 5a,c). The results show a larger intercept in the same-sides subset compared to the different-sides subset for both simulated agents and participants. In the simulated data,

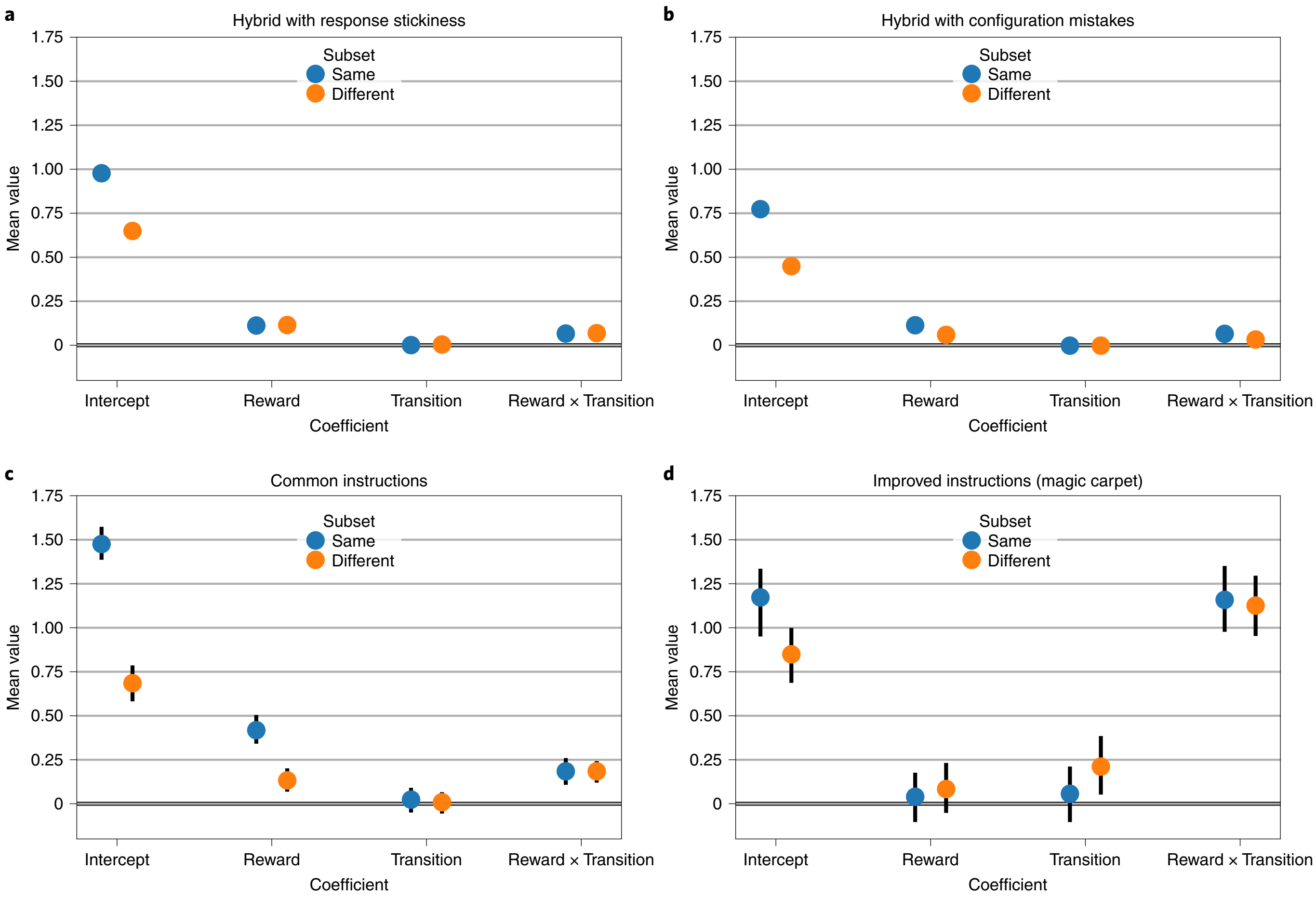


Fig. 5 | Simulated behaviour of agents and real participants can be influenced by irrelevant changes in stimulus position. a–d, All four panels show behaviour in the two-stage task on trials in which the position of the first-stage stimuli remains the same (blue) across consecutive trials compared to 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, transition and the interaction between reward and transition on the previous trial. In contrast to previous studies, we analysed 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 ($n=5{,}000$) performing 1,000 trials of the two-stage task. The median parameter values from the common instructions dataset[21] were used in these simulations. **b**, Results from a simulation of hybrid agents that occasionally made configuration mistakes ($n=5{,}000$) performing 1,000 trials of the two-stage task. The median parameter values from the common instructions dataset[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 reanalysis of the common instructions dataset[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 dataset. **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 **c** and **d** represent the 95% HDI.

this effect was caused by response stickiness. However, human participants also showed a larger reward coefficient in the same-sides subset (mean 0.41, 95% HDI [0.35, 0.47]) versus the different-sides subset (mean 0.14, 95% HDI [0.08, 0.19]), with the posterior probability that the reward coefficient is larger in the former being >0.9999 (Bayes factor >60,000).

