Model-free decision making resists improved instructions and is enhanced by stimulus-response associations

Raúl Luna[1, 2, *], Miguel A. Vadillo[2], David Luque[2, 3, *]

[1] Institute of Optics. Spanish National Research Council (CSIC).
[2] Department of Basic Psychology. Faculty of Psychology. Universidad Autónoma de Madrid.
[3] Department of Basic Psychology and Speech Therapy. Faculty of Psychology. Universidad de Málaga.

* Corresponding authors.
E-mail address: raul.lunadelvalle@gmail.com (Raúl Luna)
E-mail address: david.luque@gmail.com (David Luque)

Abstract

Human behaviour is driven by two types of processes running in parallel: goal-directed and habitual, each supported by different computational-learning mechanisms, model-free and model-based respectively. In model-free strategies, stimulus-response associations are strengthened when actions are followed by a reward and weakened otherwise. In model-based learning, previous to selecting an action, the current values of the different possible actions are computed based on a detailed model of the environment. Previous research with the two-stage task suggests that participants' behavior usually shows a mixture of both strategies. But, interestingly, a recent study by da Silva and Hare (2020) found that participants deploy a purely model-based behavior when they are given detailed instructions about the structure of the task. In the present study, we reproduce this essential experiment using a larger sample size (N=59). However, our results do not suggest a sole model-based behaviour, but rather a hybrid one. Furthermore, an additional experiment shows that slight changes in the task, like a consistent stimulus-response mapping, can encourage reliance on model-free strategies, even if participants are presented with improved instructions. This suggests that the model-free marker, as measured by the two-stage task, is related to S-R learning.

Keywords: *Two-stage task, reinforcement learning, model-based, model-free, habits, goal-directed*

Introduction

It is often assumed that behaviour is based on two types of processes: goal-directed and habitual. From a computational point of view, each of these processes corresponds to two different reinforcement-learning (RL) strategies: model-free and model-based, respectively. In the case of model-free strategies, stimulus-response (S-R) associations are strengthened when actions or responses are followed by a reward and become weakened otherwise (Sutton & Barto, 1988). On the other hand, model-based learning

generates behaviour determined by the ongoing values of all available actions. The computation of these values is based on a model of the environment, that is to say, a sort of "cognitive map" in a non-spatial domain (Dolan & Dayan, 2013). In contrast to model-free representations, these maps consider not only if current actions lead to immediate rewards, but also if they lead to new states in which other actions may produce other (better) rewards.

There is an extensive literature exploring the model-free vs. model-based dichotomy using a particular experimental paradigm: the two-stage task or two-choice Markov decision task (Daw, Gershman, Seymour, Dayan & Dolan, 2011; Decker, Otto, Daw & Hartley, 2016; Miller, Botvinick & Brody, 2017; Kool, Gershman & Cushman, 2018). Each trial in this task requires participants to go through two sequential stages. In the first stage participants are asked to select one of two options. This choice is followed by a second stage with two possible scenarios or states. In most trials, a specific option in the first stage causes a transition to a determined state in the second stage (i.e., a common transition), but in a minority of trials it may also cause a transition to the alternative state (i.e., a rare transition). At the second stage, participants are asked again to choose between two options, each leading to a different reward. The specific options presented to participants during the second stage depend on the scenario or state in which the second stage takes place. Model-free strategies lead to the repetition of actions that have previously been rewarded, regardless of the type of transition that brought the participant to a certain second-stage state in past trials. Because model-based strategies include knowledge about whether rewards were obtained as a result of an unlikely transition, they can lead to the selection of the opposed first-stage action in future trials to obtain the same reward in the second stage. Therefore, actions not leading to reward in a current trial may still be executed in future trials if the transition was rare.

Because model-free agents are prone to repeating a first-stage action that ended up in a reward irrespective of the transition they experienced, these are expected to show a positive main effect of reward in the previous trial. In contrast, model-based agents are expected to show a reward × transition interaction. This is because, based on a cognitive map of the task, the most rational decision is to select the first-stage action that will most likely lead to the largest second-stage reward. Of course, it is possible to combine both strategies. Such "hybrid" agents should show both a main effect of

reward and a reward × transition interaction. Past studies have revealed the widespread use of hybrid strategies in healthy adult humans (Daw et al., 2011; Decker et al., 2016). More specifically, research suggests a prevailing use of model-based strategies, but also a significant presence of model-free behaviour, even under experimental conditions specifically designed to favour model-based learning (Kool, Cushman & Gersham, 2016; Kool, Gershman & Cushman, 2017; Kool, Gershman & Cushman, 2018).

Contrary to these findings, da Silva & Hare (2020) demonstrated that it is possible to observe performance consistent with an only-model-based strategy when participants are provided with accurate instructions about the task, so that they can create a correct cognitive model about it. That is, these authors argue that the model-free component evidenced in past studies is the result of an inaccurate internal model of the task, produced by a poor understanding of the instructions and the experimental paradigm. In addition to gathering empirical evidence supporting this view, they also conducted a computational-modelling analysis showing that an incorrect internal model of the task can give rise to the main effect of reward that is often taken as evidence of model-free strategies.

In their version of the two-stage task, da Silva & Hare (2020) used first- and second-stage options that randomly swapped sides from trial to trial. This aspect of the procedure may prevent the formation of strong associations between these stimuli and specific motor commands (Molinero et al., 2021; Verleger et al., 2016, 2018). This is an aspect that may hinder the execution of habitual responses. Consistetly with this, Hardwick, Forrence, Krakauer & Haith, 2019, found that training specific S-R associations always executed with the same motor command produced habits. Also, Neal, Wood, Wu & Kurlander, 2011, showed that previously formed habits disappear when changing the response pattern. Also, Luque et al. (2020) have shown that habits formed after extended and consistent S-R training interfere with new S-R mappings. Therefore, changing response option positions at random, as it is usual in the two-stage task, may favour the operation of the goal-directed system and overshadow any possible implication of the habit system. A question that remains unsolved is whether presenting the response options at fixed locations throughout the two-stage task would enhance model-free learning—even when the participants are provided with detailed instructions so they have an accurate internal model of the task.

Using a larger sample size, we attempted to replicate da Silva & Hare's (2020) results using the same task and the same improved instructions. Furthermore, and for the first time, we tested whether displaying response options at fixed locations leads to stronger evidence of model-free strategies.

