Cognitive, Affective, & Behavioral Neuroscience
https://doi.org/10.3758/s13415-025-01350-9

RESEARCH ARTICLE

# When models matter: Environmental demand guides the arbitration between model-based and model-free control

**Leslie K. Held[1] · Elise Lesage[1] · Wouter Kool[2] · Senne Braem[1]**

Received: 19 March 2025 / Accepted: 25 August 2025


**Abstract**
As humans, we often repeat previously rewarded actions without thinking, but we also possess the ability to plan ahead and simulate actions based on an internal model of the environment. These two types of control are commonly conceptualized as model-free versus model-based control. While there is a body of research on interindividual differences in using either strategy, we aimed to test whether people can learn to regulate which strategy to use based on environmental demand. We used a two-stage decision-making task where participants tracked the drifting rewards associated with two second-stage states. Each trial started with one of two possible first-stage states, each offering two choices that deterministically led to one of the second-stage states. Successful generalization between first-stage options indicated model-based control, while mere repetition of previously rewarded choices reflected model-free behavior. We manipulated how often participants ($n = 140$) were exposed to alternations versus repetitions of first-stage states. When these states frequently repeat, there is a reduced need to consult the transition structure, because it pays off to adopt model-free control and simply retake previously rewarded actions. Conversely, when first-stage states frequently alternate, it is more beneficial to adopt model-based control, considering the transition structure and generalizing reward outcomes between them. In line with our hypothesis, we show that participants exposed to more first-stage state alternations were more model-based in a test phase than participants exposed to more first-stage state repetitions. These findings suggest that people learn to arbitrate between different reinforcement-learning strategies consistent with a cost–benefit analysis sensitive to environmental demands.

**Keywords** Reinforcement learning · Model-based · Model-free · Dual-system RL · Two-step task

## Introduction

When an action leads to a desired outcome, we tend to repeat it almost automatically, yet we can quickly adapt our behavior to changing circumstances when needed. For example, we can almost blindly take the same route to work daily but also plan a detour if a traffic jam is announced. The distinction between model-free and model-based reinforcement learning provides a mathematical formalization of these adaptive features of human behavior (Daw et al., 2005, 2011). While model-free choices rely on the reward history of previous actions, directly reinforcing those leading to success, model-based control guides behavior by planning based on an internal causal model of the environment (Dolan & Dayan, 2013; Doll et al., 2015; Gläscher et al., 2010). Extensive research has investigated their neural correlates, predominantly linking distinct activity in the ventral striatum to model-free and model-based control and activity in executive control regions, such as the prefrontal and cingulate cortex, to model-based control (see Huang et al., 2020 for a whole-brain meta-analysis). Both strategies trade off against each other, as model-free control is more efficient and cognitively cheap, and model-based control is more flexible but computationally more demanding (Kool et al., 2017, 2018a, b). However, little is known about how people learn to navigate this trade-off based on environmental demands. Therefore, inspired by previous theories and studies on the learning of cognitive control (Abrahamse et al., 2016; Braem & Egner, 2018; Braem et al., 2024; Doebel, 2020), we aimed to test whether people can dynamically learn to adjust their reliance on model-based versus model-free

✉ Leslie K. Held
leslie.held@ugent.be

[1] Department of Experimental Psychology, Ghent University, Ghent, Belgium

[2] Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

Published online: 13 October 2025

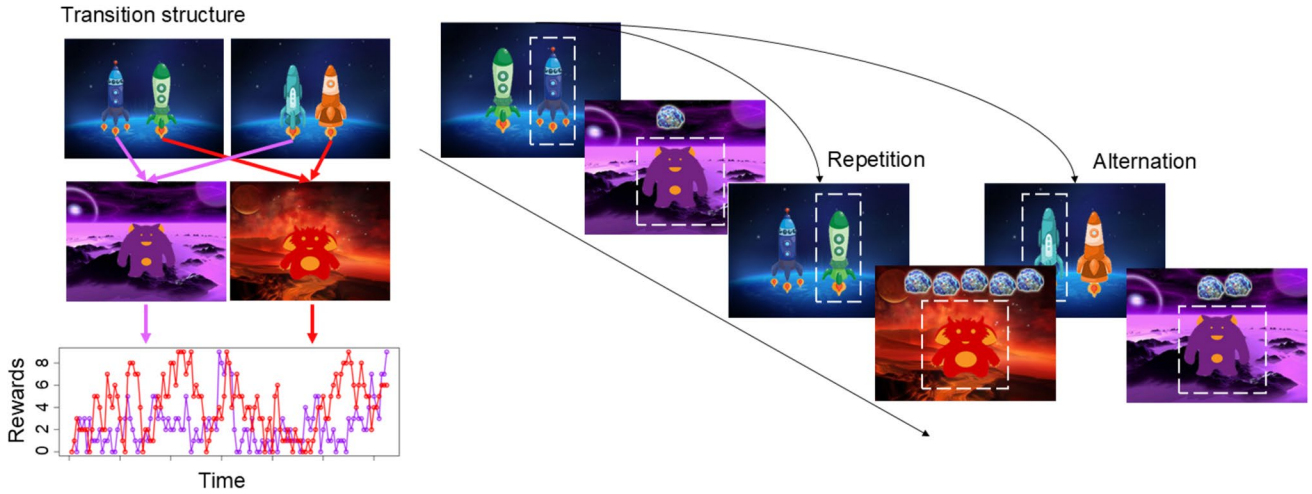


**Fig. 1** Transition structure and task procedure. Participants performed an inducer phase in which they encountered 80% first-stage state repetitions or 80% first-stage state alternations, followed by a test phase in which they encountered 50% first-stage state repetitions and alternations, respectively

control depending on the environment-specific demands and benefits of using these different strategies.

