Supplemental Material for “The Neurocognitive basis of model-based decision making and its metacontrol”

Supplemental Methods

Cognitive tasks

Inhibition

SSRT task. In the SSRT, participants have to press in response to a visual go-cue as fast as possible (Figure 1a) but withhold a response when a stop-signal appears (Figure 1b) (Matzke et al., 2018). During the task, participants were asked to press the left arrow key when seeing the go-signal (i.e., a honey pot) on the left side of the screen and the down arrow key when the signal appeared on the right side. Ten practice trials were administered before 80 trials of the main task. Each trial started with the presentation of a fixation cross of 1250ms. On 25% of the trials, a stop signal (i.e., a picture of bees) was presented after the honey pot. If participants saw the stop signal, they were instructed to not press any key. The stop signal delay (SSD) started at 200ms and decreased by 50ms after a successful stop trial and increased by 50ms after an unsuccessful stop trial. Participants had to respond within 6-seconds, or the trial timed-out. To derive a measure of inhibition, the mean Stop-Signal Reaction Time (SSRT) was calculated using the integration method (Verbruggen et al., 2019). For this study, the SSRT was inversely coded to mean that a larger score indicates better inhibition (“SSRT”).

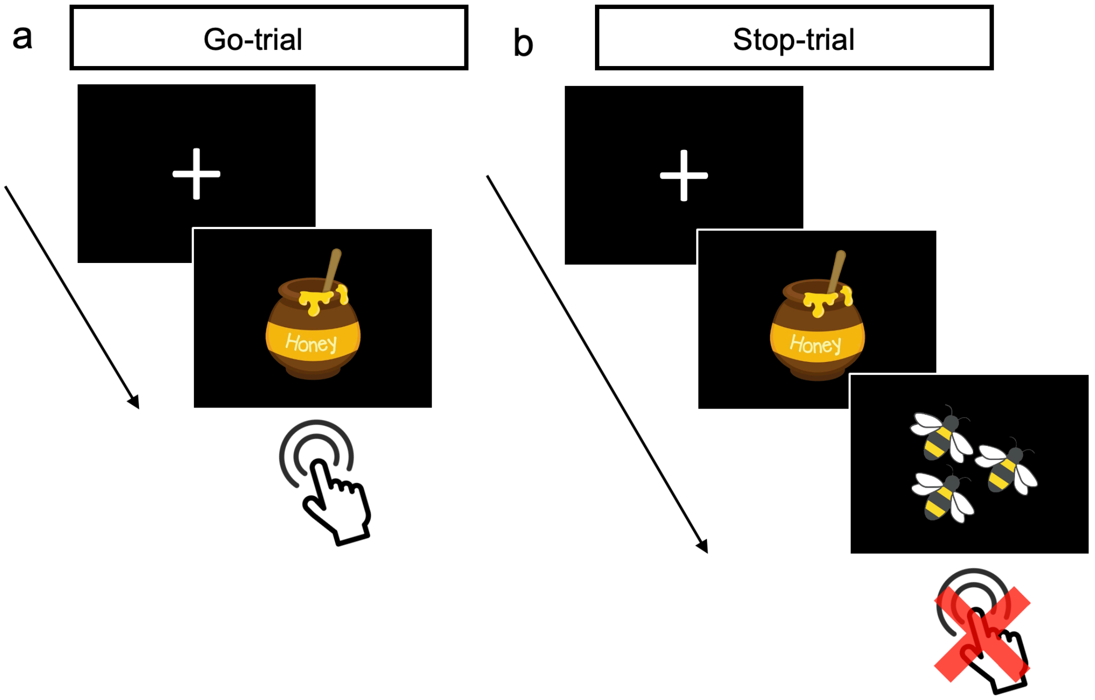

Figure 1. Stop Signal Delay (SSD) task.

During a go-trial (a), participants were instructed to react as fast as possible to the go-signal (honey pot), by pressing an arrow key depending on whether the stimulus was depicted on the left or right side of the screen (left and down arrow key). However, during a stop trial (b), the stop-signal was presented after a short delay (SSD), and participants were instructed to withhold their response.

Stroop task. A second measure of inhibition was a child-adapted Stroop task, where participants had to respond to congruent and incongruent trials with an auditory cue (Williams et al., 2007). Participants were asked to match animals to where they live (e.g., a frog to a pond). Four animal habitations were presented in the four corners of the screen throughout the game, and participants had to move their mouse pointer to the habitation of the animal on the current trial. At the start of every trial, a cartoon of an animal was displayed in the center of the screen. Participants were told that sometimes the animals wore disguises, and therefore to only respond to an auditory cue that indicated the animal type (e.g., frog – *"ribbit"*). On congruent trials, both auditory cues and visual cues matched (e.g., frog presented on screen and *"ribbit"* sound played) (Figure 2a). On incongruent trials, auditory cues and visual cues did not match (i.e., dog presented on screen and *"ribbit"* sound played) (Figure 2b). Participants completed 72 trials in total, with a 50/50 ratio of congruent and incongruent trials. Participants had to respond within 3 seconds, or the trial timed out. At the start of each trial, the mouse pointer location was reset to the center of the screen, and participants were presented with a blank screen in the center of the trial for 1000ms. For Stroop performance, the difference between reaction time and error rates were calculated separately for incongruent and congruent trials. Then the reaction time and error rates for each trial type were z-scored and summed. The performance measure used was the difference score between the incongruent minus the congruent trials, where a positive score indicated better performance on the incongruent trials. A lower score indicated less difference in the performance between the incongruent and congruent trials ("Stroop").

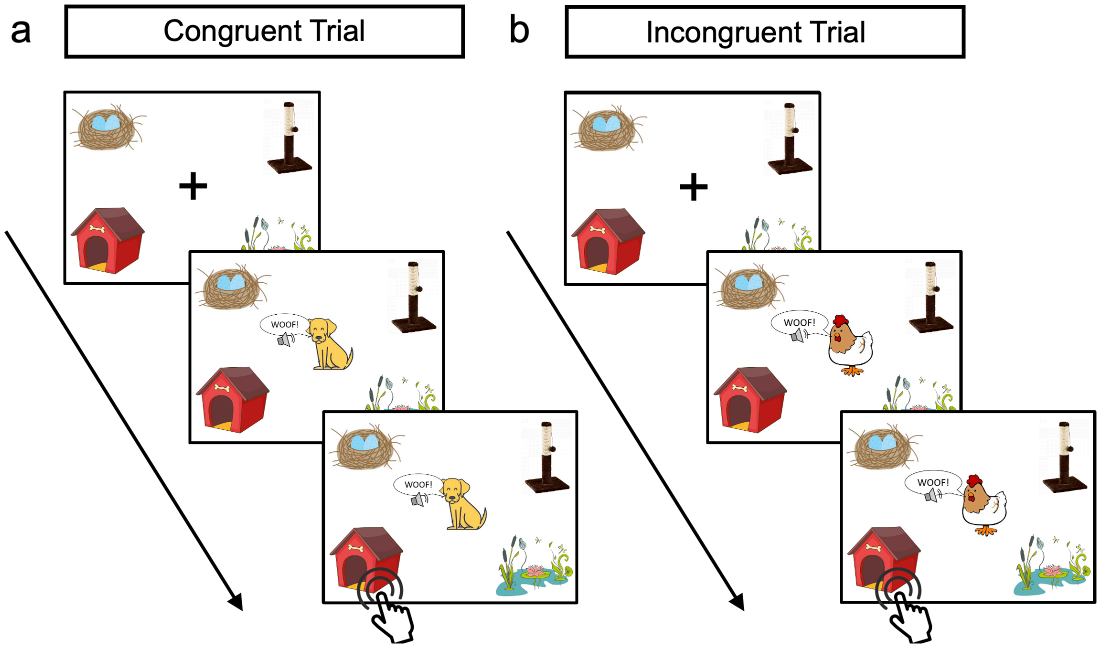

Figure 2. Animal Stroop task.