There are several potential explanations for these side-specific results. It could be that model-free learning is sensitive to stimulus locations or specific responses and 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, the reward effect for the different-sides subset should be zero but it is not. 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 latter possibility, we fit these data with a hybrid model that included an extra parameter quantifying the probability of making a configuration mistake and pressing the wrong key after a location change (see Methods ‘Fitting of reinforcement learning models’). Note that this configuration-mistake parameter is distinct from decision noise or randomness because it quantifies response probabilities when symbols have switched places from one trial to the next and thus only decreases effect sizes in the different-sides subset (Fig. 5b).

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 configuration-mistake probability higher than 0.9. The goodness of fits for the two models were equivalent on average for the 111 low-configuration-mistake participants (mean score difference, 0.0; s.e.m., 0.1), As expected, the configuration-mistake model fit better for the 51 high-configuration-mistake participants (mean score difference, −34.1; s.e.m., 6.2). These results suggest that most participants rarely made configuration mistakes but ~25% made mistakes more than nine out of ten times when the symbols switched sides.

However, it is possible that the high-configuration-mistake participants were not truly making mistakes. Configuration mistakes decrease all effect sizes in the different-sides subset, including the reward × transition interaction effect (Fig. 5b) but this was not observed in the common instructions dataset. Rather, some participants may have instead mismodelled the task—for example, they may have believed a stimulus's location was more important than its identity. In any case, these results suggest that different participants conceive of the two-stage task in different ways and that many participants misconceptualized basic aspects of the task.

Smaller 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 showed little difference in the logistic regression coefficients between the same-side and different-side subsets (Fig. 5d). They also made few configuration mistakes, with only one participant having a mistake probability >0.05. Thus, the enhanced magic carpet instructions vastly increased the probability that participants would act correctly when equivalent choices were presented on different sides of the screen.

Discussion

We show that simple changes in the two-stage task instructions led healthy adult humans to behave in a highly model-based manner. This is in contrast to most studies, 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 incorrect models, analysis results can falsely indicate an influence of model-free learning and mistakenly classify them as hybrid. If participants use incorrect models of the task, none of the traditional markers of model-based or model-free behaviour derived from logistic regression or hybrid algorithms indicate the ground-truth about participants' strategies. Therefore, our work, together with other recent reports on the limitations of hybrid model fits[26,27], indicates the need to reconsider aspects of both the empirical tools and theoretical assumptions that currently pervade the study of reward learning.

Human behaviour does not necessarily fall into the dichotomy of simple model-free and correct model-based learning assumed in the design and analysis of two-stage tasks[28–35]. Instead, agents can act on a multitude of strategies (Fig. 6). It is known that this multidimensional space includes complex model-free algorithms that can mimic model-based actions in some cases[29,31]. We show that it also includes model-based strategies that can appear hybrid because they are inconsistent with the environment. This adds uncertainty to any attempt to classify strategies in the two-stage task.

Concluding that behaviour is a mix of two exemplar algorithms is especially tenuous. The observed choices do not match the predictions of either algorithm; arguably the most plausible reason for this is that participants are not using either algorithm. Still, it is possible that both algorithms operate in parallel and jointly influence decisions. Indeed, this has been the usual conclusion, probably because of strong prior beliefs in the existence of dual learning systems. However, it relies on the strong assumption that participants have an accurate task model, which we show may not hold in many cases.

In line with the possibility that apparently hybrid behaviour is, instead, driven by misconceptions of the task, we found drastic shifts toward (correct) model-based behaviour in healthy adult humans when participants received more elaborate instructions. It comes as no surprise that people don't always understand or follow instructions well. However, the impact of these misconceptions on our ability to make accurate inferences about reward learning was unexpected to us. In both our local samples and openly available data, we found strong indications that people misconstrued the two-stage task. Among them are the location effects detected in the common instructions data[21]. Those participants completed the task online but similar behaviour has been reported by Shahar et al.[36] from participants in a laboratory. Shahar et al.[36] proposed that location effects are due to model-free credit assignment to outcome-irrelevant task features, such as stimulus locations and key responses. However, improved instructions for our tasks greatly reduced side-switch errors, suggesting that model-based misconceptions were more likely the cause.

Our spaceship task provides empirical evidence that healthy adult humans may use partially incorrect models—even with improved instructions. Behaviour on that task was almost correct model-based. However, there was a small reward effect in a logistic regression analysis of consecutive trials with the same or different initial states. Such an effect is assumed to be evidence of model-free learning[7] but model-free learning is thought to be unable to generalize between distinct state representations[15,16,21,29]. If it drove the small reward effect in the spaceship task, it would be able to generalize not only between distinct state representations but also distinct actions. Therefore, the observed effect was more likely generated by imprecise model-based learning. It is not clear how the participants misunderstood the spaceship task or even if they all misunderstood it in the same way. The fact that there are unlimited incorrect model-based models and some seem to mimic a hybrid agent is the core obstacle to drawing accurate conclusions from behaviour in the two-stage task. We also found that when the hybrid model doesn't explain human choices well, it is biased toward indicating more model-free influence. These findings corroborate previous reports that a measure of understanding, the number of attempts required to pass a comprehension quiz, was associated with a lower model-based weight[6]. Overall, the data show that, in the absence of sufficient understanding, the standard analysis methods overestimate model-free influences on behaviour.

