Supplementary methods

Behavioural modelling

We constructed 32 models that each captured different potential strategies. We considered strategies in which participants performed mental arithmetic to compute the expected value of approaching versus avoiding ("path appraisal by calculation"; models 1 to 9). We considered that participants might forgo mental arithmetic and, instead, learn the overall value of each path from experience and use these cached values to compute expected value ("path appraisal by caching"; models 10 to 27). We also considered strategies in which participants learnt the value of approaching versus avoiding from experience ("Q-learning"; models 38 to 31). Finally, we included a null model in which participants simply had an overall preference towards approaching or avoiding, irrespective of path values or probabilities ("null"; model 32). For strategies where paths were appraised (either by calculation or by caching learned values), each model differed according to whether both paths were considered (models 1, 10, and 19) or only one path was considered (path 1, path 2, a randomly selected path, or whichever path was perceived to be rewarding or punishing). Overall, this model space allowed us to determine whether participants prioritised mentalisation (i.e., a conscious simulation of the sequence) of one path more than another to ease computational burden, which could indicate a potential confound for any observed neural replay during planning.

Rational choice model

In each modelled strategy, a choice to approach or avoid depended on how the expected value of approaching was estimated. In the two-path calculation model (model 1), the expected value of approaching was determined by calculating the cumulative sum of each path (taking into account the odd rule) and weighting the sum by the respective path transition probability, and finally summing the weighted sum of each path:

$$EV_{app} = R_1 P_1 + R_2 P_2 \tag{4}$$

This value was then compared to the expected value of avoiding, EV_{av}, as defined by a threshold parameter, γ:

$$EV_{av} = \gamma \tag{5}$$

The probability of choosing to approach versus avoid was computed by a softmax function parameterised by inverse temperature, τ, to incorporate an element of stochasticity:

$$P_{app} = e^{EV_{app} \cdot \tau} / (e^{EV_{app} \cdot \tau} + e^{\gamma}) \quad (6)$$

Overall, this model represents rational choice behaviour when parameterised with $\gamma = 1$ (as this was the value of choosing to avoid in all trials) and τ approaching ∞ so that choices were entirely guided by value. In our null model (model 28), participants made choices according to an overall preference towards approaching or avoiding, where $EV_{app} = \gamma$ and $EV_{av} = 0$.

Single-path calculation models

Models 2 to 9 differed in how EV_{app} was calculated, such that only one path was taken into account on each trial. This was either path 1 (models 3 and 6), path 2 (models 4 and 7), or a randomly selected path (models 2 and 5). We also varied these models by whether a single threshold parameter, γ, was used on all trials (models 2 to 4) or whether different threshold parameters were used (models 5 to 7) depending on whether the calculated value of a path (EV_{app}) was positive (γ_{pos}) or negative (γ_{neg}).

In models 8 and 9, participants only calculated the path they perceived to be rewarding (model 8) or punishing (model 9). As there was consistency across blocks as to which path was rewarding and which was punishing, participants may have learned this and used it as a strategy to reduce mental arithmetic. On each trial, the learnt value of each path (V) is updated according to:

$$V(i_t) = V(i_{t-1}) + \alpha(R_i - V(i_{t-1})), \quad (7)$$

where i is path 1 or path 2, t is the current trial, α is the learning rate ($0 \leq \alpha \leq 1$), and R is the observed outcome of transitioning to path i. Note that V is set to 0 for all paths at the beginning of the experiment. Choosing to avoid precluded any value updating, as no path was experienced. V was used to select which path to sequentially calculate, as given by $V(i_t) > 0$ in model 8 or $V(i_t) < 0$ in model 9.

Learning path values from experience

Our family of learning models were identical to those specified above, except that the value of paths was learnt from experience rather than by calculating the cumulative sum of points along each path, incorporating the odd rule. Path values were updated either via equation 8 (models 10 to 18), or by equation 9 below (models 19 to 27):

$$V(i_t, j_t) = V(i_{t-1}, j_{t-1}) + \alpha(R(i_t, j_t) - V(i_{t-1}, j_{t-1})), \quad (8)$$

where i is path 1 or path 2, j is the state number the odd rule was applied to along path i, t is the current trial, and R is the observed outcome of a transition to path i with odd rule position j. This was to accommodate a potential strategy in which participants cached specific values for each path depending on which state the odd rule was applied to (i.e., up to 3 values per path).