During the congruent trials (a), the animal depicted in the center and the noise emitted matched, while during the incongruent trials (b) the sound did not match the animal depicted. Participants were instructed to ignore the visual center stimulus and respond to the auditory stimulus.

Flanker task (inhibition component). I also used an adapted and child-friendly Flanker task that included an inhibition component. For this component, participants were shown a row of five fishes in the center of the screen. Participants were told to press an arrow key depending on the direction the central visual cue (the middle fish) was facing, and to ignore the direction of the distractor stimuli (the flanking four fishes). On congruent trials, the central visual goal cue was facing the same direction as the flanking distractor stimuli (Figure 3a), while in incongruent trials, the visual cue was facing the opposite direction from the distractor stimuli (Figure 3b). Participants first completed six practice trials, and then completed 40 trials in the inhibition component, with congruent and incongruent trials at a 50/50 ratio. At the start of the trial, participants saw a fixation cross for 500ms, and the central visual cue and the flanking distractor stimuli were shown simultaneously and for 700ms. After this time, the screen became blank, but participants had up to 2.5 seconds afterwards to make a response. Responses made before 100ms after stimulus onset were not recorded. The ITI was jittered and ranged from 800ms to 2400ms. For Flanker inhibition performance, the difference between reaction time and error rates were calculated separately for incongruent and congruent trials. Then the reaction time and error rates for each trial type were z-scored and summed. The performance measure used was the difference score between the incongruent minus the congruent trials, where a positive score indicated better performance on the incongruent trials. A lower score indicated less difference in the performance between the incongruent and congruent trials ("Flanker_Inhib").

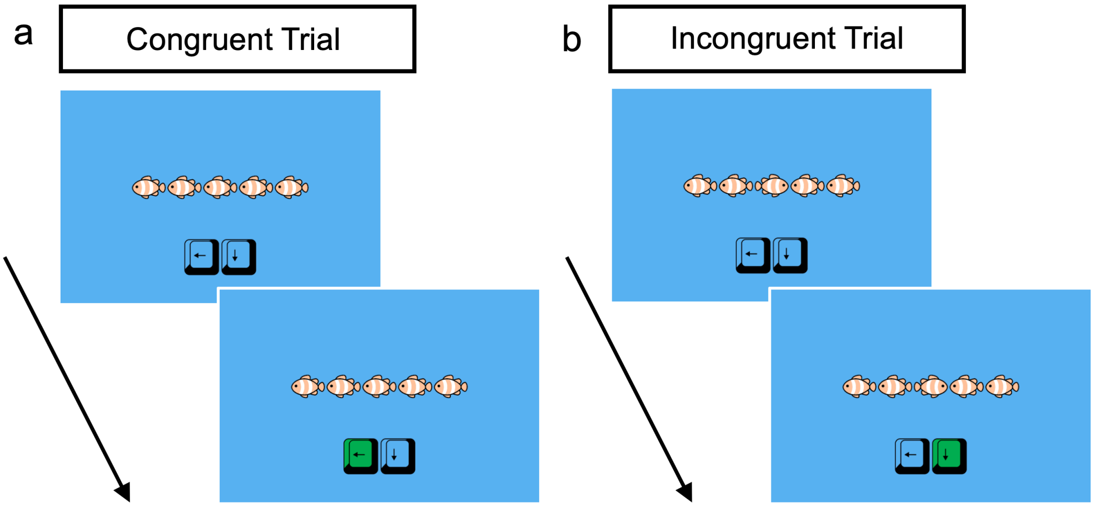

Figure 3. Flanker inhibition task.

During congruent trials (a) the central target stimulus was facing the same direction as the flanking distractor stimuli, while during the incongruent trials (b) the central target stimulus was facing the opposite direction from the flanking distractor stimuli. Participants were instructed to always focus on the central target stimulus and respond with a key press in the direction the stimulus was facing (left or down arrow key).

AX-CPT task. Lastly, inhibition was also measured via the AX-CPT task. The AX-CPT task measures participant's tendency to use more reactive or proactive control (Cooper et al., 2017). An A or B cue (i.e., dog or cat) were presented in the middle of the screen for 500ms followed by an inter-stimulus interval of 750ms and then a probe X or Y stimulus (orange or apple) during which participants had to make their response. Participants had 6000ms to respond until the trial timed out. Participants were instructed to press the left arrow key whenever an X followed an A (i.e., AX trials) (Figure 4a) and to press the down arrow key for the presentation of all other cue-probe combinations (Figure 4b). Trials were presented randomly and 40% of the trials were AX trials, and all other trials (i.e., AY, BX, BY trials) were presented 20% each (Richmond et al., 2015). Participants first completed ten practice trials with feedback followed by 60 main trials. To measure proactive control. I measured the difference in error rates and response times for the AY trials and the BX trials and calculate a composite score by deducting the BX trial performance from the AY trial performance and dividing that value by the sum of the AY and BX performance. I then created a composite score by z-scoring these measures and then taking the average. When there were zero error rates, these error rates were recoded to 1/2N where N is the number of trials. This measure is the Proactive Behavioral Index (PBI) and reflects the degree of proactive control displayed during the task, where a higher score reflects more proactive control ("AXCPT") (Gonthier et al., 2016).

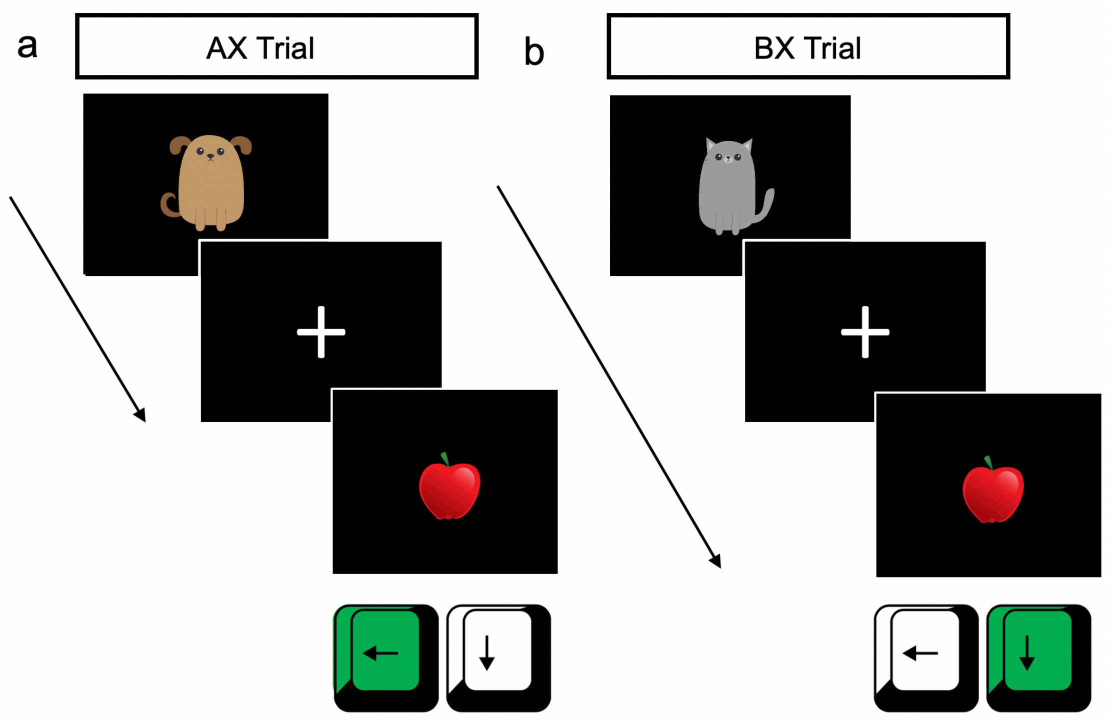

Figure 4. AX-CPT task.