Previous findings have shown that the regulation of model-based and model-free control can be driven, among others, by cognitive load (Gershman et al., 2014; Otto et al., 2013a), opportunity costs (Pezzulo et al., 2013), time (Keramati et al., 2011), stress (Otto et al., 2013b; Radenbach et al., 2015) or amplification of outcomes (Kool et al., 2017). In this study, we aimed to test whether people can also learn the value of model-based versus model-free control from more subtle contingencies in the environment that make one or the other strategy more beneficial. To test this question, we implemented a simple manipulation in the deterministic two-step task developed by Kool et al. (2017) and Keramati et al. (2011). On each trial of this task, participants start in one of two first-stage states. Each of these states is associated with a unique pair of spaceships. For each of these two pairs, the two spaceships deterministically lead to one of two planets, each associated with an alien that provides rewards that drift slowly over time (Fig. 1). In our version of this task, we varied levels of environmental demand by changing the frequency of repeating versus alternating first-stage states (i.e., the initial spaceship pairs to choose from). We reasoned that if spaceship pairs frequently alternate, it is worth mobilizing cognitive resources to update the current reward values of each planet associated with the spaceships in each pair (i.e., to be more model-based). However, if spaceship pairs frequently repeat, generalizing between pairs has less added benefit, and simply repeating previously rewarded actions (i.e., to act more model-free) may be sufficient. We tested whether people are sensitive to these regularities in a subsequent test phase with an equal number of first-stage state repetitions and alternations. Here, we predicted that people who initially encounter more alternations between the first-stage states would learn to be more model-based, while those subjected to more first-stage state repetitions would learn to be more model-free. Importantly, using a balanced test phase that was the same for both groups allowed us to test whether they adopted different strategies that could not be attributed to local differences in trial-by-trial dynamics (for a similar reasoning, see Simoens et al., 2024, 2025; Wen et al., 2023; Xu et al., 2024). We hypothesized that this would be captured by a significantly larger mean parameter estimate for model-based control in a dual-system reinforcement learning model, as well as by a stronger effect of previous reward prediction errors on subsequent spaceship choice (across first-stage states) in the group that experienced more first-stage state alternations.

## Methods

### Participants

We recruited 144 participants from Ghent University's recruitment platform, who were first-year psychology students participating for course credit. Of these, four were excluded based on more than 40 missing responses out of 250 trials.[1] This number was based on previous research by Kool and colleagues (2017), who used it as a threshold in a 200-trial version, thus presenting a slightly more

[1] To test the robustness of our main test of interest, we also ran it including these outliers, which did not change the significance or pattern of results.

conservative criterion (corresponding to 16% of all trials). For statistical comparisons of demographic variables between the two groups, see Table 1. A post-hoc power analysis using G*Power (Faul et al., 2007) indicated 93% power to detect a medium effect size ($d = 0.5$) and 72% power to detect a small effect size ($d = 0.2$) for a one-tailed t-test with $\alpha = 0.05$. Participants were not given bonus payments based on performance, but they were told that the goal of the experiment was to collect as many points as possible.

## Materials and procedure

The experiment was programmed in jsPsych (de Leeuw, 2015), retrieved and adapted from Kool and colleagues (Kool et al., 2016, 2017), and run online. The task procedure was similar to their version, with the only change being the manipulation of the first-stage state sequences described later. In each trial, participants were presented with one of two screens (first-stage states), each containing two spaceships (Fig. 1). Participants had to choose between these spaceships, which were presented side-by-side using the "F" and "J" keys on the keyboard within 1,500 ms. This choice determined which of two planets (second-stage states), a red or a purple one, would then be visited. Importantly, the choices in each first-stage state afforded the opportunity of transitioning to either plane: one ship always led to the purple planet and the other to the red planet. In this second-stage state, participants encountered a single alien that would give the rewards in the form of "space treasure" after a button press. The rewards provided by the aliens were determined according to independent and slowly drifting reward distributions (Gaussian random walk with rewards varying between 0 and 9 ($\sigma = 2$)).

Because the choices between spaceships are equivalent between the first-stage states, this task allows us to distinguish between model-based and model-free strategies. This is possible because only the model-based system transfers experiences learned in one starting state to the other starting state. For a model-based agent, each second-stage outcome affects subsequent first-stage preferences equivalently, i.e., independent of whether the next trial starts with the same pair of spaceships as on the previous trials. This is because model-based agents plan towards the second-stage goals. In contrast, a pure model-free strategy does not transfer experiences between first-stage states, because they only learn action-reward associations (Doll et al., 2015). In other words, while a model-free agent simply repeats choices of spaceships that were previously linked to high reward, i.e., assigning the credit to the specific spaceship leading to the planet with reward, a model-based agent generalizes the reward between spaceships, i.e., assigning the credit using the shared transition structure.

Before starting the main experiment, participants underwent extensive training to learn the reward manipulation and transition structure, i.e., which spaceship leads to which planet, followed by 25 complete practice rounds. They were instructed that the goal was to collect as many points as possible throughout the main experiment. Next, they proceeded to the main experiment, which consisted of two phases of 125 trials each (to ensure a sufficient number of trials for model estimation, based on Kool et al., 2016): an inducer phase and a test phase. Importantly, and in contrast to the original version by Kool and colleagues (2016; 2017), participants in the first-stage "repetition" group were presented with 80% repetitions of first-stage states in the first phase of the experiment (i.e., inducer phase). That is, on 80% of trials, they encountered the same pair of spaceships as on the previous trial. Conversely, participants in the first-stage "alternation" group encountered first-stage state alternations 80% of the time. In both groups, we ensured participants saw both first-stage states 50% of the time. In the second phase (i.e., test phase), both groups underwent another 125 trials, in which they saw 50% first-stage state repetitions or alternations, respectively. At no point in the experiment were participants instructed on the frequency manipulation.