To foreshadow, contrary to da Silva & Hare (2020), our results failed to show purely model-based behaviour in the two-stage task even with their improved instructions. What we found instead is that participants exhibit hybrid strategies, in agreement with classical results (Daw et al., 2011; Decker et al., 2016). In addition, we provide evidence that fixing the location of response options potentiates a model-free component in participants' behaviour, suggesting that the model-free marker measured by the two-stage task is related to S-R learning. Importantly, the methods and analysis plan of the present studies were pre-registered before data collection (https://osf.io/x9sya).

Results

As explained above, one objective of this study was to replicate the results from da Silva & Hare (2020), that is, we sought to find evidence of purely model-based behaviour in the two-stage task with improved instructions. Additionally, this study aimed to test whether fixing the state option locations across trials potentiates a model-free component in the same task, even with detailed instructions. Two experimental groups, Replica and Fixed-Locations, were formed to achieve these objectives. The only difference between them was that, in the Fixed-Locations condition, the location of the two response options in each state remained unchanged across trials, whereas in the Replica condition, locations changed randomly.

Logistic regression analysis

Consecutive trial pairs were analyzed through logistic regression, where the probability of repeating the same first-stage action as in the previous trial (i.e., the probability of "staying") was a function of reward and the type of transition in the previous trial. Reward was coded as +1 if the previous trial had been rewarded and -1 otherwise.

Transition was coded as +1 if the participant's response in stage 1 had led to the common state in stage 2 (i.e., common transition) and -1 otherwise (i.e., rare transition). Figure 1A displays the predicted stay probabilities separately for each Reward and Transition condition, both for the two groups of the present study and for the original experiment by da Silva and Hare (2020). Figure 1B shows the estimated logistic regression model coefficients for each study. According to da Silva & Hare (2020), the reward × transition interaction, indicative of model-based behaviour, takes place when participants have a good mental representation of the two-stage task induced by improved instructions. On the contrary, the main effect of reward is evidence of a model-free strategy. As can be seen in Figure 1, we failed to replicate the results from da Silva & Hare (2020). More specifically, the coefficient value of the reward × transition interaction was substantially lower in our Replica condition than in the original study by da Silva & Hare (2020), as shown by an independent samples t-test ($t(81)=-2.3828$, $p=0.0195$, two-tailed, $d=-0.5553$, 95% CI [-0.7819, -0.0703]). We did not find significant differences between both studies in any of the other coefficients (Intercept: $t(81)=-0.7158$, $p=0.4761$, two-tailed, $d=-0.1723$, 95% CI [-0.4636,0.2183]; Reward: $t(81)=1.3664$, $p=0.1756$, two-tailed, $d=0.3417$, 95% CI [-0.0496, 0.2671]; Transition: $t(81)=-0.4145$, $p=0.6796$, two-tailed, $d=-0.0889$, 95% CI [-0.2892, 0.1895]). On the other hand, the comparison of our Replica and Fixed-Locations studies revealed stronger evidence of model-free behaviour in the latter, as evidenced in the Reward parameter, which was larger in Fixed-Locations than in Replica ($t(116)=-1.7748$, $p=0.0393$, one-tailed, $d=-0.3268$, 95% CI ($-\infty$, -0.0088]). No significant differences between the studies were found in the rest of the parameters (Intercept: $t(116)=-1.2335$, $p=0.2199$, two-tailed, $d=-0.2271$, 95% CI [-0.4162, 0.0967]; Transition: $t(116)=-0.7980$, $p=0.7868$, one-tailed, $d=-0.1469$, 95% CI [-0.2074, ∞); Reward × Transition: $t(116)=-0.5091$, $p=0.6942$, one-tailed, $d=-0.0937$, 95% CI [-0.3000, ∞))[1].

[1] When we designed our research, we expected participants in the Replica study to show solely model-based behaviour, replicating the results from da Silva & Hare (2020). This is acknowledged in the pre-registration protocol of this work (https://osf.io/x9sya). Therefore, we did not expect any differences between coefficients in any direction, and all the analyses in this regard are consequently two-tailed. However, when we compared the Replica study with the Fixed-Locations one, we did expect the Fixed-Locations study to detect a model-free component not present in the Replica study. Therefore, the *t*-test for the Reward coefficient is one-tailed. We also performed one-tailed tests on the Transition and Reward