During AX trials (a), participants had to respond by pressing the left arrow key, while during BX trials (b) (and all other trial combinations, e.g., AY, BY), participants had to respond by pressing the down arrow key.

Cognitive flexibility

Two measures of cognitive flexibility were used, the cognitive flexibility task and the cognitive flexibility component of the Flanker task.

Dimensional switching task. This task assessed participants ability for rule switching across dimensions (using sound cues (*"animal", "size"*) to respond to either the animal (cat or dog) or to the size of the animal (big or small) (Karbach & Kray, 2009). For every trial, a small or big image of a cat or dog was shown in the center of the screen, along with an image of the keys that could be pressed (left and down arrow key), and the audio cue was played. Underneath each arrow key, the options for the relevant dimensions for that trial were displayed in text (e.g., "small" and "big", or "cat" and "dog"). Participants had 10 seconds to respond before the trial timed out, during which the stimuli remained on the screen. Responses made before 200ms after stimulus onset were not recorded. The ITI was jittered and ranged from 1000ms to 1200ms. Stay trials were preceded by a trial in the same dimension (i.e., participants had to respond to the type of animal twice in a row) (Figure 5a), while during switch trials, the current trial was preceded by a trial in a different dimension (i.e., participants had to first respond to the size of the animal but now to the size) (Figure 5b). Participants completed 20 single dimension trials in two blocks, and 40 mixed trials in one block, and completed separate practice sessions for both single and mixed trials with four practice trials were three out of four trials had to be correct to progress. During the single dimension blocks, participants only had to respond to the same dimension (e.g., they only had to respond to the size of the animal), while in the mixed blocks, the two dimensions were mixed. Switch trials were controlled to only occur after either two or three preceding stay trials. Performance on the cognitive flexibility task was captured by the difference in speed and accuracy between switch and stay trials in the mixed blocks ("CogFlex"). A higher positive score indicated better performance on switch trials relative to stay trials.

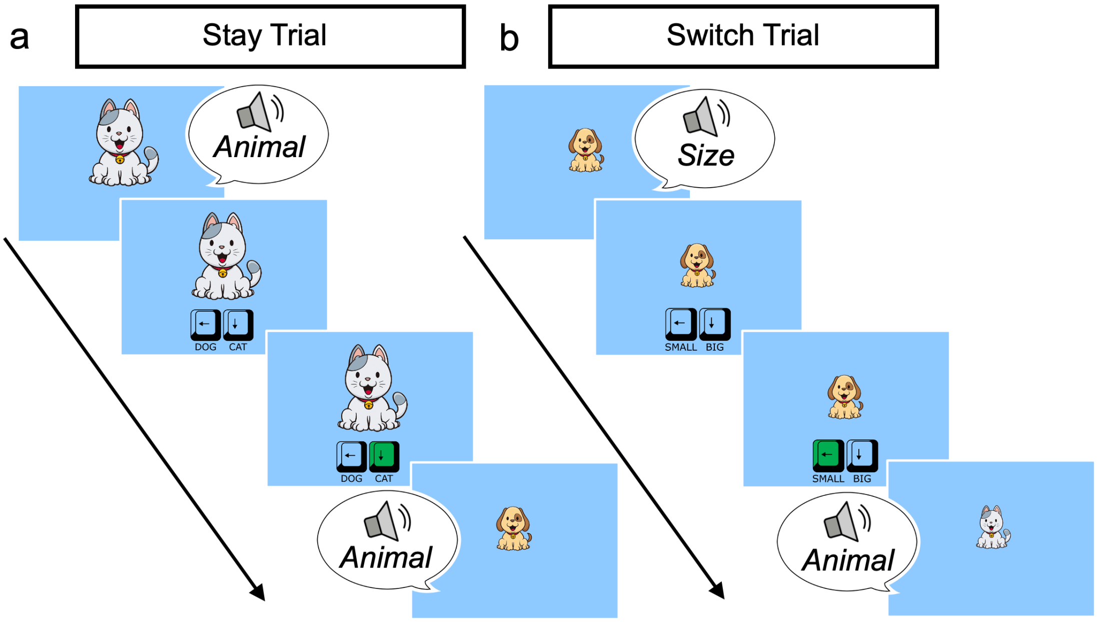

Figure 5. Dimensional switching task.

During stay trials (a), the previous trial was in the same dimension as the current trial, i.e., the participants had to respond to the type of animal displayed, and not to the size. During switch trials (b), the previous trial was a different dimension than the current trial, i.e., participants had to previously respond to the size of the animal displayed but now have to respond to the type of animal. After a short delay, an image of arrow keys with the current dimension (i.e., animal or size) was displayed under the central target stimulus.

Flanker task (cognitive flexibility component). Participants completed six practice trials before completing 40 trials across two conditions. In the stay condition, participant had to press the arrow key to match the direction the visual stimuli were facing (the row of five fishes, always all facing the same direction) (Figure 6a). In the switch condition, as indicated by all five fishes changing color, participants had to press for the opposite direction from the way the stimuli were facing (Figure 6b). Stimuli were presented for 700ms, and all responses made before 100ms were not recorded. The ITI was jittered and ranged from 800ms to 2400ms, and participants had 2.5 seconds to respond before the trial timed out. For switching performance, the difference between reaction time and error rates were calculated separately for switch and stay trials. Then the reaction time and error rates for each trial type were z-scored and summed. The performance measure used was the difference score between the switch minus the stay trials, where a positive score indicated better performance on the switch trials. A lower score indicated less difference in the performance between the switch and stay trials ("Flanker_Switch").

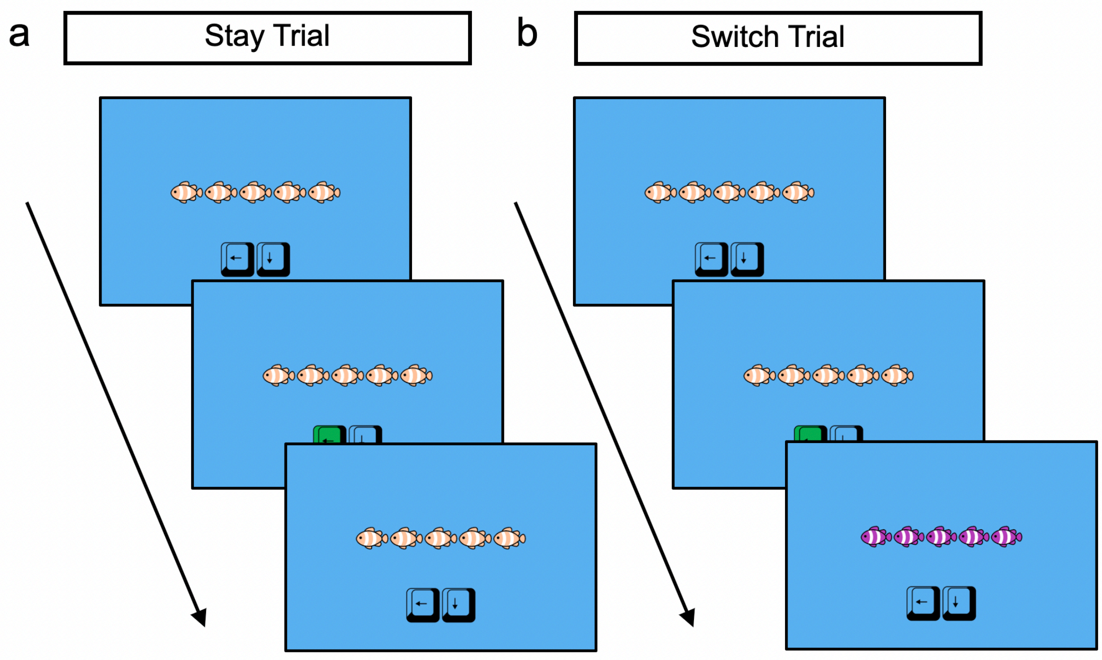

Figure 6. Flanker switching task.