**Table 1** Demographics per group

| Variable | Repetition group ($n = 69$) | Alternation group ($n = 71$) | Test statistic | $p$ |
|---|---|---|---|---|
| Age (years) | M = 19.29, SD = 2.88 | M = 18.63, SD = 1.02 | $t(82.89) = 1.789$ | .077 |
| Gender (men/women/other) | 47/20/1 | 57/14/0 | $\chi^2(2) = 2.957$ | .228 |
| Handedness (left/right/ambidexter) | 8/60/0 | 3/66/2 | $\chi^2(2) = 4.496$ | .106 |

For age, Welch's modification to degrees of freedom was used due to inequality of variances. For one participant in the repetition group, demographic data were not saved

## Reinforcement model

We formally quantified model-based and model-free contributions to behavior using a dual-system reinforcement learning model developed by Kool et al. (2016). This model was fit to behavioral data from the test phase, where both groups performed the same task (i.e., with 50% first-stage state repetitions and alternations). We only fit the model to this test phase, as a critical test of whether a generalizable strategy was learned independent of the local trial-by-trial changes.

In the following, we provide a narrative explanation of the dual-system reinforcement learning model with a non-exhaustive conceptual overview of the key equations. A comprehensive overview is provided in Appendix A. The

model contains two strategies, one model-based and the other model-free. Each strategy (model-based and model-free control) learns a $Q$ function that maps each state-action pair ($s$, $a$) to an estimate of expected future return using different algorithms.

The model-free learner uses a temporal difference learning algorithm (SARSA, Sutton & Barto, 1998), increasing values for state-action pairs with positive reward prediction errors (PE) and decreasing values for those with negative reward prediction errors across both stages. At each stage, the PE is computed based on the difference between the observed outcome (obtained reward and estimated future returns) and the expected return according to

$$PE = \left(r + Q'\right) - Q\left(a_{chosen}\right),$$

where $r$ denotes the reward (only received at the second stage), $Q'$ the estimated future return (only present at the first stage, carried over from the second-stage values of the according state-action pair at the last encounter), and $Q(a_{chosen})$ the observed value of the action. The chosen action is then updated according to

$$Q\left(a_{chosen}\right) \leftarrow Q\left(a_{chosen}\right) + \alpha * PE,$$

where $\alpha$ reflects the learning rate (the degree to which new information is integrated into existing estimates of future reward). Hence, the model-free learner learns all first and second-stage state-action values separately.

The model-based algorithm combines the task's transition structure (which first-stage action leads to which second-stage state) with the current estimate of the model-free second-stage $Q$ values to compute its first-stage $Q$ values. In other words, the model-based values of each spaceship are identical to the second-stage model-free value of its corresponding planet.

These model-based and model-free action values are then combined using a weighting parameter $w$ according to

$$Q_{net} = w * Q_{mb} + (1 - w) * Q_{mf,S1}$$

Model-based control was indexed by weights closer to 1, whereas model-free control was indexed by weights closer to 0.

In addition to $\alpha$ and $w$, we estimated an inverse temperature parameter ($\beta$), which determined the exploitation/exploration trade-off between the two choice options given their difference in value. Concretely, the higher the value, the more likely the agent is to choose the option with the highest value; the closer it is to 0, the more likely the agent is to choose uniformly. In other words, this parameter estimates how much responses can or cannot be attributed to the learned values of the different options. While lower values can be attributed to a tendency for exploration, it is important to note that they may also reflect the noisiness of decision-making, or an inability of the model to detect other systematic response strategies (i.e., poor model fit). This parameter is part of the SoftMax function that is used for choice selection by scaling $Q_{net}$. We also included an eligibility trace parameter ($\lambda$) representing the degree to which information of the second-stage prediction errors is used to update first-stage action values. Mathematically, this parameter is used to discount the product of the learning rate and second-stage PE when updating the first-stage action value. Finally, we added choice ($\pi$) and response stickiness ($\rho$) parameters, capturing participants' persistence in choosing particular spaceships or pressing response keys unrelated to reward. These parameters were implemented by multiplying their values by a binary indicator of whether they repeated or alternated, and adding their product to $Q_{net}$ prior to entering the SoftMax function.

The model was fit to behavior in MATLAB using maximum *a-posteriori* estimation and parameter priors reported by Gershman (2016). For more details on this procedure, see Appendix A. To assess model fit, we compared the full computational model to three baseline models using Akaike's Information Criterion (AIC). First, we compared it to a model-free learner with $\alpha$ and $\lambda$ set to 1, i.e., assuming perfect learning and retention, and $\beta$ set to 0.5. All other parameters, i.e., $w$, $\pi$, and $\rho$, were fixed at 0. Second, we compared it to a reduced model similar to the full model, where only $\lambda$ was fixed at 1 and $\pi$ and $\rho$ at 0. Third, we compared it to the full model but without the eligibility trace parameter, which commonly has lower recoverability (see Appendix B and Kool et al., 2016). To test whether participants in the alternation group showed more model-based control than the repetition group, we conducted $t$-tests to compare the individual parameter estimates between groups after model fitting.

## Simulating the environment-specific marginal gains of using model-based control

Before running the main analyses on participants' performance, we ran simulations to confirm whether adopting a model-free strategy effectively paid off more when first-stage states repeat vs. alternate frequently (in terms of total reward earned). To this end, we simulated data from 125 trials for 500 agents per reward group based on 500 plausible parameter combinations.