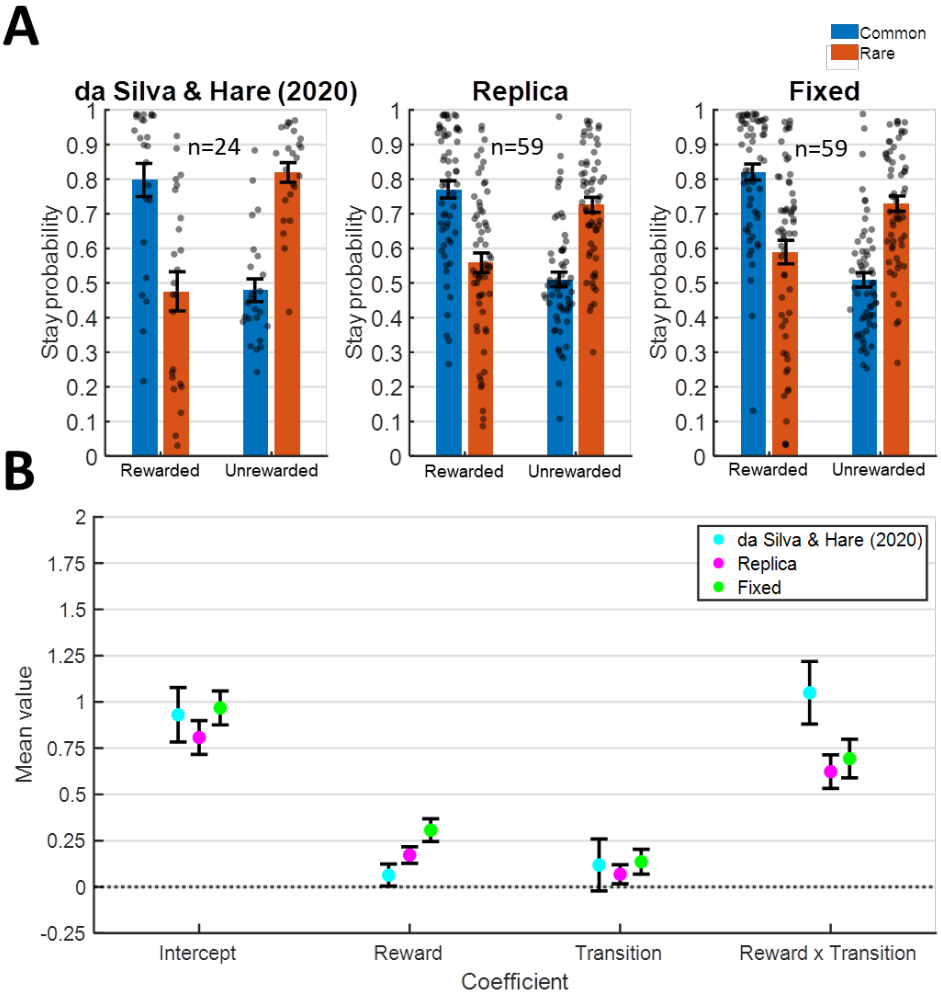

Figure 1. Results of the logistic regression analysis. **A.** Stay probabilities (probability of repeating the same response as in the previous trial) are shown in the cases when the transition in the previous trial had been Common (blue) or Rare (red). Results further distinguish whether the previous trial had been Rewarded or Unrewarded. The left panel shows the results from the original magic carpet experiment in da Silva & Hare (2020), the middle panel shows the results for the Replica condition and the right panel shows them for the Fixed-Locations condition. Individual results are shown as well as the mean ± SEM. **B.** Coefficients for each of the logistic regression parameters, which were used to calculate the stay probabilities shown in the upper panels. The mean ± SEM is depicted

It could be argued that despite our use of improved instructions, some participants might still have failed to understand the structure of the task. Consequently, their

× Transition coefficients. This is because we expected a larger model-based component in the Replica study than in the Fixed-Locations one. The same logic was applied to all subsequent analyses.

behaviour under an inaccurate cognitive map of the two-stage task may have biased our results towards a model-free component. To address this possibility, participants were asked to complete a questionnaire at the end of the task (See the "Materials and Methods" section, "*Procedure*", and Supplementary material, S3 Appendix). To rule out the possibility that our results are biased by the inclusion of participants who did not understand the instructions, in the Supplementary Material (see S1 Appendix) we report additional logistic regression analyses excluding participants who failed any question in any of the questionnaires. We apply this same exclusion criterion to the participants in the experiment by da Silva & Hare (2020). The exclusion of these participants did not make a meaningful difference in the results.

Hybrid reinforcement learning model fits

To analyze the extent to which participants showed model-based vs model-free behaviour in the two-stage task, we fitted the standard hybrid reinforcement learning model proposed by Daw et al. (2011) to their data. This model combines the model-free SARSA (λ) algorithm with a model-based learning algorithm, weighted by parameter w ($0 \leq w \leq 1$). This parameter can be interpreted as a model-based weight, with a value of 1 indicating the use of purely model-based strategies and 0 indicating a sole model-free behaviour. Figure 2 shows the estimated w weights for the magic carpet experiment in da Silva & Hare (2020) as well as our the Replica and Fixed-Locations studies. Importantly, our Replica study failed to replicate the results from da Silva & Hare (2020) (w: $t(81)=3.1706$, $p=0.0021$, two-tailed, $d=0.8209$, 95% CI [0.0941, 0.4112]). The model suggests a hybrid model-free/model-based behaviour, rather than a sole model-based one, despite the use of improved instructions. In addition, in consonance with the hypothesis that the Fixed-Locations condition would promote the use of model-free strategies, the model-based weight in this case was slightly lower than in the Replica condition, indicating a somewhat larger model-free behaviour. However, the difference between both conditions is negligible and fails to reach statistical significance (w: $t(116)=0.1379$, $p=0.4453$, one-tailed, $d=0.0269$, 95% CI [-0.1014, ∞)).

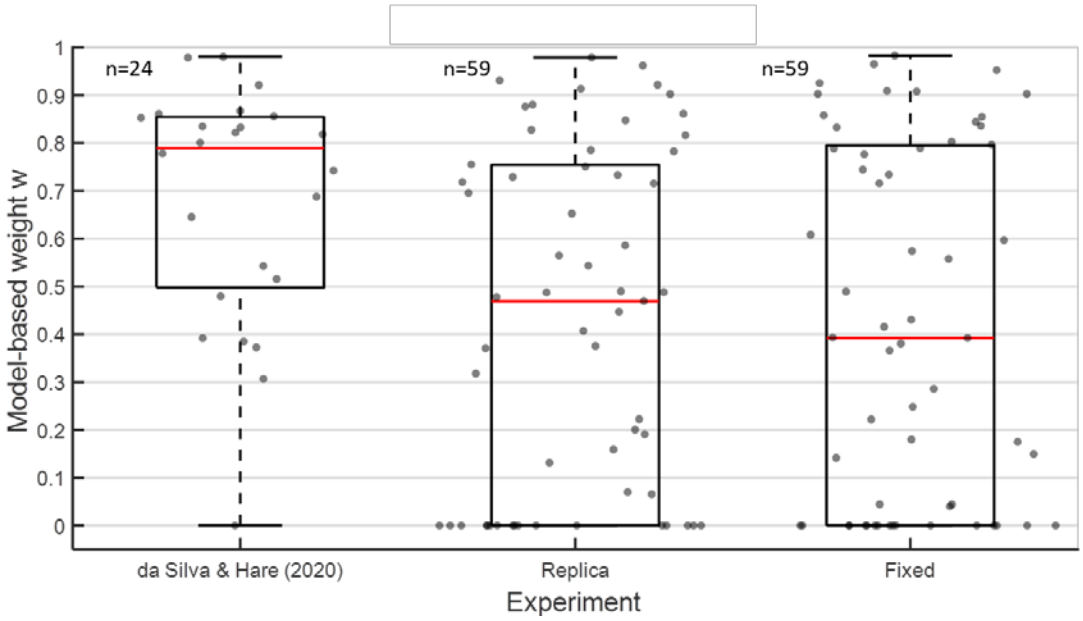

Figure 2. 25%, 50% (median), and 75% percentiles of estimated parameters from the hybrid reinforcement learning algorithm as well as individual estimates. Data is shown for the *w* model-based weight parameter.

Differences between both groups in model-free behaviour can be detected in the logistic regression analyses (i.e., in the the reward coefficient) but not in the hybrid reinforcement learning model fits (i.e., in the *w* parameter). This may be the consequence of the single *w* parameter not being as sensitive as the multiple logistic regression coefficients for discriminating model-free vs model-based behaviour. In line with this interpretation, da Silva & Hare (2020) concluded that the logistic regression model is better than the hybrid model at explaining first-stage choices in this task.