During stay trials (a), participants had to respond by pressing the direction the fishes were facing as fast as possible, while during switch trials (b), participants had to press the opposite direction from which the fishes were facing. In this task, all fishes were always facing the same direction.

Working memory

Working memory span and manipulation was assessed via two tasks.

CORSI block-tapping task. This task measures visuo-spatial working memory span with a higher value indicating a higher working memory span (Farrell Pagulayan et al., 2006). This task was designed as a frog jumping between nine potential locations designed as lily pads (Figure 7a), with the participants following the jumps by clicking on the lily pads in forward sequence (Figure 7b). The task consisted of three practice trials with feedback, and 2/3 had to be correct to continue to the main task. The main task had 14 trials, and difficulty changed in a stepwise manner designed as a 1-up, 2-down adaptive staircase. Which means one correct answer adds one jump, and two wrong answers remove one jump. Trials commenced with a count-down from three to one, and then the stimulus of the frog jumping was shown for 600ms for every jump. The ISI was fixed to 600ms. The final measure of interest was the highest number of correctly repeated consecutive jumps, referred to here as working memory span (“WM_Span”).

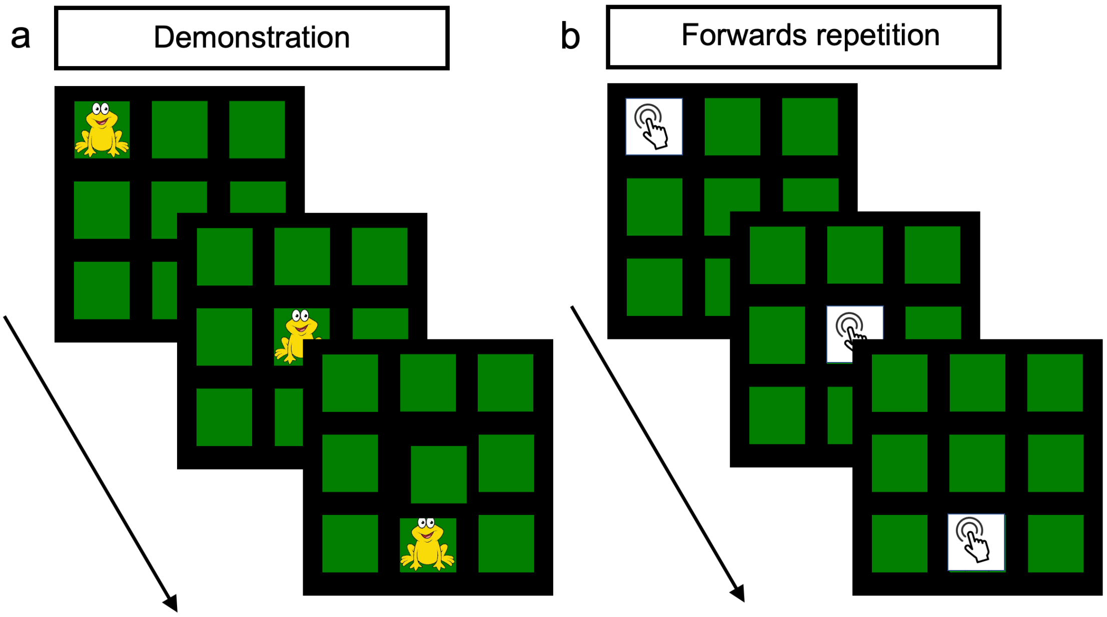

Figure 7. Corsi block tapping task.

For each trial, participants first observed the target stimulus “jumping” between lily pads (a), afterwards, participants were required to repeat the forward sequence of jumping by clicking on the corresponding “lily pads” (b).

N-back task. In addition, the n-back task was used to measure working memory manipulation (Chen et al., 2008). For every trial, participants observed a sequence of dinosaur images (center of the screen). In the 1-back condition, participants had to press the spacebar if the current dinosaur on the screen was the same dinosaur as the dinosaur previously (Figure 8a). In the 2-back condition, participants had to press the spacebar if the current dinosaur on the screen was the same dinosaur as two dinosaurs previously (Figure 8b). Participants completed 80 trials in total, 40 for each n-back condition. Each dinosaur was shown for 500ms and was followed by a 1500ms Inter-Stimulus-Interval (ISI). Responses made before 100ms after stimulus onset were not recorded, and participants had to make their response before the onset of the next stimulus presentation to be within the response window. Final measures included were the d-prime for both the 1-back and 2-back condition (“WM_1back”, “WM_2back”).

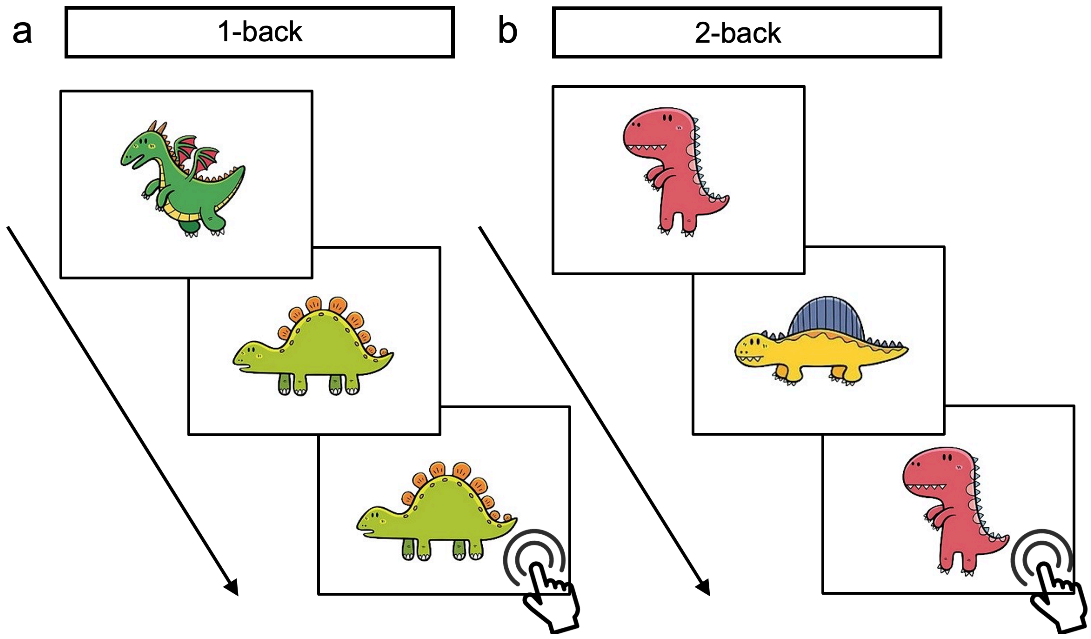

Figure 8. N-back task.

Participants completed two blocks, a 1-back block, and a 2-back block. During the 1-back block (a), participants had to respond by pressing the spacebar if they saw the same dinosaur twice in a row. Stimuli were presented sequentially and only one dinosaur was visible at the time in the center of the screen. During the 2-back block (b), participants had to respond by pressing the spacebar if the current dinosaur was the same dinosaur as two stimuli previously. Participants were instructed to press as quickly as possible.

Intelligence

In addition to the EF tasks, I used two sub-tests of the WASI-II to measure intelligence.