Learning action values from experience

Models 28 to 31 encapsulated a Q-learning approach, in which the value of making different actions (approach or avoid) in different states (here, the information presented on-screen about the odd rule positions and the path probabilities) is learned over time:

$$Q(s_t, a_t) = Q(s_{t-1}, a_{t-1}) + \alpha(R_t - Q(s_{t-1}, a_{t-1})), \quad (9)$$

where Q is the value of making action a in state s on trial t, and R is the observed outcome (i.e., the number of points gained or lost on a trial). Note that Q was set to 0 for all states at the beginning of the experiment. Models differed by whether only one "state" was used, or whether there were different "states" for: i) each possible path transition probabilities that could be displayed on-screen (10-90%, 30-70%, 50-50%, 70-30%, or 90-10%), ii) each possible combination of positions the odd rule could be applied to ($3^2 = 9$), or iii) a combination of (i) and (ii) giving 45 unique "states".

Parameter optimisation

In all models, γ was bound between -12 and 12, as this was the maximum range of probability-weighted values of any path seen by all participants. Learning rate, α was restricted to $0 \leq \alpha \leq 1$. Inverse temperature, τ, was restricted to $0 < \tau$.

Parameter optimisation was conducted via the 'patternsearch' function in MATLAB without linear constraints, which looks for a minimum based on an adaptive mesh aligned with coordinate directions. To guard against local minima, we conducted parameter optimisation 3^k times where 3 starting values per parameter were uniquely recombined, where k is the number of parameters in a model.

Model fit was evaluated by computing the negative log-likelihood, L, of each parameter-optimised model. Models were compared according to the Bayesian Information Criterion (BIC), as computed by:

$$BIC = -2L + n_p \cdot log(n_t), \quad (10)$$

where n_p was the number of free parameters in a model, and n_t was the number of trials for a participant. BIC scores were computed per participant so that the minimum BIC across models was subtracted from all models. For group-level model comparison, BIC scores were summed over participants.

Results

Model recovery

We assessed the specificity of each model by simulating responses using each of the 32 models (100 iterations each) with a range of different parameter values. We then fit each model to each simulated data set, where a high fit between a simulated data set and the model used to generate that simulated data indicates a true positive.

We observed high accuracy (M = 96.01%, range = 84.10% to 96.92%) and high specificity (M = 0.97, range = 0.84 to 0.99), as well as moderate to high sensitivity (M = 0.70, range = 0.38 to 0.98). The lower sensitivity was driven by the null model, which was sometimes erroneously fit to data generated by other models (**Extended data Fig. 5A**). Importantly, these findings indicate we can be confident in our ability to distinguish between mental arithmetic and learning strategies, as well as between two-path and single-path evaluation strategies.

Expected performance given each strategy

We assessed the maximum performance that could be expected from each of the modelled strategies. To do this, we fit each model to rational choice behaviour — that is, only approaching when the expected value of approaching was greater than 1. We then computed the average accuracy of each model across each experimental protocol.

Using the optimal two-path calculation model as a benchmark, we found that the next best strategy was to learn the value of each path depending on which state the odd rule was applied to ($\alpha = 0.92$) and then only make decisions based on its probability-weighted value against a threshold of 0.91 (mean accuracy = 84.93%, BIC = 204, which is 162 above the next optimal model with BIC 42; **Extended data Fig. 5B**). Thus, the next best strategy after a rational two-path mental arithmetic approach was to learn the value of each path based on visual cues provided on each trial.

A 2-path calculation strategy is the winning model

Finally, we evaluated the fit between different strategy models and each subject's behavioural data. At the group level, the winning model was the optimal two-path calculation model (BIC = 148, 239 less than next best model with BIC 387; **Extended data Fig. 5C**). Thus, participants were more likely overall to be implementing the intended evaluation process than a simpler heuristic. The next best model was a two-path evaluation model where path values were learned from experience according to the position of the odd rule in each path (i.e., caching three values per path in a given block).