Discussion

The present study attempted to replicate the results by da Silva & Hare (2020), who found that, when provided with improved instructions, participants employed exclusively model-based strategies during the two-stage task. However, we failed to provide favourable evidence of pure mode-based behaviour when participants are provided with their improved instructions. In agreement with classical results (Daw et al., 2011; Decker et al., 2016), our data suggests that participants behave according to a hybrid model, employing both model-free and model-based strategies. Importantly, we

recruited a larger sample (n=59) than da Silva & Hare (2020) (Magic carpet task: n=24, Spaceship task: n=21), allowing for larger statistical power.

One may wonder whether our participants did not reach a good understanding of the improved instructions. This would have led them to build an incorrect model of the task, and thus it could be argued that the model-free component observed in our analyses could perhaps be the consequence of incorrect models, and not genuinely model-free computations. However, eliminating from the analyses any participants who could potentially have not correctly understood the task does not alter our main results (See Supplementary Material, S1 Appendix). In addition, like in da Silva and Hare (2020), the vast majority of our participants had a good understanding of the task (i.e. 1 participant out of n=24 [0.04%] was removed from the magic carpet study by da Silva & Hare (2020), and 6 participants out of 59 [1%] were removed from our Replica study).

We also conducted a different experimental group fixing the locations of response options, in contrast to the classical two-stage task, where response options swap positions randomly across trials (e.g. Daw et al., 2011; and also da Silva & Hare, 2020). Our logistic regression analysis suggests that fixing option locations induces a more pronounced model-free component. In other words, this manipulation triggered habitual processes, facilitating the association between specific stimuli and specific motor commands. In agreement with our results, using a different task, but manipulating the consistency of response mappings, Molinero et al. (2021) found that reward-related cognitive prioritization was stronger when a constant response pattern was kept. Also, Hardwick, Forrence, Krakauer & Haith (2019) found that habits were produced when specific S-R associations were always executed with the same motor command. Additionally, Neal, Wood, Wu & Kurlander (2011) provided evidence that previously formed habits disappeared when the response pattern was changed.

Thus, our results suggest that the model-free marker, as measured by the two-stage task, seems to be related to S-R learning. This result is important because it has been questioned whether the two-step task really taps into the habitual component of behaviour. For instance, if so, then this marker should increase with training, and not the contrary—and that was the result found in the only study that has manipulated the amount of training using the two-stage task to date (Economides et al., 2015).

Moreover, there is empirical evidence showing that such model-free parameter does not correlate with habit strength measured by the canonical outcome devaluation test (Gillan et al., 2015). To this evidence, we should add the da Silva & Hare, 2020 study itself. All these results apparently conflict with our suggestion that we were measuring, to some extent, habit strength using the two-stage task. We would argue that the conflict can be explained. As we show in our study, to tap into the functioning of the habit/model-free system it is essential that the motor response remains the same from trial to trial, given the same discriminative stimulus. This aspect of the design was not present in da Silva & Hare's (2020) study. Also, the null result from Gillan et al., 2015 could be produced because their habit task was insensitive to the functioning of the habit system; indeed, other learning tasks previously used for studying habits have shown a lack of sensitivity for detecting them when they were further tested (Buabang et al., 2022; de Wit et al., 2018; de Houwer et al., 2018). Economides et al. (2015) found that extended training promoted model-based behaviour. Because they did not use the improved instructions as in da Silva & Hare (2020), it seems reasonable that a number of participants started with a wrong model of the task. Thus, extended training allowed these participants to learn and apply the correct model; hence, for these participants, model-based behavior could be only available at the end of training. Learning the correct model of the task through training might overshadow the effect of S-R learning on participants' behaviour. Future research should investigate the effect of the amount of S-R learning on model-free parameters in a task with improved instructions—ideally by manipulating both factors' instructions (classic vs improved) and the amount of learning (little vs extended).

Our results concern to habitual responses thought as specific motor patterns that are activated after an S. It is important to note that that is not the only conceptualization of the "R" in S-R habits. For instance, Gadner and colleagues understand these responses as "impulses to act" whereas the act itself can change from instance to instance (e.g., Gardner et al., 2015). Our conception is different and concerns low level specific motor commands to achieve a certain goal (e.g., Du et al., 2022; Yang et al., 2022)

To sum up, the present study converges with the conclusions of previous research showing the widespread presence of hybrid model-based/model-free strategies in the two-stage task. Importantly, such hybrid behaviour can still be observed even after ensuring that participants do not have an incorrect model of the task. In addition, we

found that model-free behaviour can be promoted through invariably linking specific stimuli to certain motor commands, providing evidence that the model-free marker measured by the two-stage task is linked to S-R learning.

Materials and Methods

Pre-registration

The methods and analysis plan employed in this study were pre-registered. The pre-registered procotol is publicly available at https://osf.io/x9sya

Participants

Following the indications by Brysbaert. (2019) about minimum sample size in psychological research, we set our minimum sample to be one-hundred participants to ensure properly inter-group comparisons in our experiments. In total, one-hundred-and-eighteen participants from the Autonomous University of Madrid (UAM) were randomly assigned to the Replica condition (9 males, mean age: 22.29 years $\pm$ 5.54 SD; 50 females, mean age: 20.04 years $\pm$ 1.53 SD) or to the Fixed-Locations condition (9 males, mean age: 19.35 years $\pm$ 0.86 SD; 50 females, mean age: 20.35 years $\pm$ 2.13 SD). The best 3 participants in each group obtaining the largest scores in the task received 25€. Procedures were approved by the UAM ethics committee, and participants signed an informed consent before taking part in the experiment and were treated in accordance with the Helsinky declaration. All of them had normal vision or vision corrected to normality.

Apparatus

Participants were tested in individual cubicles, each with a standard PC and a monitor. Stimuli were presented using MATLAB with Psychtoolbox extensions (Brainard, 1997; Pelli, 1997; Kleiner, Brainard, & Pelli, 2007). Responses were collected using custom keyboards.

Task

Despite a few, but significant differences, a similar two-stage task was employed in the Replica and Fixed-Locations studies. The task, with its common features between studies will now be described, and whenever a difference between them exists, this will be made explicit.