Model-free learning can explain many animal electrophysiology and human neuroimaging studies. A highly influential finding is that dopamine neurons respond to rewards consistently with the reward prediction errors obtained by temporal difference learning, a form of model-free learning[37–39]. However, in the studies showing this response, there was no higher-order structure to the task—often, there was no instrumental task. Thus, model-based behaviour was indistinguishable from model-free. Conversely, when the task is more complicated, it has been found that the dopamine signal reflects knowledge of task structure (for example, refs. [40,41]). A recent optogenetic study showed that dopamine signals are necessary and sufficient for model-based learning in rats[42] 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 errors[7,43]. Moreover, although there is evidence that anatomically distinct striatal systems mediate goal-directed and habitual actions[44], to date there is no evidence for anatomically separate model-free and model-based systems.

Importantly, the two-stage task does not directly measure habits and a lack of evidence for model-free learning does not imply humans do not form or follow habits. Initial theoretical work proposed the model-based versus model-free distinction to formalize the distinction between goal-directed and habitual control[4].

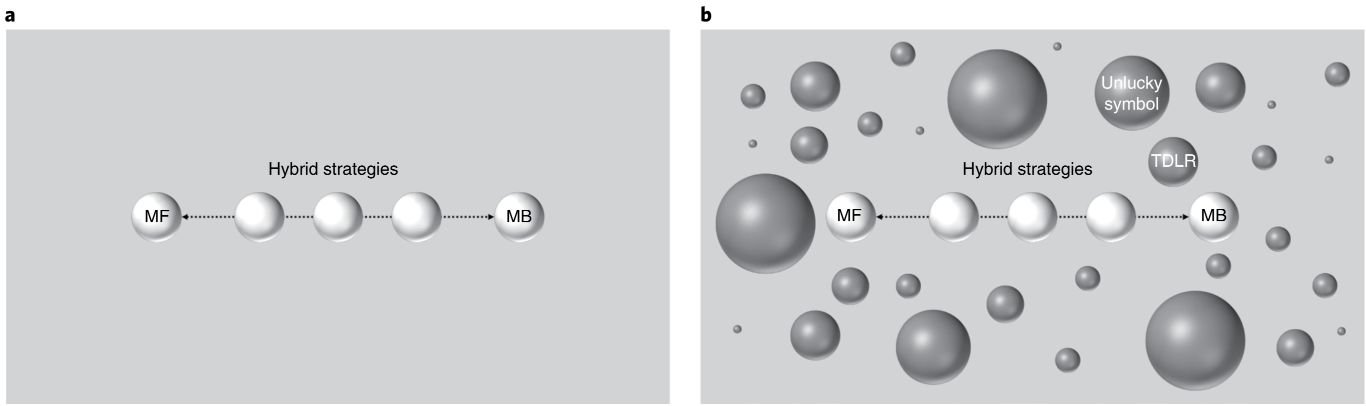


Fig. 6 | Simplified diagrams representing the strategy space in the two-stage task. a, The commonly used analysis methods for two-stage task behaviour assume that the space of all strategies agents can use to perform this task contains only the MF and correct MB strategies, as well as intermediate hybrid strategies on the line between them. **b**, However, the true strategy space also contains other purely model-based strategies, such as the TDLR and unlucky-symbol strategies, among countless others. If these other strategies are projected onto the line segment between MF and correct MB, they will yield the false conclusion that participants are using a hybrid MF/MB strategy. This is essentially what the standard analyses of the two-stage task do. In reality, a human participant who misconceives of the two-stage task in some way may use any of the potential strategies that exist in a multidimensional space. Importantly, if people misconceive the task in different ways, then they will use different incorrect models that correspond to their current mental representation of the task. Moreover, people may well switch between different strategies over the course of the experiment if they determine that their conceptualization of the task was wrong in some way. 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 MF and MB learning a daunting, if not impossible, task.

However, it is generally assumed that goal-directed actions can be based on model-free learning too. Similarly, there is an ongoing debate as to whether habits arise exclusively from model-free learning[1,4,45–52]. In any case, two-stage task studies indicate that participants classified as primarily model-free are probably not acting on habits.

Apparently model-free participants behave inconsistently with the habitual tendencies that model-free learning supposedly indexes. A study by Konovalov and Krajbich[53] combined eye-tracking with two-stage task choices to examine first-stage fixation patterns as a function of learning type. They divided participants into model-free and model-based, based on a median split of their estimated model-based weight ($w = 0.3$). They reported that when first-stage symbols were presented, model-based learners tended to look once at each symbol, as if they had already decided which to choose. In contrast, model-free learners tended to make more fixations and their choices related more closely to fixation duration. These results suggest that model-free participants made goal-directed rather than habitual comparisons at the first stage. This is because similar patterns of head movements, presumably analogous to fixations, are seen when rats initially learn to navigate a maze[54]. The head movements accompany hippocampal representations of reward in the direction the animal faces and are seen as evidence that the animals deliberate over choices in a goal-directed fashion. Humans also make more fixations per trial as difficulty increases in goal-directed paradigms[55]. Notably, head movements and hippocampal place cell signalling cease once animals are extensively trained and act on habits[54].

In contrast to habits, apparent model-free behaviour decreases with extensive training on the two-stage task. In general, the frequency and strength of habits increase with experience. However, Economides et al.[56] showed that estimated model-free influence in human participants decreases over 3 d of training on the two-stage task. They also found that, after 2 d of training, participants remain primarily model-based even when performing a Stroop task in parallel. Both results raise questions about the effortfulness of model-based learning in the two-stage task. After all, its transition model may be tricky to explain but it is easy to follow once understood. Rats also show primarily model-based behaviour after receiving extensive training on the two-stage task[23]. Moreover, inactivation of the dorsal hippocampus or orbitofrontal cortex in the rats impaired model-based planning but did not increase model-free influence, which remained negligible[23]. Thus, while it is possible that humans and other animals may use model-free strategies in some cases, these results are difficult to reconcile with the idea of a competition between model-based and model-free systems for behavioural control.