Fluid reasoning. For the fluid reasoning measure, I used the WASI-II Matrix Reasoning subtest (Wechsler, 2011). The Matrix Reasoning subtest was conducted offline and one-on-one by a researcher with a participant and a WASI-II booklet, but after the Covid-related lockdown, was administered online via PyschoPy (Peirce, 2007). Participants were asked to choose the image from five options that would complete the missing picture in a sequence of images (Figure 9). The task measured pattern recognition, and the correct missing image completed or adhered to the pattern visualized in the sequence of images. The task continued until the participant had three consecutive incorrect answers, or until they attained the maximum number of items for their age group. Afterwards, their raw scores were converted to standardized scores by age as instructed in the WASI-II manual (“WASI_Matrix”).

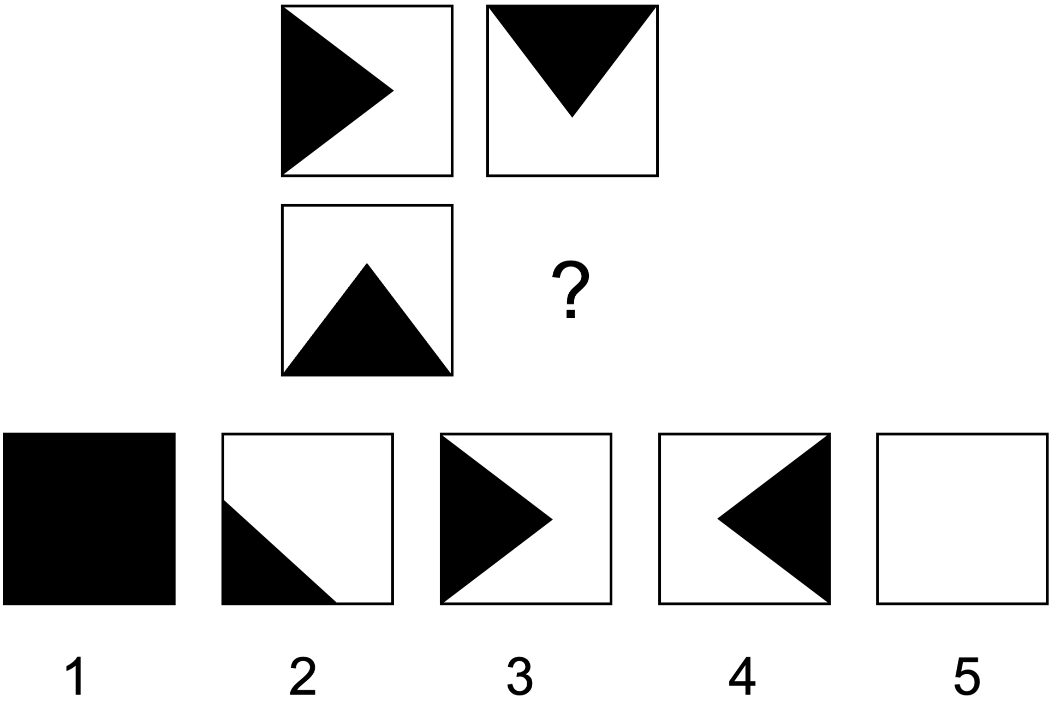

Figure 9. Matrix reasoning example.

Toy example of a matrix reasoning problem. Participants had to complete the sequence by pressing the image that best fits from five options displayed at the bottom. This example uses a simple rotation rule, and the correct answer would be 4.

Crystallized intelligence. The WASI-II vocabulary subtest was used to measure crystallized intelligence (Wechsler, 2011). The Vocabulary subtest was only conducted offline and one-on-one by a researcher with a participant and a WASI-II booklet. This task was not part of the online battery. Participants were asked to describe a word, for example “shirt”, which had several two-point (e.g., “clothing”), one-point (e.g., “keeps you warm”), and zero-point answers (e.g., “*points at shirt*”) (Figure 10). The task continued until the participant had three consecutive zero-point answers, or until they attained the maximum number of items for their age group. Afterwards, their raw scores were converted to standardized scores by age as instructed in the WASI-II manual (“WASI_Vocab”).

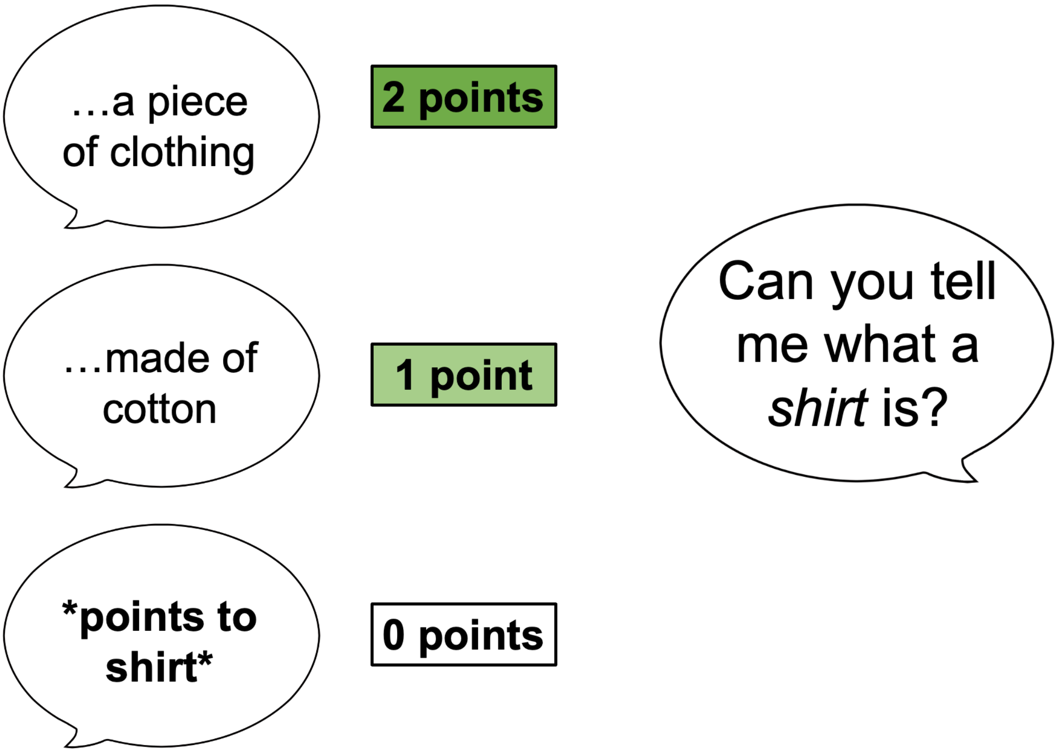

Figure 10. Crystallized intelligence example.

Participants were asked to explain what a word meant. In this case, the prompt was “shirt”. In the text balloons are the left examples of 2-point, 1-point and 0-point answers are depicted.

Two-step Task Design (additional information)

To be manageable for the younger children in our sample, our task consisted of 102 trials (compared to 140 trials in Smid et al. 2022). We conducted parameter recovery analyses of the current task with 100 trials, to ensure that the model-based contribution (*w)* parameter had good recoverability for the trial numbers completed by participants in our sample.

The amount of treasure that could be collected from each planet ranged between 0-9 treasure pieces and changed independently throughout the game following a Gaussian random walk with a standard deviation of two, see Figure 1b. Such drifting reward rates have been shown to promote learning and continuous monitoring of rewards won at each planet, in essence allowing a model-based system to capitalize on faster changes in rewards compared to the traditional two-step task (Kool et al., 2016; for full details on the task such as timings, see Smid et al. 2022).

Instruction Phase

Before starting the main task, all participants completed an identical instruction phase, which took approximately 20 minutes. The main task itself took approximately 20 minutes to complete. During the instruction phase participants learned a) the deterministic transition structure (e.g. that one spaceship always went to the same planet; see Figure 1a), and participants were required to pass a criterion of four correct consecutive transitions to the red and purple planet respectively to continue the task, b) that the amount of treasure changed over time (the drifting reward rates; see Figure 1b), c) how to progress through a trial (e.g. first choose a spaceship, then collect treasure at a planet), and d) the difference between high- and low-stake trials. This phase was identical for children and adults. No rewards were gained during the instruction phase and practice trials were not used for further analysis. For more details on the instruction phase, see the *Supplementary Material*.