## Parameter recovery based on fitted values

As a post-hoc descriptive check, we performed a parameter recovery based on the fitted values reported in the Results section, which resulted in robust recovery of most parameters (the eligibility trace decay parameter showed weak

recoverability, consistent with prior work). Correlations between true and simulated values are reported in Appendix B. For a more extensive procedure based on randomly selected values, see Kool et al. (2016). To assess the model's ability to capture key patterns in the observed data, we additionally conducted the same behavioral analyses reported below on the model-simulated data. The simulated results closely matched the empirical findings, supporting the model's fit (see Appendix C for details).

### Behavioral signatures of model-based and model-free control

To obtain a behavioral estimate of whether participants in the group encountering more repetitions in the inducer phase showed more model-based behavior in the test phase, we also ran a logistic mixed effects model in brms[2] (Bürkner, 2017, 2018, 2021). This model predicted whether participants stayed with or switched the second-stage state from the previous trial as a function of the sign of the second reward prediction error (PE) on the last trial (retrieved from the computational model), first-stage state sequence (repetition or alternation) and group (in line with Kool et al., 2017), following a maximum random effects structure:

$$Stay \sim Sign(PE_2) * Transition * Group + (1 + Sign(PE_2) * Transition | Subject).$$

In logistic models for this task, a main effect of previous PE reflects model-based control (because it reflects a tendency to visit previously rewarded second-stage states regardless of starting state), whereas an interaction between previous PE and first-stage state reflects model-free control (because the effect of previous outcome depends on whether the same first-stage state is encountered). We predicted that we would observe a three-way interaction between group, the sign of the previous PE, and the first-stage state sequence. Specifically, we expected that participants in the group encountering more first-stage state alternations would show more stay (leave) behavior following positive (negative) prediction errors on both repetition and alternation trials (where one has to generalize across first-stage states, i.e., be model-based), whereas participants in the group encountering more first-stage state repetitions to show more stay (leave) behavior following positive (negative) PE, but only on repetition trials (where no generalization is necessary).

[2] Note that we used a Bayesian approach rather than Frequentist methods by default, as it often facilitates model convergence while reaching similar qualitative conclusions. (see also Figner et al., 2024).

## Results

### Simulations

As expected, our simulated data showed a significant correlation between the model-based weighting parameter and the total rewards earned in the group encountering more first-stage state alternations (Pearson's $r(498) = 0.168$, $p < .001$), indicating that it paid off to be model-based in this group in terms of total rewards earned. However, this correlation was not observed in the group encountering more first-stage state repetitions (Pearson's $r(498) = 0.036$, $p < .425$). A Fisher's z-test indicated a significant difference between these two correlations, $p = .017$, suggesting there was a reduced need to be model-based in this second group, and a cheaper model-free strategy was sufficient regarding total pay-off.

### Comparing model-based control in high- versus low-alternation environments

Model comparison of our full model against the baseline models revealed the full model had the lowest (AIC = 17,273.62) compared to the model-free learner (AIC = 23,303.95, ΔAIC = 6030.33), the reduced model (AIC = 18,284.05, ΔAIC = 1010.42), and the model with the eligibility trace parameter fixed at 1 (AIC = 17,319.89, ΔAIC = 46.26). This suggests that our full model showed the best model fit and best explained the data.

As expected, results from our *t*-tests for group comparisons revealed that people in the group encountering more first-stage state alternations showed increased model-based control compared with people in the group encountering more first-stage state repetitions ($t(128.97) = -2.374$, $p < .01$; *Cohen's d* = 0.4, one-sided, Welch's modification to degrees of freedom used owing to inequality of variances). Moreover, they were less persistent in their choices, as indicated by the choice stickiness parameter ($t(138) = 2.013$, $p = .046$; *Cohen's d* = 0.34). All other parameter estimates did not significantly differ between groups (all $ps > .167$; Fig. 2; Table 2). We note that the temperature and learning rate parameters often trade off against each other in reinforcement learning models (Gershman, 2016), and were likewise negatively correlated in our dataset ($r = -0.87$, $p < .001$). This trade-off is important to acknowledge (Wilson & Collins, 2019). Importantly, however, our primary inferences did not concern these parameters, and we observed no group differences in either parameter. Correlations between all parameter estimates are displayed in Appendix D.