Humans have been reported to arbitrate between model-based and model-free strategies on the basis of both their accuracy and effort. Several lines of evidence indicate that humans and other animals generally seek to minimize physical and mental effort[57]. Model-based learning is thought to require more effort than model-free learning and a well-known aspect of the original two-stage task[7] is that model-based learning does not lead to greater payoffs than model-free learning[21,29]. This has been hypothesized to lead participants to use a partially model-free strategy[21,29]. Previous studies tested this hypothesis by modifying the original two-stage task so that model-based strategies achieve more rewards[21,22,24]. Participants appeared more model-based when a model-based strategy paid off more and thus they concluded that participants will use model-based learning if it is advantageous in a cost–benefit tradeoff between effort and money. Our results and those from studies with extensive training[23,58] cannot be explained by such tradeoffs. The magic carpet and spaceship tasks led to almost completely model-based behaviour but had the same tradeoffs as the original two-stage task[7]. Similarly, the profitability of model-based learning does not change with experience. If anything, more experience should allow the agent to learn that the model-based strategy is no better than the model-free one if both are being computed in parallel. However, a possibility that merits further study is that giving causes for rare transitions reduced the subjective effort of forming the correct model.

Seemingly model-free behaviour may be reduced in all three sets of experiments through better understanding of the task. Obviously, improved instructions and more experience can lead to better understanding. In modified two-stage tasks that make model-based learning more profitable, the differential payoffs also provide clearer feedback to participants about the correctness of their models. Conversely, if both correct and incorrect models lead to the same average payoffs, participants may be slow to realize their mistakes. Of course, increased understanding and changes in cost–benefit ratios may jointly drive the increases in (correct) model-based behaviour and additional data are needed to tease these possibilities apart.

It is not clear what is being measured by the two-stage task. Studies showed that intelligence quotient, working memory, processing speed and dorsolateral prefrontal cortex function are associated with model-based weights in the standard two-stage task with common instructions[10–12,59]. Therefore, determining and employing the correct model-based strategy probably relies on working memory and other prefrontal cortex (PFC)-mediated cognitive functions. Apparent model-free behaviour has been reported to correlate with compulsive symptoms[6,14,60]. Given our current results, however, the conclusion that model-free learning and compulsive behaviours are linked in healthy populations or those with obsessive compulsive disorder should be drawn with caution (see also Supplementary Discussion).

There are some remaining open questions and potential limitations of the current work that we wish to highlight. One potential limitation of this study is that our instructions may have simply encouraged participants to be more model-based, rather than alleviating confusion. However, while they might encourage model-based behaviour by explicitly describing and providing reasons for the transitions, the near-elimination of location-based mistakes is direct evidence that the instructions also facilitated the formation and use of more accurate task models. Another potential limitation is that our improved instructions decreased, but did not eliminate, the use of partially incorrect models. It is unknown how these models deviate from the correct model and why participants formed them. The possibility that participants may have used multiple models hinders the design of a computational model that can accurately describe two-stage task behaviour. Unfortunately, it is unclear how one might modify the instructions or task design to limit the space of models participants could potentially use to a small set of testable alternatives. Lastly, thus far we have only examined the influence of our improved instructions on behaviour in healthy young adults. How these instructions might change behaviour in younger or older individuals and clinical populations remains to be determined.

Reward learning is one of the most widely studied processes in the neural and behavioural sciences. Researchers must use tools that can ascertain their mechanistic properties, their potential neural substrates and how their dysfunction might lead to suboptimal behaviour or psychopathology. The specification of algorithms for model-free versus model-based learning has advanced the study of reward learning. 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’[28]. We believe that our current results are another strong indication that the time to move beyond oversimplified binary frameworks has come.

Methods

The magic carpet and spaceship tasks. Twenty-four 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. No statistical methods were used to predetermine sample sizes but our sample sizes were based on our previous pilot studies using the two-stage task[61]. Data collection and analysis were not performed blind to the conditions of the experiments. 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 Swiss Francs (CHF)0.37 or CHF0.29 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 (https://github.com/carolfs/muddled_models), 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 Fig. 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 (Fig. 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; that is, they knew which transition was common and which was rare. Also, during the tutorial flights, the participants saw a transition screen (Fig. 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 than 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. Which carpet more frequently flew to which mountain was randomized for each participant.

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 (Fig. 3a). The task design 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[29]. 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 could 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. Each trial's flight announcement was randomly selected with uniform probability.

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 (Fig. 1b). This transition screen was displayed during both the practice trials and the task trials (Fig. 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 (randomized for each participant), (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.

Simulations of model-based agents. The model-based agents described in the Results were simulated and their decisions analysed by reinforcement learning model fitting. A total 1,000 agents of each type (original hybrid, unlucky symbol and TDLR) 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[62], was run ten times with random seeds and for 5,000 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 analyse the results of the two-stage task.