At the individual level, 15 of 26 participants were best explained by strategies involving mental arithmetic (**Extended data Fig. 5D**). 10 participants learned path values from experience (7 participants cached one value per path, and 3 participants cached up to three values per path — one per odd rule position). One participant was best explained by a null model, and this was one of the participants who

was excluded from path-specific replay analyses due to low performance. If we group strategies according to whether one or two paths contributed to an expected value calculation, then 20 participants used a strategy in which both paths were evaluated per trial, while the remaining 5 participants (excluding the participant with a winning null model) used a strategy in which only one path was evaluated: just path 1 (2 participants) or just the path learned to be punishing (3 participants, one of whom was the second participant excluded from path-specific replay analyses due to poor performance).

Supplementary tables

Model 1: Choice ~ (Reward magnitude × Loss magnitude × Transition probability) + Certainty + RT + (1 | Subject)

Fixed Effect	**ß**	**SEM**	**p**	
(Intercept)	0.03	0.17	0.852	
Reward magnitude	0.11	0.02	3.220E-8	***
Loss magnitude	0.05	0.02	0.011	*
Transition probability	6.46	0.25	1.187E-145	***
Certainty	0.32	0.07	1.682E-5	***
RT	0.01	0.01	0.477	
Reward magnitude × Loss magnitude	0.01	0.01	0.202	
Reward magnitude × Transition probability	0.50	0.09	2.717E-8	***
Loss magnitude × Transition probability	-0.03	0.09	0.744	
Reward magnitude × Loss magnitude × Transition probability	0.02	0.03	0.464	

Variable inflation factor = 1.03 to 1.55, Durbin-Watson = 1.86, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 2: Sequenceness ~ (Replay type × Choice) + RT + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	0.01	0.00	0.009	**
Replay type	0.01	0.00	5.808E-7	***
Choice	0.00	0.00	0.309	
RT	-0.00	0.00	0.008	**

Replay type × Choice	-0.01	0.00	7.347E-6	***

Variable inflation factor = 1.01 to 3.46, Durbin-Watson = 1.76, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 3: Sequenceness ~ (Recency × Replay type × Transition probability) + RT + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	0.01	0.00	0.006	**
Recency	-0.00	0.00	0.001	***
Replay type	0.00	0.00	0.009	**
Transition probability	-0.00	0.00	0.049	*
RT	-0.00	0.00	0.01	**
Recency × Replay type	0.01	0.00	3.150E-9	***
Recency × Transition probability	0.01	0.00	0.155	
Replay type × Transition probability	0.00	0.00	0.74	
Recency × Replay type × Transition probability	0.01	0.01	0.01	**

Variable inflation factor = 1.02 to 2.05, Durbin-Watson = 1.76, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 4: Sequenceness ~ (Transition probability × Choice) + RT + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	0.01	0.00	0.001	***
Transition probability	-0.00	0.00	0.83	
Choice	-0.00	0.00	0.003	**
RT	-0.00	0.00	0.008	**
Transition probability × Choice	-0.00	0.00	0.19	

Variable inflation factor = 1.01 to 2.49, Durbin-Watson = 1.76, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 5: Choice ~ (Expected value × Differential replay) + Certainty + RT + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	-0.04	0.10	0.735	
Expected value	0.43	0.01	0	***
Differential replay	-0.71	0.13	9.336E-8	***
Certainty	0.27	0.02	3.304E-33	***
RT	-0.00	0.00	0.646	
Expected value × Differential replay	0.13	0.05	0.008	**

Variable inflation factor = 1.02 to 1.05, Durbin-Watson = 1.87, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 6: Choice ~ (Expected value × Rewarding path replay) + (Expected value × Punishing path replay) + Certainty + RT + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	-0.04	0.10	0.735	
Expected value	0.43	0.01	0	***
Rewarding path replay	-1.23	0.19	3.557E-11	***
Punishing path replay	0.19	0.19	0.313	
Certainty	0.27	0.02	5.023E-33	***
RT	-0.00	0.00	0.598	
Expected value × Rewarding path replay	0.12	0.07	0.09	
Expected value × Punishing path replay	-0.15	0.07	0.031	*

Variable inflation factor = 1.01 to 1.05, Durbin-Watson = 1.87

Model 7: Choice ~ (Expected value × Differential replay × Risk-aversion) + (Expected value × Differential replay × Anxiety) + Certainty + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	-0.03	0.10	0.745	
Expected value	0.43	0.01	0	***
Differential replay	-0.62	0.14	6.569E-6	***