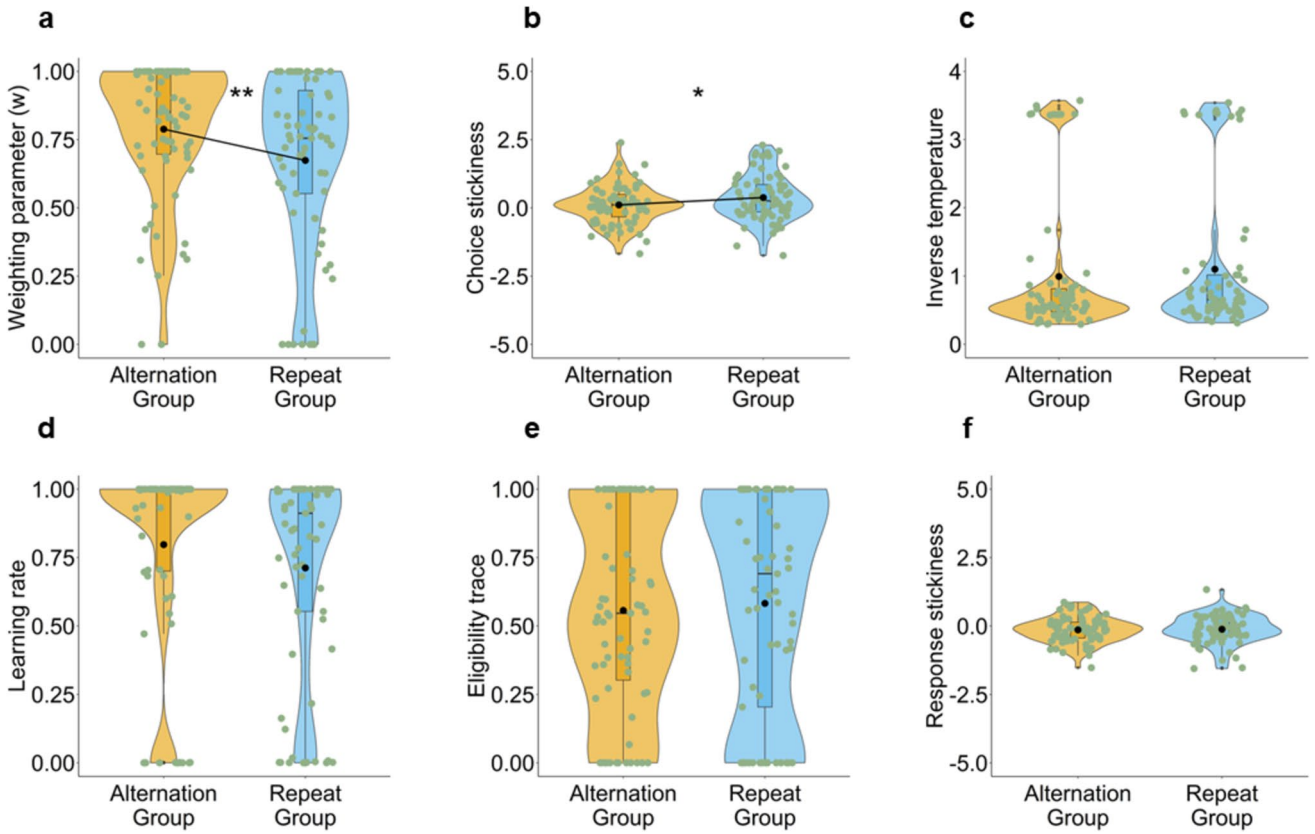


**Fig. 2** Parameter differences between the two groups in the test phase. Participants in the group encountering more first-stage state alternations (80%) were more model-based in the test phase (50% first-stage state alternations and repetitions) than participants in the group encountering more first-stage state repetitions (**a**) and less persistent in their choices (**b**). No group differences were found for any of the other parameters (d-f). *$p < .05$; **$p < .01$

**Table 2** Parameter estimates per group

| | $\beta$ | $\alpha$ | $\lambda$ | $\pi$ | $\rho$ | $w$ |
|---|---|---|---|---|---|---|
| Repetition group | | | | | | |
| Mean | 1.10 | 0.71 | 0.58 | 0.38 | −0.12 | 0.67 |
| SD | 1.04 | 0.37 | 0.40 | 0.85 | 0.52 | 0.32 |
| 25th percentile | 0.50 | 0.55 | 0.20 | −0.15 | −0.35 | 0.55 |
| Median | 0.65 | 0.91 | 0.69 | 0.26 | −0.10 | 0.76 |
| 75th percentile | 1.02 | 1 | 1 | 0.86 | 0.29 | 0.93 |
| Alternation group | | | | | | |
| Mean | 0.99 | 0.80 | 0.56 | 0.11 | −0.14 | 0.79 |
| SD | 1.01 | 0.35 | 0.36 | 0.71 | 0.47 | 0.25 |
| 25th percentile | 0.48 | 0.70 | 0.30 | −0.33 | −0.44 | 0.70 |
| Median | 0.57 | 1.00 | 0.55 | 0.11 | −0.11 | 0.85 |
| 75th percentile | 0.81 | 1 | 1 | 0.49 | 0.14 | 1 |

$\beta$ = inverse temperature; $\alpha$ = learning rate; $\lambda$ = eligibility trace; $\pi$ = choice stickiness; $\rho$ = response stickiness; $w$ = weighting parameter

## Differences in choice behavior in high- versus low-alternation environments

As described, we next ran a logistic regression model to test the presence of a significant interaction between group, the sign of the previous PE, and the first-stage state sequence, as a marker of qualitative differences in terms of model-based versus model-free control in both groups. As expected, we

found main effects of previous PE ($b = -0.406$, 95% CI [−0.519, −0.403]), with participants being more likely to stay following a positive PE, and of transition ($b = 0.148$, 95% CI [0.098, 0.197]), with participants being overall more likely to choose the same second-stage state again when first-stage states repeated. We also saw an interaction between the sign of the previous PE and transition ($b = -0.084$, 95% CI [−0.126, −0.039]), with participants being more likely to stay following a positive PE on repetition as compared to alternation trials. This interaction was not further modulated by group ($b = -0.004$, 95% CI [−0.046, 0.035]) (Fig. 3). We also did not see an interaction between the sign of the previous PE and group ($b = 0.023$, 95% CI [−0.033, 0.083]), although the alternation group showed numerically less stay behavior following negative prediction errors.

## Discussion

The goal of this project was to investigate whether people can learn to arbitrate adaptively between two reinforcement-learning strategies, i.e., model-free and model-based control, depending on features of the environment. While others have shown that people can learn to arbitrate between both strategies based on explicit incentives (Kool et al., 2017; Patzelt et al., 2019), we showed that they can also learn to regulate these control modes based on uninstructed learned environmental features. Specifically, we manipulated how frequently participants encountered first-stage state alternations or repetitions in the two-step task, thereby influencing which strategy is more effective in terms of a cost–benefit trade-off (Kool et al., 2016, 2017; Otto & Daw, 2019; Shenhav et al., 2013). As model-based control pays off more when the first-stage states alternate frequently due to having to update values between the two spaceship pairs, we expected and found that participants eventually learned to act more model-based in a subsequent, separate test phase. Our finding provides important evidence for the ability of humans to learn adaptive control settings over time, without explicit instructions or task cues, in line with a learning perspective on cognitive control (Abrahamse et al., 2016; Braem et al., 2024; Held et al., 2024; Simoens et al., 2024; Xu et al., 2024). This perspective encompasses the idea that cognitive control functions, traditionally assumed to describe an independent, hierarchical set of top-down functions, are grounded in a broader, heterarchical network susceptible to associative learning. The fact that such learning translates to reinforcement learning modes further hints at a general self-regulatory mechanism that can explain adaptive parameter settings without necessitating a homunculus or "loan of intelligence" (Boureau et al., 2015; Dennett, 1981).