Initially, at time $t = 1$, the model-free (MF) values $Q_1^{\mathrm{MF}}(s, a)$ 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 $\delta_t^2 = r_t - Q_t^{\mathrm{MF}}(s_2, a_2)$, the difference between the chosen action's current value and the received reward r_t. The update is performed as

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

where α_2 is the second-stage learning rate ($0 \le \alpha_2 \le 1$). 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

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

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

The model-based (MB) value $Q_t^{\mathrm{MB}}(s_2, a_2)$ of each action a_2 performed at second-stage state s_2 is the same as the corresponding model-free value, that is, $Q_t^{\mathrm{MB}}(s_2, a_2) = Q_t^{\mathrm{MF}}(s_2, a_2)$. 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

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

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, $\mathcal{S} = \{\text{pink}, \text{blue}\}$ is the set of second-stage states, max is the maximum function 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 ($0 \le w \le 1$), according to a soft-max distribution as follows

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

where β_1 is the first-stage's inverse temperature parameter, which determines the exploration–exploitation tradeoff 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 $\mathrm{rep}_t(a')$ is defined as 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 equation (4) becomes

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

where the variable $\mathrm{resp}_t(a')$ is defined as 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%. We reasoned that it is possible that an agent 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 with each symbol but also how they believe the symbol affects the reward probabilities of second-stage choices.

This model-based algorithm has three parameters: $0 \le \alpha \le 1$, 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

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

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

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

where $\mathrm{unlucky}(a) = \eta$ if the agent thinks action a is unlucky and $\mathrm{unlucky}(a) = 1$ 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

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

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

The 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. The TDLR algorithm was inspired by debriefing comments from participants in a pilot study, which suggested that they assign greater importance to outcomes observed after common (expected) relative to rare transitions.

The TDLR algorithm has three parameters: α_c, the higher learning rate for outcomes observed after common transitions ($0 \le \alpha_c \le 1$); α_r, the lower learning rate for outcomes observed after rare transitions ($0 \le \alpha_r < \alpha_c$); and $\beta > 0$, an inverse temperature parameter that determines the exploration–exploitation tradeoff. 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

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

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

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

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.

The hybrid algorithm with a mistake probability. We developed a hybrid model with a mistake probability to analyse the symbol location effects in the common instructions dataset. 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

$$P_t(s_1, a_1) = (1 - \rho)P_t^h(s_1, a_1) + \rho(1 - P_t^h(s_1, a_1)) \tag{13}$$

This model also assumes that the participant realized their mistake after making one and that action values were updated correctly.

Analysis of the common instructions data. In ref. [21], 206 participants recruited via Amazon Mechanical Turk performed the two-stage task for 125 trials. The behavioural data were downloaded from the first author's Github repository (https://github.com/wkool/tradeoffs) and reanalysed by logistic regression and reinforcement learning model fitting, as described next.

Logistic regression of consecutive trials. This analysis was applied to all behavioural datasets. Consecutive trial pairs were analysed together or first divided into subsets, depending on the presentation of first-stage stimuli, then analysed separately. The analysis used a hierarchical logistic regression model whose parameters were estimated through Bayesian computational methods. The predicted variable was p_{stay}, the stay probability, 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

$$p_{\text{stay}} = \frac{1}{1 + \exp[-(\beta_0 + \beta_r x_r + \beta_t x_t + \beta_{r\times t} x_r x_t)]} \tag{14}$$

For each trial pair, the variable y was equal to one if the agent chose in the next trial the same first-stage action as in the previous trial (a 'stay' choice) or equal to zero if the agent chose a different first-stage action (a 'switch' choice). The distribution of y was Bernoulli (p_{stay}). The distribution of the $\boldsymbol{\beta}$ vectors was $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. The hyperparameters $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$ were given vague prior distributions based on preliminary analyses—the $\boldsymbol{\mu}$ vectors' components were given a $\mathcal{N}(\mu = 0, \sigma^2 = 25)$ prior and the $\boldsymbol{\sigma}^2$ 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 modelling language[63,64] and used the PyStan Python package[62] 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 reinforcement learning models. The hybrid and correct model-based reinforcement learning model proposed by Daw et al.[7] were fitted to all datasets (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 (Fig. 3a).

The parameters fitted for the standard hybrid and correct model-based algorithms were as follows. 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

$$(\text{logit}(\alpha_1^i), \text{logit}(\alpha_2^i), \text{logit}(\lambda^i), \log(\beta_1^i), \log(\beta_2^i), \text{logit}(w^i), p^i)$$

were given a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ and obtained for each participant. These transformations 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 correct model-based algorithm had the same parameters except α_1, λ and w. 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 Σ was decomposed into a diagonal matrix τ, whose diagonal components were given a Cauchy prior distribution with mean 0 and variance 1 and a correlation matrix Ω, which was given an Lewandowski–Kurowicka–Joe (LKJ) prior[65] with shape $\nu = 2$ (ref. [64]). This model was coded in the Stan modelling language[63,64] and fitted to each dataset using the PyStan interface[62] to obtain a chain of 40,000 iterations (warmup 20,000) for the common instructions dataset and 80,000 iterations (warmup 40,000) for the magic carpet and spaceship datasets. 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 datasets. For that model, the data from each participant were described by a vector

$$(\text{logit}(\alpha_1^i), \text{logit}(\alpha_2^i), \text{logit}(\lambda^i), \log(\beta_1^i), \log(\beta_2^i), \text{logit}(w^i), p^i, \text{logit}(\rho^i))$$

where the additional parameter, logit(ρ^i), is the mistake probability after a side switch.