After the instruction phase, participants were told they would go on four missions to collect treasure during the main part of the experiment. Children were told that the more treasure they collected in the game, the bigger their present would be at the end of the study (Smid et al., 2020).

We examined participants' understanding of the task by asking them to report the deterministic transition structure of the spaceships to the planets after the preparation phase. Children seemed to learn the task structure well, as 96% of children accurately reported the task structure after the instruction phase. Missed trials were excluded from the analysis as participants did not receive rewards on these trials and therefore could not learn from them. Previously, participants were excluded if they missed more than 30% of the trials. For the current study, children missed 0.05% of the trials on average, and the highest percentage of trials missed was 17%. Thus, no participants had to be excluded from the analysis.

Dual reinforcement Learning Modeling Approach

Computational Model

We used an established dual-systems reinforcement learning model, which has been tested previously with adults (e.g., Daw et al., 2011; Kool et al., 2016, 2017), and with a developmental sample (Smid et al. 2022) to estimate the parameter solutions used to determine the degree of model-based decision making in the behavior of the participants. Model-fitting was conducted using the *mfit* package in Matlab (Gershman, 2018). In computational models, *priors* can be used which are values used to initialize the parameters of a model. If priors are kept "vague", they do not influence the parameter solution strongly, and only have a minimal effect on parameter solutions. Using priors helps with the accuracy of model-fitting, and we therefore used the same vague priors as used in a previous study investigating age effects in model-based decision making and metacontrol in aging adults (Bolenz et al., 2019; Gershman, 2016), and our recent developmental study (Smid et al., 2020). We used Beta(2,2) priors for all model parameters bounded between 0 and 1 (learning rate (α), eligibility trace (λ), and the mixing weight(s) w), and a Gamma(3,0.2) prior for the inverse Softmax temperature (β), and for the two choice stickiness parameters (π and ρ) we used Normal(0,1) priors (Bolenz et al., 2019).

The paradigms consist of four states across two stages, (the two pairs of spaceships and the two planets), with two available actions at the first-stage states between the spaceships (a_A and a_B), and one action at the second-stage state, to collect the treasure (a_C). The reinforcement-learning model consists of a model-based and a model-free system that both learn different values for actions and states, denoted as $Q(s, a)$, which map each state-action pair to its expected discounted future return. On trial t, the first-stage state is denoted by $s_{1,t}$, the second-stage state by $s_{2,t}$, the first and second stage actions by $a_{1,t}$ and $a_{2,t}$, and the first and second stage rewards as $r_{1,t}$ (which is always zero, since only on the second stage reward is attained) and $r_{2,t}$.

Model-free agent. The model-free agent relies on the state-action-reward-state-action (SARSA) temporal difference learning algorithm, which uses reward prediction errors, the learning rate, and the eligibility trace to update the values for each state-action pair *(s,a)* at stage *i* and trial *t* according to:

$$Q_{MF}(s,a) = Q_{MF}(s,a) + \alpha\delta_{i,t}e_{i,t}(s,a)$$

where

$$\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t})$$

Is the reward prediction error for trial *t* at stage *i*, *a* is the learning rate parameter, which determines to which degree new information is incorporated, and $e_{i,t}(s,a)$ is an eligibility trace parameter, and which is set equal to 0 at the beginning of each trial and updated according to:

$$e_{i,t}(s_{i,t}, a_{i,t}) = e_{i-1,t}(s_{i,t}, a_{i,t}) + 1$$

before the *Q* value update. The eligibilities of all state-action pairs are then decayed by *λ* after the update.

For the current paradigm, this learning rule applies in the following way. The reward prediction error is different for the first two levels of the paradigm. Since at the first stage where they choose the spaceships, there is no reward, $r_{1,t}$ is always zero. The reward prediction at the first stage is instead driven by the value of the selected second stage action $Q_{MF}(s_{2,t},a_{2,t})$:

$$\delta_{1,t} = Q_{MF}(s_{2,t}, a_{2,t}) - Q_{MF}(s_{1,t}, a_{1,t})$$

This means that the predicted reward from choosing the spaceships is tied to the reward attained at the planet stage. Since there is no third stage, the second stage prediction error is driven by the reward $r_{2,t}$:

$$\delta_{2,t} = r_{2,t} - Q_{MF}(s_{2,t}, a_{2,t})$$

Both the first- and second-stage values are updated at the second stage, with the first-stage values receiving a prediction error that is down-weighted by the eligibility trace decay *l*. When *l* = 0, only the values of the current state get updated, rather than the values in the past.

Model-based agent. The model-based agent uses the same reward prediction errors and learning rate as the model-free agent, but in addition, uses the transition map of the paradigm to calculate values of each choice. For this paradigm, it means that a model-based agent, but not a model-free agent, can generalize over choices in the two different starting states. To get an intuition for how this leads to different forms of behavior, say, for example, that a participant chooses the blue spaceship, which then transitions to the red planet, and this leads to a large reward. In the next trial, the participant is presented with the other starting state, the one that does not have the previously chosen blue spaceship. Now, the model-based system will realize that the orange spaceship also transitions to the red planet, and because it has just learned that this planet has become better, it will increase its preference for this choice option. A model-free agent is not able to make such generalizations since it relies on direct learning from action-reward contingencies. Therefore, it will not be more likely to pick the orange spaceship over the light blue spaceship in the other starting state. In short, a model-free agent would generate four separate values for all the spaceships, while a model-based agent would only generate two, correctly learning that two spaceships transition to the same planet.

The model-based values are defined in terms of the Bellman's equation, which specifies the expected values of each first-stage action using the transition structure *P*, which means knowing how the spaceships transition to the planets, and which is assumed to be known to the agent:

$$Q_{MB}(s_A, a_j) = P(s_B|s_A, a_j) \underset{a\in\{a_A, a_B\}}{max} Q_{MF}(s_B, a) + P(s_C|s_A, a_j) \underset{a\in\{a_A, a_B\}}{max} Q_{MF}(s_C, a)$$

where we have assumed these are recomputed at each trial from the current estimates of the transition probabilities and second-stage reward values.

Decision rule. To connect the model-based and model-free values to choices, the Q-values are then mixed according to a weighting parameter *w*:

$$Q_{net}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1 - w)Q_{MF}(s_A, a_j).$$

Where a value closer to 1 means the agent is more model-based, and a value closer to 0 means the agent is more model-free. To accommodate our stake manipulation, we defined two different weights that operated on different trial types. We set $w = w_{low}$ on low stake trials and $w = w_{high}$ on high stake trials.

In the second stage, the decision is made using only the model-free values. We used the Softmax rule to translate the weighted Q-values to actions. This rule computes the probability for an action, reflecting the combination of the model-based and model-free action values weighted by an inverse temperature parameter. At both states, the probability of choosing action *a* on trial *t* is computed as:

$$P(a_{i,t} = a|s_{i,t}) = \frac{exp(\beta Q_{net}(s_{i,t}, a))}{\sum_{a\prime} exp(\beta Q_{net}(s_{i,t}, a\prime))}$$

where the inverse temperature *b* determines the randomness of choice or the exploitation/exploration trade-off. Specifically, when *b* approaches infinity, the probability of choosing the action with the highest expected value tends to be 1, whereas, for *b* approaching 0, the probabilities over actions become equally likely across all options. The indicator variable *rep(a)* is defined as 1 if *a* is a first-stage action (choosing a spaceship) and is the same one as was chosen in the previous trial, so the participant chose the same rocket, zero otherwise. Multiplied with the ‘stickiness’ parameter *p*. This captures the degree to which participants show perseveration (when $p > 0$) or switching ($p < 0$) at the first stage. The indicator variable *resp(a)* is defined as 1 if *a* is a first-stage action selecting the same response key as the key that was pressed on the previous trial, zero otherwise. Multiplied with the response stickiness parameter *r*, this captures the degree to which participants repeated ($r > 0$) or alternated ($r < 0$) key presses at the first stage (e.g., whether they pressed the left key twice in a row). These two stickiness parameters were used since the locations of the spaceships changed per trial, and participants could therefore show perseveration or alternation bias towards the spaceships, button presses, or both.