The experimental task replicated that of da Silva & Hare (2020), which in turn was similar to Daw et al. (2011), except for some minor changes. As in da Silva and Hare (2020), the task was supported by a cover story causally explaining each transition and nuisance, so that a good understanding of the structure of the task was ensured (see the "Magic carpet task description" section in da Silva & Hare, 2020). Participants played the role of musicians living in a fantasy land, and obtained gold coins by playing the flute for an audience of genies living inside magic lamps in two different mountains, the Blue and the Pink mountain. Each mountain held two genies and participants were told that each lamp, with each genie inside, had a symbol written (Tibetan character) with the genie's name in the local language. When participants arrived at a given mountain they had to choose a lamp, pick it and rub it. If the genie inside was in the mood for music then he would come out, listen to a song and give a gold coin to the musician. On such occasions, a genie with a coin was displayed on top of the lamp just chosen for 1,5 sec. Otherwise, a crossed "0" was displayed also for 1,5 sec. Participants were told, however, that the genies' interest in music might change over time. Furthermore, in the Replica sutudy, they were told that the lamps in each the Blue and Pink mountains could swap their positions between visits to a mountain. This was because every time they picked a lamp, they might leave it in a different place. In the Fixed-Locations study, participants were not told anything in this regard. This is because lamps' positions were fixed across trials (positions were counterbalanced across participants).

To go to a certain mountain, participants had to choose between two magic carpets which would bring them there. The carpets had previously been enchanted by a magician so that each would fly to a different mountain. They had symbols written on them in the local language that meant "Blue Mountain" or "Pink Mountain", depending on the destination of each magic carpet. Normally, carpets flew to their destination (common transitions). However, on rare occasions (rare transitions), travelling to the mountain of destination was too dangerous due to strong winds happening there. On

such rare occasions, magic carpets were forced to land in the remaining mountain. Once more, in the Replica study, carpets could change sides between trials due to musicians putting them down and unrolling them on a different side of the room. On the contrary, in the Fixed-Locations study carpets remained in the same position across trials (positions were counterbalanced across participants). No specification on the location of the carpets was given during instructions.

In short, the task had a general main structure consisting of two stages happening in each trial (See Figure 3). In the first stage, participants needed to choose between two magic carpets that took them to either of two mountains (second stage). 70% of the times, a given carpet would take the participant to its assigned destination (common transition). However, the remaining 30% of the times, the carpet would bring the participant to its non-assigned mountain of destination (rare transition). Which carpet most probably flew to which mountain was randomized across participants. The position of the carpets (left or right) in the first stage changed randomly across trials in the Replica study and remained fixed in the Fixed-Locations study. In the second stage, participants were presented with two more options or states (lamps) and needed to choose one. These also changed their position (left or right) randomly across trials in the Replica study and remained fixed in the Fixed-Locations study. Finally, each state had a reward probability that varied between trials through a Gaussian random walk (mean 0, SD .025; with reflecting boundaries at 0.25 and 0.75) so that ongoing learning was encouraged. A pool of 20 Gaussian random walks was generated out of which, for each subject, 4 different random walks were selected at random to represent the reward probabilities of the total 4 second-stage state options of the 2 possible mountains in the second stage.

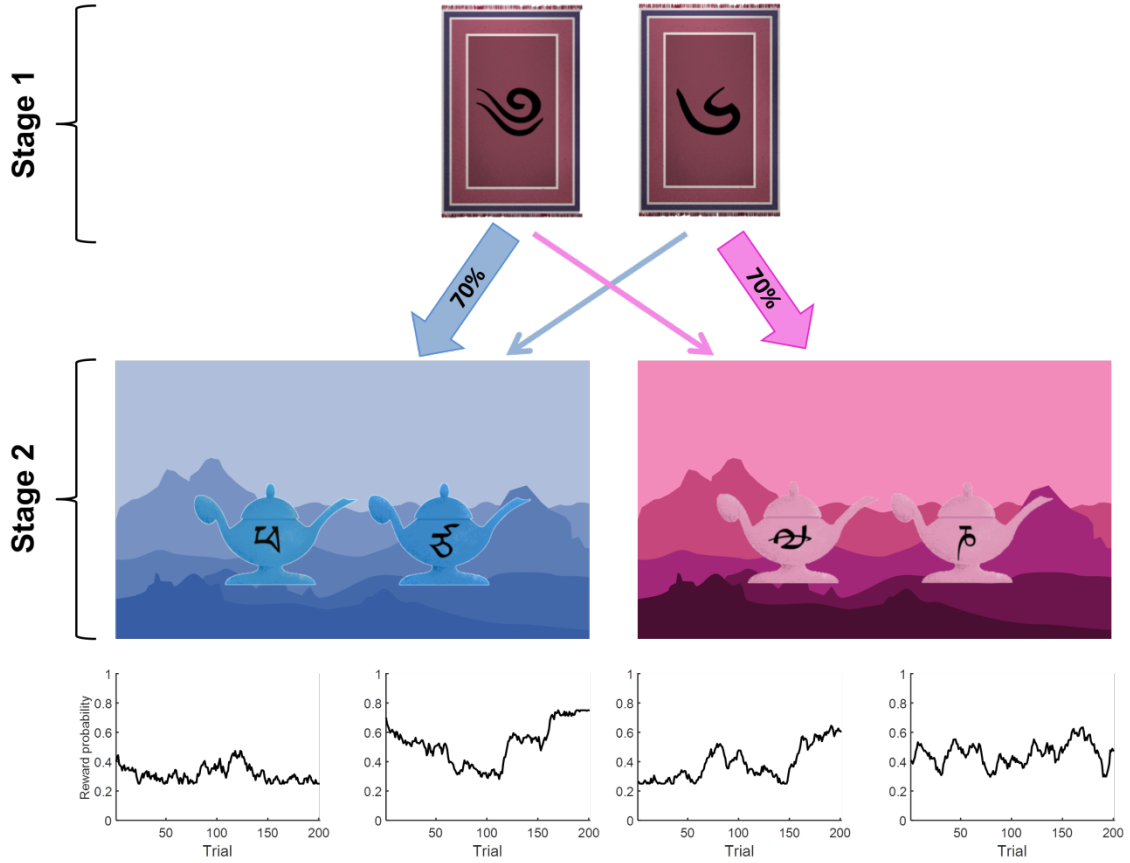

Figure 3. Two-stage task structure. During the first-stage state, participants need to choose between two options (magic carpets) that will bring them to two possible different states (Blue mountain [left] and Pink mountain [right]) in stage two. Transitions to second-stage states are probabilistic. A given magic carpet will transition to a given second-stage state with 70% probability, and it will transition to the remaining one with 30% probability. Once in a second-stage state, participants need to choose between two options (lamps). Each option has a reward probability that changes throughout trials by means of a Gaussian random walk (lower panels). Pictures of magic carpets and lamps are taken from da Silva & Hare (2020), and used in the studies of the present manuscript.

Participants were asked to always use the same finger for each response (left or right). That is, options on the left were selected using the left index finger, and options on the right were selected with the right index finger.