The hybrid and model-based algorithms were also fitted to data by maximum likelihood. They were first coded in the Stan modelling language[63,64] and fitted 1,000 times (for robustness) to each participant's choices using LBFGS algorithm implemented by Stan through the PyStan interface[62].

Model comparisons. To calculate PSIS-LOO scores for model comparison, each model was first fitted simultaneously to all behavioural data in each condition, as described above. Then, the log-likelihood of every trial was obtained for each iteration and used to calculate the PSIS-LOO score (an approximation of leave-one-out cross-validation) of each model (considering all trials from all participants together) or of each model for each participant (considering only the trials from that participant). To this end, the loo and compare functions of the loo R package were used[66].

To calculate AIC scores, we fit the hybrid and the model-based reinforcement learning models to all datasets by maximum likelihood estimation for each participant, as described above. The logistic regression model was fitted to the data from each participant using the statsmodels library[67]. AIC scores were then calculated from the log-likelihood of the model fit for each participant and summed across each dataset and model.

For the comparisons between the correct model-based, hybrid and logistic regression models, the AIC and PSIS-LOO scores were computed for the first-stage choices only. This is because the logistic regression model is only fit to first-stage choices and the model-free and model-based algorithms only generate different behaviour at the first stage but not at the second.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data obtained from human participants are available at https://github.com/carolfs/muddled_models

Code availability

All the code used to perform the simulations, run the magic carpet and the spaceship tasks, and analyse the results are available at https://github.com/carolfs/muddled_models

Received: 15 July 2019; Accepted: 30 May 2020;
Published online: 6 July 2020

References

1. Ceceli, A. O. & Tricomi, E. Habits and goals: a motivational perspective on action control. *Curr. Opin. Behav. Sci.* **20**, 110–116 (2018).

2. Redish, A. D., Jensen, S. & Johnson, A. Addiction as vulnerabilities in the decision process. *Behav. Brain Sci.* **31**, 461–487 (2008).
3. Rangel, A., Camerer, C. & Montague, P. R. A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545–556 (2008).
4. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
5. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction* (A Bradford Book, 1998).
6. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
7. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
8. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine enhances model-based over model-free choice behavior. *Neuron* **75**, 418–424 (2012).
9. Eppinger, B., Walter, M., Heekeren, H. R. & Li, S.-C. Of goals and habits: age-related and individual differences in goal-directed decision-making. *Front. Neurosci.* **7**, 253 (2013).
10. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA* **110**, 20941–20946 (2013).
11. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning. *Psychol. Sci.* **24**, 751–761 (2013).
12. Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 914–919 (2013).
13. Sebold, M. et al. Model-based and model-free decisions in alcohol dependence. *Neuropsychobiol.* **70**, 122–131 (2014).
14. Voon, V. et al. Disorders of compulsivity: a common bias towards learning habits. *Mol. Psych.* **20**, 345–352 (2015).
15. Doll, B. B., Shohamy, D. & Daw, N. D. Multiple memory systems as substrates for multiple decision systems. *Neurobiol. Learn. Mem.* **117**, 4–13 (2015).
16. Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).
17. Cushman, F. & Morris, A. Habitual control of goal selection in humans. *Proc. Natl Acad. Sci. USA* **112**, 13817–13822 (2015).
18. Otto, A. R., Skatova, A., Madlon-Kay, S. & Daw, N. D. Cognitive control predicts use of model-based reinforcement learning. *J. Cogn. Neurosci.* **27**, 319–333 (2015).
19. Deserno, L. et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc. Natl Acad. Sci. USA* **112**, 1595–1600 (2015).
20. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From creatures of habit to goal-directed learners: tracking the developmental emergence of model-based reinforcement learning. *Psychol. Sci.* **27**, 848–858 (2016).
21. Kool, W., Cushman, F. A. & Gershman, S. J. When does model-based control pay off? *PLoS Comput. Biol.* **12**, e1005090 (2016).
22. Kool, W., Gershman, S.J. & Cushman, F.A. Cost–benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
23. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
24. Kool, W., Gershman, S. J. & Cushman, F. A. Planning complexity registers as a cost in metacontrol. *J. Cogn. Neurosci.* **30**, 1391–1404 (2018).
25. FeherdaSilva, C. & Hare, T. A. A note on the analysis of two-stage task results: how changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. *PLoS ONE* **13**, e0195328 (2018).
26. Shahar, N. et al. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Comput. Biol.* **15**, e1006803 (2019).
27. Toyama, A., Katahira, K. & Ohira, H. Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *J. Math. Psychol.* **91**, 88–102 (2019).
28. Daw, N. D. Are we of two minds?. *Nat. Neurosci.* **21**, 1497–1499 (2018).
29. Akam, T., Costa, R. & Dayan, P. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Comput. Biol.* **11**, e1004648 (2015).
30. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
31. Momennejad, I. et al. The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
32. Dayan, P. Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624 (1993).
33. Dayan, P. & Berridge, K. C. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* **14**, 473–492 (2014).
34. Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).
35. Radulescu, A., Niv, Y. & Ballard, I. Holistic reinforcement learning: the role of structure and attention. *Trends Cogn. Sci.* **23**, 278–292 (2019).
36. Shahar, N. et al. Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl Acad. Sci. USA* **116**, 15871–15876 (2019).
37. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
38. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
39. Caplin, A. & Dean, M. Axiomatic methods, dopamine and reward prediction error. *Curr. Opin. Neurobiol.* **18**, 197–202 (2008).
40. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus–habenula–dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* **104**, 1068–1076 (2010).
41. Sadacca, B. F., Jones, J. L. & Schoenbaum, G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife* **5**, e13665 (2016).
42. Sharpe, M. J. et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).
43. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
44. Balleine, B. W. & O'Doherty, J. P. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69 (2010).
45. Dezfouli, A. & Balleine, B. W. Habits, action sequences and reinforcement learning. *Eur. J. Neurosci.* **35**, 1036–1051 (2012).
46. Dezfouli, A. & Balleine, B. W. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Comput. Biol.* **9**, e1003364 (2013).
47. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
48. Dezfouli, A., Lingawi, N. W. & Balleine, B. W. Habits as action sequences: hierarchical action control and changes in outcome value. *Philos. Trans. R. Soc. Lond. B* **369**, 20130482–20130482 (2014).
49. Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
50. Balleine, B. W., Dezfouli, A., Ito, M. & Doya, K. Hierarchical control of goal-directed action in the cortical-basal ganglia network. *Curr. Opin. Behav. Sci.* **5**, 1–7 (2015).
51. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
52. Morris, A. & Cushman, F. Model-free RL or action sequences? *Front. Psychol.* **10**, 2892 (2019).
53. Konovalov, A. & Krajbich, I. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nat. Commun.* **7**, 12438 (2016).
54. Redish, A. D. Vicarious trial and error. *Nat. Rev. Neurosci.* **17**, 147–159 (2016).
55. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).
56. Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R. J. Model-based reasoning in humans becomes automatic with training. *PLoS Comput. Biol.* **11**, e1004463 (2015).
57. Shenhav, A. et al. Toward a rational and mechanistic account of mental effort. *Annu. Rev. Neurosci.* **40**, 99–124 (2017).
58. Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R. J. Model-based reasoning in humans becomes automatic with training. *PLoS Comput. Biol.* **11**, e1004463 (2015).
59. Schad, D. J. et al. Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Front. Psychol.* **5**, 1450 (2014).
60. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).
61. Feher da Silva, C., Yao, Y.-W. & Hare, T.A. Can model-free reinforcement learning operate over information stored in working-memory? Preprint at *bioRxiv* https://doi.org/10.1101/107698 (2018).
62. Stan Development Team. *PyStan: the Python interface to Stan* http://mc-stan.org (2017).
63. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* http://www.jstatsoft.org/v76/i01/ (2017).
64. Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*, Version 2.16.0 (2017).