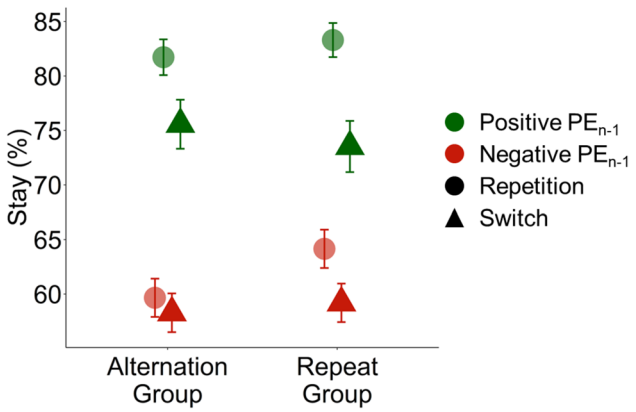


**Fig. 3** Differences in choice behavior. Probability of staying following a negative or positive prediction error (PE) on alternation and repetition trials. We see a trend of participants in the repetition group being less likely to stay (alternate) following a positive (negative) prediction error on alternation trials (not significant). Overall, they were more likely to stay following first-stage state repetitions

This perspective also aligns with a recent study showing that fixed stimulus locations in the first-stage states of the two-step task bias behavior towards model-free control (Luna et al., 2023). This is hypothesized to occur through the formation of stimulus–response associations, which marks another instance of how relatively subtle environmental features can determine arbitration between reinforcement learning systems. These results mirror our finding that the group subjected to more first-stage state repetitions also showed higher choice stickiness, indicating that this more stable environment promoted more stimulus-reward learning. In a similar vein, the frequent alternation of first-stage states in the alternation group may have hindered the formation of such habits or model-free control, in line with the notion that context changes benefit goal-directed behavior but hinder habit formation (Bouton, 2021).

Another interesting question is whether model-based learning, if acquired owing to environmental demands that disrupt habit formation, can later be unlearned. This could, for example, be observed in an experiment with a longer test phase (allowing participants to pick up on the new statistical structure) or in a within-subject design where participants experience both conditions. While such a design offers higher statistical power, we here opted for a between-subjects design as a first test of our hypothesis, as previous research from our lab suggests that learning control parameters is rather slow, and once learned, may be hard to unlearn, necessitating long experiment durations (Braem et al., 2024; Held et al., 2025).

It would also be interesting to study the neural basis of this flexible regulation to gain further insights into its mechanisms. One could test whether the increased reliance on model-based control in the group encountering more first-stage state alternations is indeed due to generalization between first-stage states as compared to, for instance, forward-planning. To this end, one could compare activity in

midbrain regions and the hippocampus, associated with generalization in value-based decisions and integrative encoding during initial learning (Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012) with areas encoding second-stage states prospectively (Doll et al., 2015). As an alternative, another strategy could be to simply add self-report questions, asking people about their strategy use to gain deeper insight, an often underrated tool in psychological research (Simon & Ericsson, 1984; Wurgaft et al., 2025; Xie et al., 2023).

Our findings may further have clinical implications. Specifically, previous research has found interesting interindividual differences in applying either strategy, highlighting an important transdiagnostic mechanism (Gillan et al., 2016). Extending these findings, others have pointed to interindividual differences in regulating both strategies based on reward incentives that participants were explicitly told to consider (Patzelt et al., 2019). In contrast, our manipulation relied on less explicit learned contingencies with environmental demands. As also argued for other disorders (Geurts et al., 2009; Van Eylen et al., 2011), the use of explicit task instructions or manipulations can often lead to a failure to detect underlying cognitive impairments in clinical disorders because (difficulties with) our daily activities or everyday life often come without clear instructions on how to act. For example, it has been suggested that people with autism have difficulty with the context-specific adjustment of control parameters (Goris et al., 2018; Palmer et al., 2015). Also, attention deficit/hyperactivity disorder (ADHD) and schizotypal traits are thought to affect environmentally guided reinforcement-learning arbitration. ADHD is linked to differences in probability tracking (Frank et al., 2007) and schizophrenia to altered belief updating, impairing flexible learning (Nassar et al., 2021). We, therefore, believe it would be interesting to study whether (mal)adaptive regulation of reinforcement strategies based on environmental demands could be a clinical marker. Similarly, it would be interesting to study this in development, meaningfully extending findings by Smid et al. (2023), who found less arbitration between both strategies in children based on explicit stake cues.

One limitation of our work concerns the current criticisms of the strict distinction between model-based and model-free control. For instance, it has been argued that this dichotomy oversimplifies more complex decision-making structures (Collins & Cockburn, 2020), that there is no sharply defined line separating them (Miller et al., 2019), or that the model-free system is hierarchically governed rather than flat, as modelled in this paper (Dezfouli & Balleine, 2012, 2013; Dezfouli et al., 2014). Another criticism states that seemingly model-free behavior in humans may arise simply from inaccurate models rather than a model-free process (Feher Da Silva & Hare, 2020). We believe these criticisms do not diminish the value of our findings. Our predictions rely on the assumed computational cost of model-based control and its increased usefulness in an environment with more first-stage state alternations. Because our main interest was to document how people learn to arbitrate the relative costs and benefits of different decision strategies (see also Otto et al., 2022), showing a similar learned arbitration between this type of model-based control and, for instance, a computationally efficient but incorrect model (a wrong model would take up resources as well, see also, Morris & Cushman, 2019) should thus be equally informative.