Procedure

The task consisted of 201 trials which were run along three blocks of 67 trials each. Participants were allowed to take a break in between blocks.

Choices in each stage were recorded, as well as response times (RTs). Participants were told that magic carpets in the first stage had to be chosen in less than 2 sec or else they would fly without them. In either case, lamps at the second stage had to be rubbed within 2 sec or the genies inside them would fall asleep and not come out. Trials in which participants failed to enter a response within 2 sec were be aborted with a message displayed on the screen: “TOO LATE! The magic carpets have flown without you” (for 7,5 sec) or “TOO LATE! The genies have fallen asleep” (for 1,5 sec). The duration of each message was such that the time spent was similar to the one that would have been spent if trials had not been aborted. We randomly selected the inter-trial interval from a uniform distribution ranging from 0.7 to 1.3 sec. During such interval, a Gaussian random noise mask (mean 0, SD 0.5) was presented to prevent possible visual aftereffects.

Previous to performing the task, participants completed 50 tutorial random flights. This was intended to make them aware of which transitions were common and which rare, and, in general, they became familiar with the narrative of the game and how to proceed in it. The only difference between the Replica and the Fixed-Locations group was that in the former participants were told that items may change places across trials (with an explanation of why that may happen) and in the latter they were not. During tutorial flights, participants were presented for 1 sec with a transition screen that made explicit to which mountain a carpet was flying. In addition, that screen showed them whether a carpet was flying to its mountain of destination (common transition) or was being flown away from it (rare transition). However, during the non-tutorial trials and because magic carpets were self-driving, musicians took a nap aboard it and only woke up upon arrival. A black screen was displayed during this period for 1 sec (not explicitly showing the mountain of destination and whether a transition had been common or not). Therefore, participants had to figure out the meaning of the symbols in each carpet for themselves. It is important to note that the mountains to which magic carpets flew during tutorial flights were different from the ones to which they flew during the main task. Namely, during tutorial flights, and in order not to interfere with the forthcoming task, magic carpets flew to the Black and Red mountains instead of the Blue and Pink mountains, where magic carpets flew during the main task. The magic carpets flying to each mountain and the lamps at the second stage used different Tibetan symbols from the

ones used during the main task. A different pool of 20 Gaussian random walks for the second-stage state options' reward probabilities was used for tutorial flights.

As explained above, the positions of the first and second-stage state options were randomized across trials in the Replica group and kept constant in the Fixed-Locations group but randomized across participants. The most likely transition through which each carpet flew to each mountain during tutorial flights was also randomized across participants.

Figure 4 shows the appearance and timing of the two-stage task both during tutorial flights (Figure 4A) and during the main task (Figure 4B).