Risk-aversion	0.03	0.06	0.644	
Anxiety	-0.07	0.07	0.284	
Certainty	0.27	0.02	4.754E-32	***
Expected value × Differential replay	0.11	0.05	0.028	*
Expected value × Risk-aversion	-0.01	0.00	0.048	*
Differential replay × Risk-aversion	0.38	0.09	1.888E-5	***
Expected value × Anxiety	0.01	0.00	0.168	
Differential replay × Anxiety	0.31	0.11	0.003	**
Expected value × Differential replay × Risk-aversion	0.02	0.03	0.54	
Expected value × Differential replay × Anxiety	-0.10	0.04	0.014	*

Variable inflation factor = 1 to 1.12, Durbin-Watson = 1.87, significance given by a two-tailed statistic using a Satterhwaite approximation

Model 8: Choice ~ (Expected value × Rewarding path replay × Risk-aversion) + (Expected value × Rewarding path replay × Anxiety) + (Expected value × Punishing path replay × Risk-aversion) + (Expected value × Punishing path replay × Anxiety) + Certainty + (1 | Subject/Lag)

Fixed Effect	*ß*	*SEM*	*p*	
(Intercept)	-0.04	0.10	0.725	
Expected value	0.44	0.01	0	***
Rewarding path replay	-1.14	0.19	1.556E-9	***
Risk-aversion	0.02	0.06	0.705	
Anxiety	-0.08	0.07	0.245	
Punishing path replay	0.05	0.19	0.794	
Certainty	0.27	0.02	9.710E-32	***
Expected value × Rewarding path replay	0.12	0.07	0.102	
Expected value × Risk-aversion	-0.01	0.00	0.116	
Rewarding path replay × Risk-aversion	0.22	0.13	0.084	
Expected value × Anxiety	0.01	0.00	0.134	
Rewarding path replay × Anxiety	0.20	0.15	0.189	
Expected value × Punishing path replay	-0.11	0.07	0.13	
Risk-aversion × Punishing path replay	-0.47	0.12	6.340E-5	***
Anxiety × Punishing path replay	-0.45	0.15	0.003	**
Expected value × Rewarding path replay × Risk-aversion	0.13	0.05	0.004	**

Expected value × Rewarding path replay × Anxiety	-0.01	0.06	0.85	
Expected value × Risk-aversion × Punishing path replay	0.08	0.04	0.059	
Expected value × Anxiety × Punishing path replay	0.20	0.05	2.133E-4	***

Variable inflation factor = 1 to 1.16, Durbin-Watson = 1.87, significance given by a two-tailed statistic using a Satterhwaite approximation

Comparison of models containing risk-aversion and/or anxiety as predictors of choice

Model definitions	
Model A	Choice ~ (EV × Replay_differential) + Certainty + RT + (1 | Subject/Lag)
Model B	Choice ~ (EV × Replay_differential) + Risk aversion + Certainty + (1 | Subject/Lag)
Model C	Choice ~ (EV × Replay_differential) + Anxiety + Certainty + (1 | Subject/Lag)
Model D	Choice ~ (EV × Replay_differential) + Risk aversion + Anxiety + Certainty + (1 | Subject/Lag)
Model E	Choice ~ (EV × Replay_differential × Risk aversion) + Certainty + (1 | Subject/Lag)
Model F	Choice ~ (EV × Replay_differential × Anxiety) + Certainty + (1 | Subject/Lag)
Model G	Choice ~ (EV × Replay_differential × Risk aversion) + Anxiety + Certainty + (1 | Subject/Lag)
Model H	Choice ~ (EV × Replay_differential × Anxiety) + Risk aversion + Certainty +(1 | Subject/Lag)
Model I	Choice ~ (EV × Replay_differential × Risk aversion) + (EV × Replay_differential × Anxiety) + Certainty + (1 | Subject/Lag)

Comparison	χ^2	*p*
B vs A	0.122	1
C vs B	0.900	1
D vs C	0.352	0.553
E vs D	19.249	6.610E-5 ***
G vs E	0	1
F vs G	3.172	0.075
H vs F	0	1
I vs H	22.031	6.426E-5 ***

Significance given by an ANOVA comparing all model chi-squared values.

122

123