65. Lewandowski, D., Kurowicka, D. & Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001 (2009).
66. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).
67. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92–96 (SciPy, 2010).

Acknowledgements

We thank G. M. Parente for the wonderful illustrations used in the spaceship and magic carpet tasks, S. Gobbi, G. Lombardi and M. Edelson for many helpful discussions and ideas, the participants at the NYU Neuroeconomics Colloquium for useful feedback and P. Dayan, W. Kool, A. Konovalov, I. Krajbich and S. Nebe for helpful comments on early drafts of this manuscript. Our acknowledgement of their feedback does not imply that these individuals fully agree with our conclusions or opinions in this paper. We also acknowledge W. Kool, F. Cushman and S. Gershman for making the data from their 2016 paper openly available at https://github.com/wkool/tradeoffs. This work was supported by the CAPES Foundation (https://www.capes.gov.br) grant no. 88881.119317/2016-01 and the European Union's Seventh Framework programme for research, technological development and demonstration under grant no. 607310 (Nudge-it). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-020-0905-y.

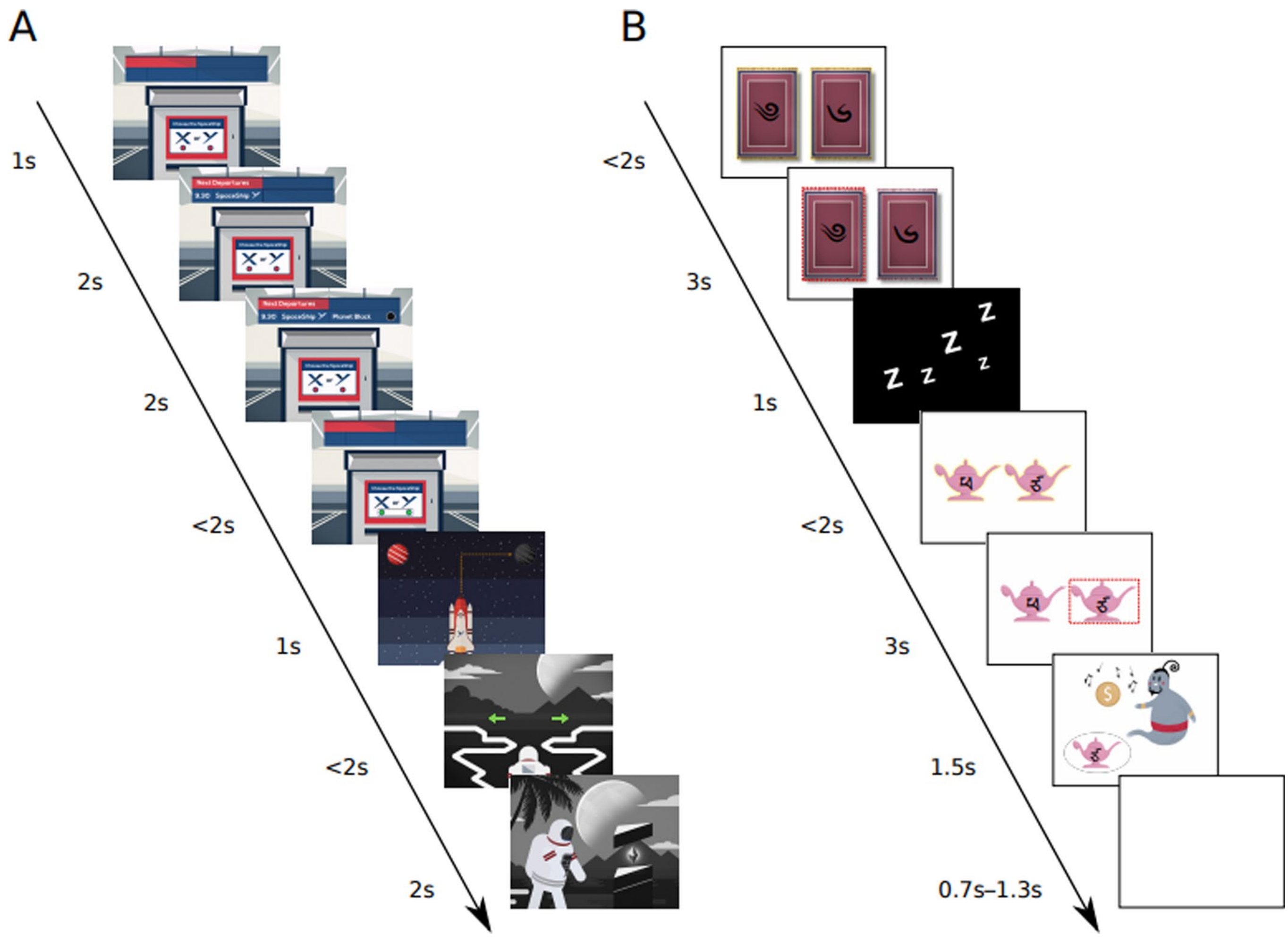
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0905-y.

Correspondence and requests for materials should be addressed to C.F. or T.A.H.

Peer review information Primary Handling Editor: Marike Schiffer.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Timelines of the spaceship and magic carpet task. Each box depicts an event within the spaceship or magic carpet tasks. The duration of each even is given in seconds on the left. A) In the spaceship task, the 1st screen simply indicates that a new trial has begun. The 2nd and 3rd screens represent the initial state. At the 4th screen, the participant has up to 2 seconds to indicate her choice. The common or rare transition is shown on the 5th screen. The second-stage state was indicated by the background colour (black, red) and the choice by the green left and right arrows on the 6th screen. The 7th and final screen in a trial revealed whether or not a reward was delivered. After feedback, the task advanced directly to the next trial. B) The magic carpet task was designed to closely mimic the original, abstract version of the two-stage task while still allowing for story-based instructions that included causes and effects for all task events. Thus, we used the same Tibetan characters from the original task, made them into labels for magic carpets and genies rather than simply identifying coloured squares. In the magic carpet task, the 1st screen represented the initial state and first-stage choice. Participants had up to 2s to make this choice. On the second screen, the chosen option was highlighted for 3 seconds. Next, a 'nap' screen was shown for 1s while the magic carpet automatically took the participant to one of the