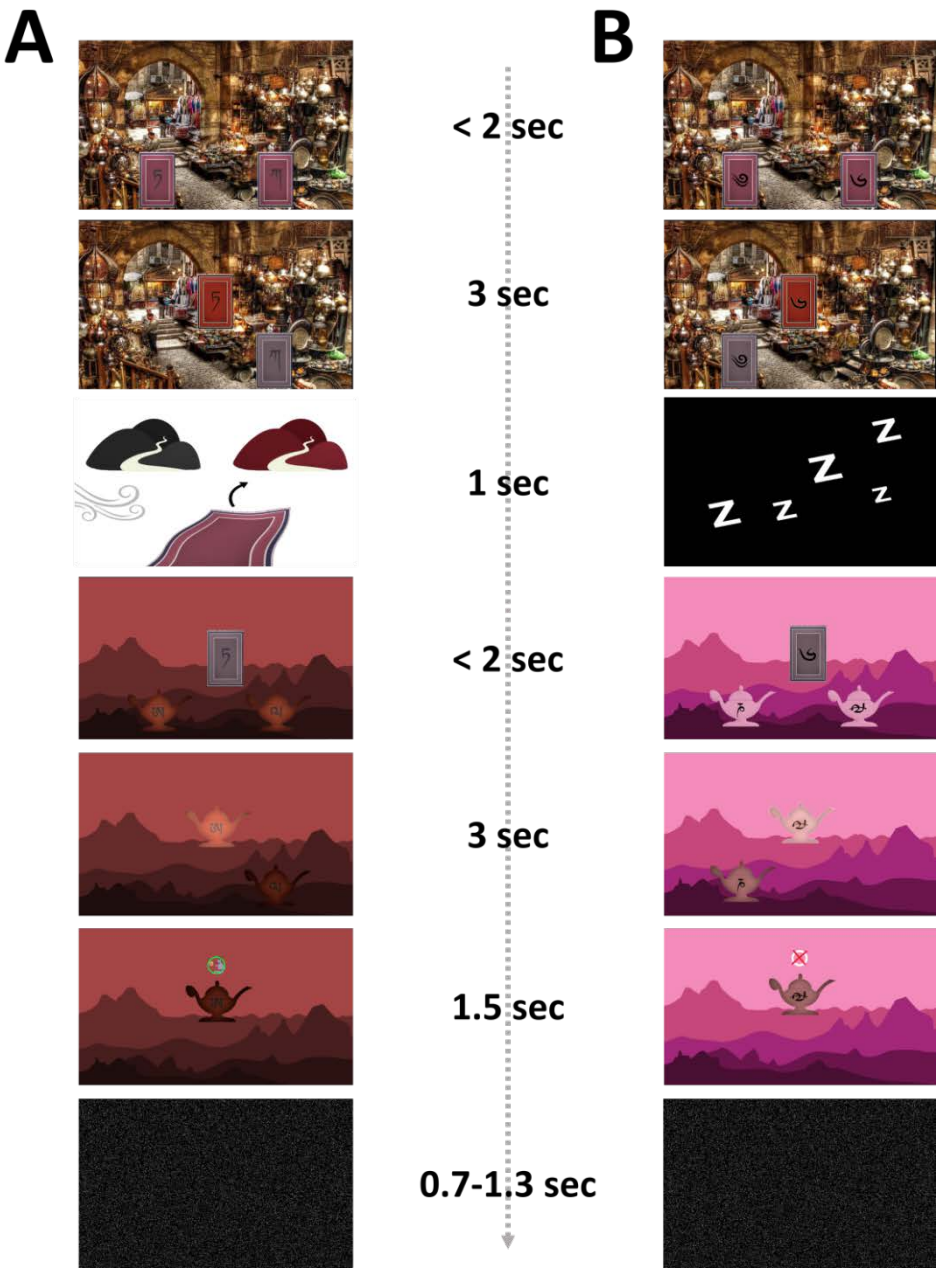

Figure 4. Appearance of the two-stage task and timing of events for tutorial flights (**A**) and the main task (**B**). First, one of two carpets needs to be selected in less than 2 sec, then the choice made is displayed for 3 sec. Afterwards, the transition to a given mountain is made (1 sec). In the case of **A**, such transition is shown to be a rare one. In **B**, this is not explicitly shown as the musician is taking a nap during the flight. Afterwards, once at a given mountain, a lamp needs to be chosen out of two different ones in less than 2 sec. Then the choice is displayed for 3 sec. Afterwards, a reward may be given (**A**) or not (**B**) depending on the interest in music at that moment of the genie inside the chosen lamp. The reward/non-reward stimulus is displayed for 1.5 sec. Finally, and right before the next trial, a Gaussian random noise mask is shown for 0.7-1.3 sec to prevent any visual aftereffects. (Note that during the main task musicians fly to the Blue and Pink mountains. However, during tutorial flights musicians fly to the Black and Red mountains) Pictures of magic carpets, lamps, genies and transition screens to a mountain are taken from da Silva & Hare (2020), and used in the studies of the present manuscript.

Importantly, before the tutorial flights, participants carefully read detailed instructions about them and completed a questionnaire with specific aspects about the task (See Supplementary material, S3 Appendix). Wrong answers received feedback with the correct answer, making sure that participants did not start the tutorial flights without having understood the task. Also, after the 201 trials in the main task, participants were asked the following questions:

1) For each carpet symbol: What was the meaning of the symbol?

2) How difficult was the game? a) very easy, b) easy, c) average, d) difficult e) very difficult.

Analyses

First, in both experimental conditions (Replica and Fixed-Locations), trials in which participants had failed to enter a response within 2 sec were omitted (Replica: mean: 3.47 trials $\pm$ 5.87 SD; Fixed-Locations: 3.54 trials $\pm$ 4.26 SD). The pre-registration of the present study (https://osf.io/x9sya) specified that participants whose response time median absolute deviation (MAD) (Leys et al., 2013) was 3 points or larger would be excluded from analyses. The analyses reported in the main text do not remove participants based on this criterion. However, the reader can find analyses excluding them in the Supplementary Material, S2 Appendix. The reason why we do not exclude these participants in the main manuscript is because this allows for a larger statistical power. Additionally, removing them does not significantly alter the results.

Logistic regression analysis

A logistic regression analysis of consecutive trial pairs was performed separately for each participant in the Replica and Fixed-Locations studies. Trial pairs including a trial performed after a break during the task were excluded from analyses. The stay probability (i.e. the probability of repeating the same first-stage action as in the previous trial) was predicted as a function of two variables: reward (was the participant rewarded on the previous trial or not?) and transition (was the previous trial's transition common or rare?) using the following equation:

$$p_{stay} = \frac{1}{1 + e^{-(\beta_0 + \beta_r x_r + \beta_t x_t + \beta_{r \times t} x_r x_t)}} \qquad (1)$$

where β_0, β_r, β_t and $\beta_{r\times t}$ are, respectively, the coefficients for the intercept and the reward, transition and reward × transition effects. x_r adopted values of +1 or -1 depending on whether the previous trial had been rewarded or not, and x_t adopted values of +1 or -1 depending on whether the previous trial had had a common or a rare transition. The model was fit to each individual subject using the Matlab function "fitglm" in a way that coefficients were obtained for every participant.

A few participants in each experimental group (i.e., 6 in Replica and 6 in Fixed-Locations) provided responses in a very consistent manner. For instance, a subject may unequivocally choose the same first-stage state option when the previous trial had been rewarded and the transition had been common. In this scenario, perfect separation between classes occurs, making it impossible for Iteratively Reweighted Least Squares methods (as used by the Matlab function "fitglm") to estimate parameter values. In these cases, we artificially changed at random only one of their choices producing a perfectly unequivocal pattern. With such consistent pattern broken, parameter estimation was made possible. We preferred this to remove participants where perfect separation of classes took place, as their cases were still informative about performance in the two-stage task. After all, this task may encourage such consistent patterns of behaviour.

Hybrid reinforcement learning model fits

The standard hybrid reinforcement learning proposed by Daw et al. (2011), combining model-based learning with the model-free SARSA(λ) algorithm, was fitted to the empirical data of each individual participant taking part in the study.

At initiation (i.e., trial $t = 1$), the model-free (MF) values of the algorithm, $Q_{t=1}^{MF}(s,a)$ are set to zero. That is, $Q_{t=1}^{MF}(s,a)$ for each possible action a that an agent can perform in each stage, s, is 0. Once an action is chosen at the end of trial t, the $Q_t^{MF}(s,a)$ value for that action performed at a certain stage is updated. In the particular case of second-stage actions, a_2, performed in a second-stage state, s_2, (i.e. Pink and Blue mountains in Figure 3), $Q_t^{MF}(s_2,a_2)$, is updated through the following formula:

$$Q_{t+1}^{MF}(s_2,a_2) = Q_t^{MF}(s_2,a_2) + \alpha_2\delta_t^2 \qquad (2)$$

were $\alpha_2$stands for the learning rate for the second stage ($0 < \alpha_2 > 1$) and δ_t^2 is the reward prediction error; namely, the current value of the action chosen, $Q_t^{MF}(s_2, a_2)$, and the reward received, r_t (0 or 1), and is defined as follows:

$$\delta_t^2 = [r_t - Q_t^{MF}(s_2, a_2)] \qquad (3)$$

Regarding the chosen first-stage action, a_1, at the first stage state s_1, the value of the chosen action, $Q_t^{MF}(s_1, a_1)$, is updated as follows:

$$Q_{t+1}^{MF}(s_1, a_1) = Q_t^{MF}(s_1, a_1) + \alpha_1 \delta_t^1 + \alpha_1 \lambda \delta_t^2 \qquad (4)$$

where δ_t^1 is the reward prediction error for the first stage, and is defined as:

$$\delta_t^1 = Q_t^{MF}(s_2, a_2) - Q_t^{MF}(s_1, a_1) \qquad (5)$$

α_1 is the learning rate for the second stage ($0 < \alpha_2 > 1$) and λ is the eligibility parameter ($0 < \lambda > 1$). This last parameter weights the effect of second-stage reward prediction error on first-stage action values.