Finally, our findings have some real-life implications. By showing that people can learn to adopt model-based strategies based on learning environments, not just explicit incentives, we open a door for intervention studies to focus on creating such optimal environments. For instance, to promote healthier choices, grocery stores could be organized in ways that favor model-based control through alternating product locations, reminding people of better goal-directed choices. More generally, everyday life unfolds in fast-changing, dynamic environments. Demonstrating that individual differences in exposure significantly affect control settings highlights the importance of this topic for future research (see also Coutrot et al., 2022; Heller et al., 2020).

In sum, our findings suggest that people can learn to adaptively regulate their reliance on model-based versus model-free control depending on environmental demands. Namely, individuals showed changes in control strategies in response to the frequency of first-stage state alternations, favoring model-based control when alternations were frequent and defaulting to the less costly model-free approach when first-stage state repetitions dominated. Future research exploring individual differences in model-based versus model-free control could further elucidate how this learning of adaptive control processes may vary across different (sub)clinical populations or stages of development.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13415-025-01350-9.

**Authors' contributions** LKH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Writing – original draft, review and editing. EL: Resources, Validation, Writing – review and editing. WK: Conceptualization, Formal analysis, Resources, Writing – review and editing. SB: Conceptualization, Supervision, Writing – original draft, review and editing.

**Funding** This work was supported by an ERC Starting grant awarded to S.B. (European Union's Horizon 2020 research and innovation program, Grant agreement 852570) and a fellowship by the Fonds Voor Wetenschappelijk Onderzoek–Research Foundation Flanders 11C2322N to L.K.H.

**Data availability** All data are available at https://osf.io/6gy9c/?view_only=994e80d9931448dea371782c5443a0af.

**Code availability** All code is available under https://osf.io/6gy9c/?view_only=994e80d9931448dea371782c5443a0af.

## Declarations

**Conflicts of interest** None.

**Ethics approval** The study was approved by the Ethics Committee of the Faculty of Psychological and Pedagogical Sciences of Ghent University.

**Consent to participate** All participants gave consent to participate.

**Consent for publication** All participants gave consent for publication of their anonymized data.

## References


Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin, 142*(7), 7. https://doi.org/10.1037/bul0000047

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences, 19*(11), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Bouton, M. E. (2021). Context, attention, and the switch between habit and goal-direction in behavior. *Learning & Behavior, 49*(4), 349–362. https://doi.org/10.3758/s13420-021-00488-z

Braem, S., Chai, M., Held, L. K., & Xu, S. (2024). One cannot simply 'be flexible': Regulating control parameters requires learning. *Current Opinion in Behavioral Sciences, 55*, Article 101347. https://doi.org/10.1016/j.cobeha.2023.101347

Braem, S., & Egner, T. (2018). Getting a grip on cognitive flexibility. *Current Directions in Psychological Science, 27*(6), Article 6. https://doi.org/10.1177/0963721418787475

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian multilevel models using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in *R* with **brms** and *Stan*. *Journal of Statistical Software*, *100*(5). https://doi.org/10.18637/jss.v100.i05

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience, 21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe, D., Dalton, R. C., Wiener, J. M., Hölscher, C., Hornberger, M., & Spiers, H. J. (2022). Entropy of city street networks linked to future spatial navigation ability. *Nature, 604*(7904), 104–110. https://doi.org/10.1038/s41586-022-04486-7

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. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711. https://doi.org/10.1038/nn1560

de Leeuw, J. R. (2015). JsPsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods,* 12. https://doi.org/10.3758/s13428-014-0458-y

Dennett, D. C. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. The MIT Press. https://doi.org/10.7551/mitpress/1664.001.0001

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience, 35*(7), 1036–1051. https://doi.org/10.1111/j.1460-9568.2012.08050.x

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology, 9*(12), Article e1003364. https://doi.org/10.1371/journal.pcbi.1003364

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: Hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1655), 20130482. https://doi.org/10.1098/rstb.2013.0482

Doebel, S. (2020). Rethinking executive function and its development. *Perspectives on Psychological Science, 15*(4), 4. https://doi.org/10.1177/1745691620904771

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*(5), 767–772. https://doi.org/10.1038/nn.3981

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feher Da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour, 4*(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y

Figner, B., Algermissen, J., Burghoorn, F., Chen, Z., Fenneman, J., Guo, M., Held, L., Khalid, A., Klaassen, F., Klein Breteler, J., Mosannenzadeh, F., Quandt, J., Vadakkedath Dharmapalan, M., & Wienicke, F. (2024). *Standard Operating Procedures (SOP) for using Mixed-Effects Models: A Principled Workflow from the Decision, Development, and Psychopathology (D2P2) Lab*.