two mountains. Although participants saw the common or rare transition screens depicted in Fig. 1d during the practice trials, the transitions were not shown during the main task to make it more comparable with previous versions. The second-stage state (blue, pink) and choice were indicated by the pink or blue lamps on the right and left side of the 4th screen. The participant had up to 2s to make her choice. The 5th screen highlighted the chosen lamp/genie for3s. The 6th and final screen in a trial revealed whether or not a reward was delivered. After reward feedback, there was a blank screen for 0.7-1.3s before the next trial began.

Corresponding author(s): Carolina Feher da Silva, Todd A. Hare

Last updated by author(s): May 21, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PsychoPy 1.90
Data analysis	We used custom Python 3.6 and R 3.6.0 code to analyze the data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data obtained from human participants are available at https://github.com/carolfs/muddled_models

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Two groups of young healthy participants performed the two-stage task in a laboratory after receiving different instructions. Computational models were fitted to our data as well as an openly available data set.

Research sample

Participants were recruited from the University of Zurich's Registration Center for Study Participants. The sample comprises healthy young participants and is thus in line with typical experiments in Psychology, including Daw et al. (2011). Information about their sex and age is not available.

Sampling strategy

No statistical methods were used to pre-determine sample sizes, but our sample sizes were based on our previous pilot studies using the two-stage task.

Data collection

Participants from each group performed the task simultaneously in a laboratory, where they had their own desk and computer. They were alone in the room during the experiment, and both the researcher and assistants remained in an adjacent room from where they could observe the participants through glass. The researcher was not blind during data collection.

Timing

Data from the magic carpet experiment were collected on 12 June 2018 and data from the spaceship experiment were collected on 25 April 2018.

Data exclusions

No participants who completed the experiment were excluded.

Non-participation

No participants declined participation. Five recruited participants did not show up for the experiment or the computer crashed and they could not complete the experiment.

Randomization

Participants were not allocated to experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

See above.

Recruitment

Participants were recruited from the University of Zurich's Registration Center for Study Participants.

Ethics oversight

Zurich Cantonal Ethics Commission

Note that full information on the approval of the study protocol must also be provided in the manuscript.