Having explained the model-free (MF) values of the algorithm, we may now explain its model-based (MB) values. $Q_t^{MB}(s_2, a_2)$ for action a_2 at second-stage state s_2 has the same meaning as the corresponding model-free value: $Q_t^{MB}(s_2, a_2) = Q_t^{MF}(s_2, a_2)$. On the other hand, for each first-stage action, model-based values are calculated as follows:

$$Q_t^{MB}(s_1, a_1) = \sum_{s_2 \epsilon S} P(s_2 \mid s_1, a_1) max_{a_2 \epsilon A} Q_t^{MB}(s_2, a_2) \qquad (6)$$

That is, model-based values for first-stage actions are computed when a decision is made from the values of second-stage actions, where $P(s_2 \mid s_1, a_1)$ stands for transition probability to state s_2 through first-stage action, $s_1.S = \{pink, blue\}$ designates the possible second-stage states, and A designates the possible actions at those states.

Agents perform first-stage choices both based on model-free and model-based values according to a softmax distribution:

$$P_t(s_1 \mid a_1) = \frac{e^{\beta_1 [w Q_t^{MB}(s_1, a_1) + (1-w) Q_t^{MF}(s_1, a_1) + p \times rep_t(a_1)]}}{\sum_{a' \epsilon A} e^{\beta_1 [w Q_t^{MB}(s_1, a') + (1-w) Q_t^{MF}(s_1, a') + p \times rep_t(a')]}} \qquad (7)$$

where w is a model-based weight whose value determines the amount of model-based influence ($0 < w > 1$). $\beta_1$is the inverse temperature parameter for the first stage, and it

models the exploration-exploitation trade-off during that stage. p is a perseveration parameter whose value has an effect on how prone agents are to repeating the previous trial's first stage action in the next trial. Finally, $rep_t(a')$ is a value defined as 1 if the first-stage action a' was performed in the previous trial (0 otherwise).

When it comes to the second stage, the probability of a given second-stage choice is computed as follows:

$$P_t(s_2 \mid a_2) = \frac{e^{\beta_2 Q_t(s_2,a_2)}}{\sum_{a' \epsilon A} e^{\beta_2 Q_t(s_2,a')}} \qquad (8)$$

where model-free and model-based values for the corresponding second-stage actions are the same. This is because no tendency to repeat the previous action or keypress is assumed.

Estimates for model parameters, α_1, α_2, β_1, β_2, w and p, were obtained through maximum likelihood estimation. To this end, participants' first-stage and second-stage responses as well as the transitions (common vs rare) that happened in each trial, together with the reward obtained were fed into the algorithm. In short, the reinforcement learning algorithm performed the same task as participants in a way that maximized the negative log-likelihood $-\log[P_t(s_1 \mid a_1)]$ to achieve each subject's parameter values. For all participants, this process was repeated throughout 1000 iterations. As in da Silva & Hare (2020), the model was coded in the Stan modelling language (Stan Development Team, 2012; Carpenter et al., 2017), and was further fitted to each subject's data using the cmdstanpy library.

References

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

Brysbaert, M. 2019 How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition, 2*(1): 16, pp. 1–38.

Buabang, E. K., Köster, M., Boddez, Y., Van Dessel, P., De Houwer, J., & Moors, A. (2022). A goal-directed account of action slips: The reliance on old contingencies. Journal of Experimental Psychology: General.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76(1).

da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053-1066.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. Psychological science, 27(6), 848-858.

De Houwer, J., Tanaka, A., Moors, A., & Tibboel, H. (2018). Kicking the habit: Why evidence for habits in humans might be overestimated. Motivation Science, 4(1), 50.

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. Journal of Experimental Psychology: General, 147(7), 1043.

Du, Y., Krakauer, J. W., & Haith, A. M. (2022). The relationship between habits and motor skills in humans. Trends in Cognitive Sciences.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312-325.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., & Dolan, R. J. (2015). Model-based reasoning in humans becomes automatic with training. *PLoS computational biology*, *11*(9), e1004463.

Gardner, B. (2015). A review and analysis of the use of 'habit'in understanding, predicting and influencing health-related behaviour. *Health psychology review*, *9*(3), 277-295.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523-536.

Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time-dependent competition between goal-directed and habitual response preparation. *Nature human behaviour*, *3*(12), 1252-1262.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception*, 36 (ECVP Abstract Supplement).

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off?. *PLoS computational biology*, *12*(8), e1005090.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, *28*(9), 1321-1333.

Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of cognitive neuroscience*, *30*(10), 1391-1404.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology, 49*(4), 764-766.

Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. E. (2020). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology: General, 149*(8), 1449-1459.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature neuroscience, 20*(9), 1269-1276.

Molinero, S., Giménez-Fernández, T., López, F. J., Carretié, L., & Luque, D. (2021). Stimulus–response learning and expected reward value enhance stimulus cognitive processing: An ERP study. *Psychophysiology, 58*(5), e13795.

Neal, D. T., Wood, W., Wu, M., & Kurlander, D. (2011). The pull of the past: When do habits persist despite conflict with motives?. *Personality and Social Psychology Bulletin*, *37*(11), 1428-1437.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442.

Sutton, R. S., & Barto, A. G. (1998). A bradford book. In *Reinforcement learning: An introduction*. The MIT Press.

Stan Development Team. (2012). Stan Modeling Language User's Guide and Reference Manual, Version 1.0.

Verleger, R., Grauhan, N., & Śmigasiewicz, K. (2016). Effects of response delays and of unknown stimulus-response mappings on the oddball effect on P3. *Psychophysiology, 53*(12), 1858-1869.

Verleger, R., Keppeler, M., Sassenhagen, J., & Śmigasiewicz, K. (2018). The oddball effect on P3 disappears when feature relevance or feature-response mappings are unknown. *Experimental Brain Research, 236*(10), 2781–2796.

Yang, C. S., Cowan, N. J., & Haith, A. M. (2022). Control becomes habitual early on when learning a novel motor skill. bioRxiv.

Funding

This work has received funding from grant PGC2018-094694-B-I00 (MCIN/AEI), grant PID2020-118583GB-I00 (MCIN/AEI), grant PID2021-126767NB-I00 (MCIN/AEI) and grant PROYEXCEL_00287, funded by the Junta de Andalucía. RL is supported by a Juan de la Cierva-Formación fellowship (FJC2020-044084-I) MCIN/AEI /10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Acknowledgements

We would like to thank Carolina Feher da Silva for her help providing the code for reinforcement-learning model fitting analysis. As well, we would like to thank Jonathan Wilson for making available Matlab code of the two-stage task in Github. Our version of the task took code from him.

Competing interests

The authors declare that the research was conducted in the absence of any competing interest.