Frank, M. J., Santamaria, A., O'Reilly, R. C., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology, 32*(7), 1583–1599.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal Of Mathematical Psychology, 71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143*(1), 182–194. https://doi.org/10.1037/a0030844

Geurts, H. M., Corbett, B., & Solomon, M. (2009). The paradox of cognitive flexibility in autism. *Trends in Cognitive Sciences, 13*(2), 74–82. https://doi.org/10.1016/j.tics.2008.11.006

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife, 5*, e11305. https://doi.org/10.7554/eLife.11305

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Goris, J., Braem, S., Nijhof, A. D., Rigoni, D., Deschrijver, E., Van De Cruys, S., Wiersema, J. R., & Brass, M. (2018). Sensory prediction errors are less modulated by global context in autism spectrum disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(8), 667–674. https://doi.org/10.1016/j.bpsc.2018.02.003

Held, L. K., Goris, J., & Braem, S. (2025). *The influence of autism and reward on balancing cognitive flexibility versus stability*. https://osf.io/r97dy

Held, L. K., Vermeylen, L., Dignath, D., Notebaert, W., Krebs, R. M., & Braem, S. (2024). Reinforcement learning of adaptive control strategies. *Communications Psychology, 2*(1), 8. https://doi.org/10.1038/s44271-024-00055-y

Heller, A. S., Shi, T. C., Ezie, C. E. C., Reneau, T. R., Baez, L. M., Gibbons, C. J., & Hartley, C. A. (2020). Association between real-world experiential diversity and positive affect relates to hippocampal–striatal functional connectivity.

*Nature Neuroscience, 23*(7), 800–804. https://doi.org/10.1038/s41593-020-0636-4

Huang, Y., Yaple, Z. A., & Yu, R. (2020). Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning. *Neuroimage, 215*, 116834. https://doi.org/10.1016/j.neuroimage.2020.116834

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

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

Kool, W., Cushman, F. A., & Gershman, S. J. (2018a). Competition and cooperation between multiple reinforcement learning systems. In *Goal-Directed Decision Making* (pp. 153–178). Elsevier. https://doi.org/10.1016/B978-0-12-812098-9.00007-3

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science, 28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288

Kool, W., Gershman, S. J., & Cushman, F. A. (2018b). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience, 30*(10), 1391–1404. https://doi.org/10.1162/jocn_a_01263

Luna, R., Vadillo, M. A., & Luque, D. (2023). Model-free decision making resists improved instructions and is enhanced by stimulus-response associations. *Cortex, 168*, 102–113. https://doi.org/10.1016/j.cortex.2023.06.009

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review, 126*(2), 292–311. https://doi.org/10.1037/rev0000120

Morris, A., & Cushman, F. (2019). Model-free RL or action sequences?. *Frontiers in Psychology, 10*, 2892.

Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M., & Frank, M. J. (2021). All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain, 144*(3), 1013–1029.

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia, 123*, 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006

Otto, A. R., Braem, S., Silvetti, M., & Vassena, E. (2022). Is the juice worth the squeeze? Learning the marginal value of mental effort over time. *Journal of Experimental Psychology: General, 151*(10), 2324.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013a). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24*(5), 751–761. https://doi.org/10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013b). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America, 110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G., & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: A predictive processing perspective. *Proceedings of the Royal Society b: Biological Sciences, 282*(1802), Article 20141557. https://doi.org/10.1098/rspb.2014.1557

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). Incentives boost model-based control across a range of severity on several psychiatric constructs. *Biological Psychiatry, 85*(5), 425–433. https://doi.org/10.1016/j.biopsych.2018.06.018

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00092

Radenbach, C., Reiter, A. M. F., Engert, V., Sjoerds, Z., Villringer, A., Heinze, H.-J., Deserno, L., & Schlagenhauf, F. (2015). The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology, 53*, 268–280. https://doi.org/10.1016/j.psyneuen.2014.12.017

Shenhav, A., Botvinick, M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), Article 2. https://doi.org/10.1016/j.neuron.2013.07.007

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron, 60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Simoens, J., Braem, S., Verbeke, P., Chen, H., Mattioni, S., Chai, M., Schuck, N. W., & Verguts, T. (2025). *Two time scales of adaptation in human learning rates*. https://doi.org/10.1101/2025.06.05.658048

Simoens, J., Verguts, T., & Braem, S. (2024). Learning environment-specific learning rates. *PLoS Computational Biology, 20*(3), e1011978. https://doi.org/10.1371/journal.pcbi.1011978

Simon, H. A., & Ericsson, K. A. (1984). *Protocol analysis: Verbal reports as data*. The MIT Press.

Smid, C. R., Kool, W., Hauser, T. U., & Steinbeis, N. (2023). Computational and behavioral markers of model-based decision making in childhood. *Developmental Science, 26*(2), Article e13295. https://doi.org/10.1111/desc.13295

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/TNN.1998.712192

Van Eylen, L., Boets, B., Steyaert, J., Evers, K., Wagemans, J., & Noens, I. (2011). Cognitive flexibility in autism spectrum disorder: Explaining the inconsistencies? *Research in Autism Spectrum Disorders, 5*(4), 1390–1401. https://doi.org/10.1016/j.rasd.2011.01.025

Wen, T., Geddert, R. M., Madlon-Kay, S., & Egner, T. (2023). Transfer of learned cognitive flexibility to novel stimuli and task sets. *Psychological Science, 34*(4), 435–454. https://doi.org/10.1177/09567976221141854

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife, 8*, e49547. https://doi.org/10.7554/eLife.49547

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science, 338*(6104), 270–273. https://doi.org/10.1126/science.1223252

Wurgaft, D., Prystawski, B., Gandhi, K., Zhang, C. E., Tenenbaum, J. B., & Goodman, N. D. (2025). Scaling up the think-aloud method. *arXiv Preprint* arXiv:2505.23931.

Xie, H., Xiong, H., & Wilson, R. C. (2023, October). *Text2Decision: Decoding latent variables in risky decision making from think aloud text*. NeurIPS 2023 AI for Science Workshop.

Xu, S., Simoens, J., Verguts, T., & Braem, S. (2024). Learning where to be flexible: Using environmental cues to regulate cognitive control. *Journal of Experimental Psychology: General, 153*(2), 328–338. https://doi.org/10.1037/xge0001488