Model-fitting Procedure

We used maximum *a posteriori* estimation, implemented using the *mfit* toolbox (Gershman, 2018), to fit the parameters for the 6 (dual-systems reinforcement learning model with one mixing weight) and 7-parameter (dual-systems reinforcement learning model with two mixing weights per stake) computational models to observed data. To avoid local optima in the estimation solution, the optimization was run 100 times for each participant with randomly selected initializations for each parameter.

Priors

We used identical priors as used in a previous study investigating model-based decision-making across different environmental contexts in an older adult sample (Bolenz et al., 2019). We used Beta(2,2) priors for all model parameters bounded between 0 and 1 (learning rate (α), eligibility trace (λ), and the mixing weight(s) *w*), and a Gamma(3,0.2) prior for the inverse Softmax temperature (β), and for the two stickiness parameters (π and ρ) we used Normal(0,1) priors.

Parameter recovery

To test whether the 7-parameter reinforcement learning model was capable of reliably identifying the contributions of both model-free and model-based decision-making on the task, we conducted parameter recovery for the 7-parameter model, by running the generative version of the model for 500 agents and for 100 trials. For each agent we randomly sampled the initial parameters from uniform distributions: for all parameters bounded between 0 and 1 (learning rate α, eligibility trace λ, *w*-low, *w*-high) we used U(0,1), for inverse temperature β U(0,2) and for the stickiness parameters π and ρ we used U(-0.5,0.5) (Bolenz et al. 2019, Kool et al. 2016). Next, we used the same model-fitting procedures as for the participant data to estimate the model parameters of the simulated data.

For 100 trials, we found substantial correlations between the estimated parameters for w-low (r = .61) and w-high (r = .60). This indicates that for the trial ranges in our sample, we could extract meaningful parameter estimates for the model-based parameters across stakes. For the other parameters, for 100 trials we found: β: r = .87, α: r = .79, λ: r = .45, π: = .44, ρ: = .58.

MRI Sequence and Analysis

MRI images were obtained with a Siemens 3.0 Tesla Prisma scanner located at the Birkbeck-UCL Centre for Neuroimaging (BUCNI) equipped with a standard whole-head coil. To limit head motion, children were requested to keep their heads as still as possible and foam inserts were placed between the head and head coil to ensure the head was snug in the coil. Visual stimuli were projected onto a screen in the magnet boar that could be viewed via a mirror attached to the head coil. Participants watched cartoons without sound during the acquisition of the structural scan. MRI images were processed with FreeSurfer (Version 6.0.0; http://surfer.nmr.mgh.harvard.edu (Fischl et al., 2002)), which is a software that can label and segment cortex and white matter. After being run through FreeSurfer, all scans were manually visually inspected for quality, and the segmentation was manually corrected in FreeSurfer if needed. Four independent inspectors analyzed the scans, and one final inspector performed a final inspection of all scans. After corrections, scans were re-segmented using FreeSurfer, until quality was adequate, or if it did not reach the final level of acceptance, excluded. Using this method, 44 MRI scans were included, while one scan was left out of further analysis, due to excessive movement or poor segmentation.

After preprocessing, sulcal and gyral features across individual subjects were aligned by morphing each subject's brain to an average spherical representation that accurately matches cortical thickness measurements across participants, while minimizing metric distortion. For whole-brain analysis, thickness data were smoothed using a 10 mm Gaussian kernel before statistical analysis. Selecting a surface-based kernel reduces measurement noise but preserves the capacity for anatomical localization, as it respects cortical topological features (Bernhardt, Klimecki, et al., 2014; Lerch & Evans, 2005).

To create the Region of Interest (ROI) of the DLPFC, the Desikan-Killiany atlas was used (Desikan et al., 2006). This atlas allows automatic division of the cortex into standard gyral-based neuroanatomical regions. This atlas divides the cortex into 34 cortical ROIs in each of the individual hemispheres. We extracted the individual cortical thickness of the ROI that most closely matches the DLPFC in the Desikan-Killiany atlas (ROIs 28 (left) and 64 (right); the Rostral middle frontal cortex) for the ROI analysis.

Cortical thickness data were analyzed using the SurfStat toolbox for Matlab [https://www.math.mcgill.ca/keith/surfstat, (Worsley et al., 2009)]. Linear regression models were used to assess the effects of age, sex, model-based decision making, and metacontrol on cortical thickness at each vertex. Findings from the surface-based analyses were controlled for multiple comparisons using random field theory (Bernhardt, Klimecki, et al., 2014; Bernhardt, Smallwood, et al., 2014; Steinbeis et al., 2012; Worsley et al., 2009). This reduced the chance of reporting a family-wise error (FWE). The threshold for significance was set to a stringent $p < 0.01$.

Mediation analysis was conducted in Python using the Pingouin package (Vallat, 2018).

Supplemental Results

Additional Behavioral Results

Neither model-based decision making ($t(65.18) = -1.20$, $d = -0.29$, $p = .236$), nor metacontrol ($t(60.81) = 1.44$, $d = 0.35$, $p = .155$) differed with sex.

We also investigated the behavioral markers of model-based decision-making and metacontrol. We used generalized linear mixed models to approximate a behavioral model-based decision-making measure, which was the probability of repeating a visit to a planet (stay probability) as a function of reward on the previous trial (Kool et al., 2016; Smid et al., 2022). Using this method, the model-based component consists of a main effect of the previous reward on the probability of staying, whereas the reduced effect of previous reward when the starting state is different (compared to when it is the same) indicates a model-free component (Kool et al., 2016). Previous reward refers to the points won by the participant on the previous trial and starting state similarity refers to whether the current starting state (the rocket pair) is the same as on the previous trial. The influence of previous reward on staying behavior approximates the transfer of experience from one starting state to the other, while the differential influence of previous reward on starting state similarity or difference can reflect a lack of transfer of experience between the starting states. Model-free and model-based systems should therefore generate different influences of starting state, as only the model-based system can effectively generalize over states (Smid et al., 2022). In addition, we included the difference in available reward across the two planets on the previous trial (a proxy of reward history) and stake (high and low stakes), and age as potential predictors of stay probability. We conducted nested model selection to find the best-fitting model to predict stay probability.

The winning model consisted of previous reward and starting state, as well as age and stake. There was a significant main effect of previous reward on stay probability (β = .36, se = .03, z = 13.20, p < .001), indicating a significant effect of the model-based component in the children's behavior. In addition, there was a main effect of stake, indicating that children were more likely to repeat a visit to the same planet for high-stake trials (β = .09, se = .03, z = 3.34, p = .001). There was a significant interaction between previous reward and age, mirroring the computational finding that with age, children showed more influence of model-based decision making (β = .18, se = .03, z = 6.43, p < .001). There was a significant interaction between previous reward and stake, indicating that for the behavioral marker, there did seem to be more model-based decision-making for higher stake trials (β = .06, se = .03, z = 2.26, p = .024). Lastly, there was a significant interaction between stake and age, indicating that with increasing age, children were more likely to repeat a visit to the same planet for high-stake trials (β = .07, se = .03, z = 2.72, p = .007) (Figure 11).

Thus, both computational and behavioral markers indicate that overall model-based decision-making seems to increase with age. Via computational makers, there was no group effect of metacontrol, nor an increase with age. However, using behavioral measures we did observe markers of metacontrol in the behavior of the children, which increased with age.

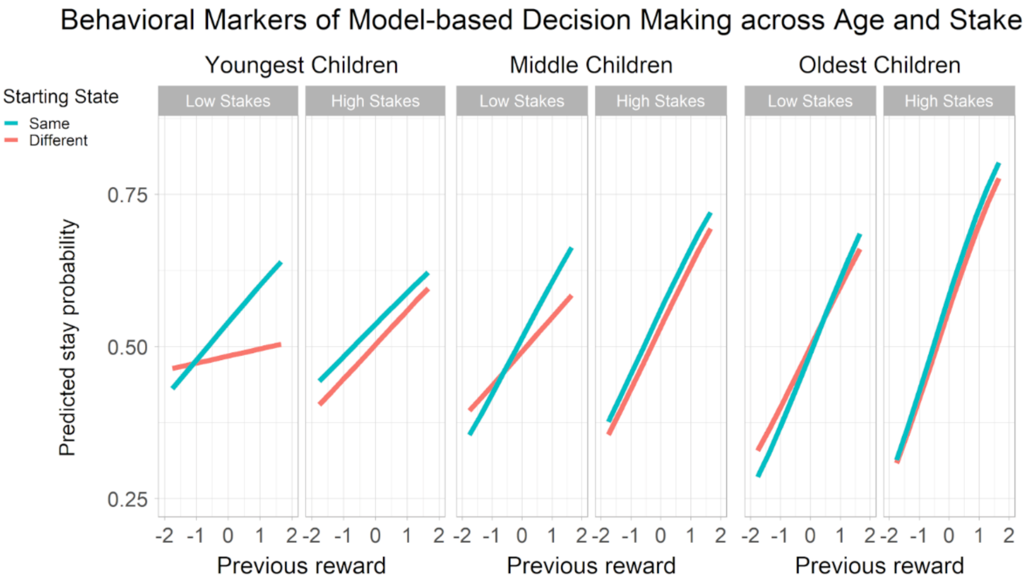

Figure 11. Behavioral markers of model-based decision-making via regression analysis.
The model-based component is reflected in a positive relation between previous reward and stay probability, regardless of starting state. The predicted stay probability is plotted over previous reward, low and high-stake trials, starting state similarity, and age.

Assessing the effect of inhibition on metacontrol and cortical thickness

To assess whether the relationship between cortical thickness and metacontrol is related to measures of inhibition, we included Flanker inhibition as a covariate in the analyses at the whole-brain level and conducted a mediation analysis with the ROI values. When entering performance on the Flanker task as a term in the whole brain model testing the unique effect of metacontrol, while simultaneously controlling for age and sex, all previously reported effects remained significant (Figure 4). In addition, we compared a model with metacontrol, controlled for by age and sex (main model), and a model with metacontrol, controlled for by age, sex, and inhibition (Flanker model), and found that the first model had a numerically better fit (MSE main model = 0.0683, MSE Flanker model = 0.0685, RMSE main model = 0.1367, RMSE Flanker model = 0.1370). The difference in the error sum of squares between the models was significant (t(20483) = 84.09, 95% CI [0.060, 0.063], p < .001). Thus, adding the Flanker term did not improve the accuracy of the model investigating the effect of metacontrol on whole-brain cortical thickness, and the previously found effects for metacontrol remained significant.

References

Bernhardt, B. C., Klimecki, O. M., Leiberg, S., & Singer, T. (2014). Structural covariance networks of the dorsal anterior insula predict females' individual differences in empathic responding. *Cerebral Cortex*, *24*(8), 2189–2198. https://doi.org/10.1093/cercor/bht072

Bernhardt, B. C., Smallwood, J., Tusche, A., Ruby, F. J. M., Engen, H. G., Steinbeis, N., & Singer, T. (2014). Medial prefrontal and anterior cingulate cortical thickness predicts shared individual differences in self-generated thought and temporal discounting. *NeuroImage*, *90*, 290–297. https://doi.org/10.1016/j.neuroimage.2013.12.040

Bolenz, F., Kool, W., Reiter, A., & Eppinger, B. (2019). Metacontrol of decision-making strategies in human aging. *ELife*, *8*. https://doi.org/10.7554/eLife.49154

Chen, Y. N., Mitra, S., & Schlaghecken, F. (2008). Sub-processes of working memory in the N-back task: An investigation using ERPs. *Clinical Neurophysiology*, *119*(7), 1546–1559. https://doi.org/10.1016/j.clinph.2008.03.003

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*(SEP), 1–16. https://doi.org/10.3389/fpsyg.2017.01482

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Farrell Pagulayan, K., Busch, R., Medina, K., Bartok, J., & Krikorian, R. (2006). Developmental normative data for the Corsi Block-Tapping task. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 1043–1052. https://doi.org/10.1080/13803390500350977

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Gershman, S. J. (2018). *“mfit”: simple model-fitting tools*. https://github.com/sjgershm/mfit

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R. A., & Braver, T. S. (2016). Inducing proactive control shifts in the AX-CPT. *Frontiers in Psychology*, *7*(NOV), 1–14. https://doi.org/10.3389/fpsyg.2016.01822

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*(6), 978–990. https://doi.org/10.1111/j.1467-7687.2009.00846.x

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLoS Computational Biology*, *12*(8), 1–34. https://doi.org/10.1371/journal.pcbi.1005090

Lerch, J. P., & Evans, A. C. (2005). Cortical thickness analysis examined through power analysis and a population simulation. *NeuroImage*, *24*(1), 163–173. https://doi.org/10.1016/j.neuroimage.2004.07.045

Matzke, D., Verbruggen, F., & Logan, G. D. (2018). The Stop-Signal Paradigm. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. https://doi.org/10.1002/9781119170174.epcn510

Richmond, L., Redick, T. S., & Braver, T. S. (2015). Remembering to prepare: The benefits (and costs) of high working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1764.

Smid, C. R., Kool, W., Hauser, T., & Steinbeis, N. (2020). *Computational and Behavioral Markers of Model-based Decision Making in Childhood*. 1–32.

Smid, C. R., Kool, W., Hauser, T. U., & Steinbeis, N. (2022). Computational and behavioral markers of model-based decision making in childhood. *Developmental Science*. https://doi.org/10.1111/desc.13295

Steinbeis, N., Bernhardt, B. C., & Singer, T. (2012). Impulse Control and Underlying Functions of the Left DLPFC Mediate Age-Related and Age-Independent Individual Differences in Strategic Social Behavior. *Neuron*, *73*(5), 1040–1051. https://doi.org/10.1016/j.neuron.2011.12.027

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Verbruggen, F., Aron, A. R., Band, G. P. H., Beste, C., Bissett, P. G., Brockett, A. T., Brown, J. W., Chamberlain, S. R., Chambers, C. D., Colonius, H., Colzato, L. S., Corneil, B. D., Coxon, J. P., Dupuis, A., Eagle, D. M., Garavan, H., Greenhouse, I., Heathcote, A., Huster, R. J., … Boehler, C. N. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *ELife*, *8*, 1–26. https://doi.org/10.7554/eLife.46323

Williams, B. R., Strauss, E. H., Hultsch, D. F., & Hunter, M. A. (2007). Reaction time inconsistency in a spatial stroop task: Age-related differences through childhood and adulthood. *Aging, Neuropsychology, and Cognition*, *14*(4), 417–439. https://doi.org/10.1080/13825580600584590

Worsley, K. J., Taylor, J. E., Carbonell, F., Chung, M. K., Duerden, E., & Bernhardt, B. (2009). *SurfStat. a Matlab toolbox for the statistical analysis of univariate and multivariate surface and volumetric data using linear mixed effects models and random field theory* (p. 47).