SI Appendix for:

Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum

Mehdi Keramati[1]*, Peter Smittenaar[2], Ray Dolan[2,3], Peter Dayan[1,3]

[1] Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, UK.

[2] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London. London, London WC1N 3BG, UK.

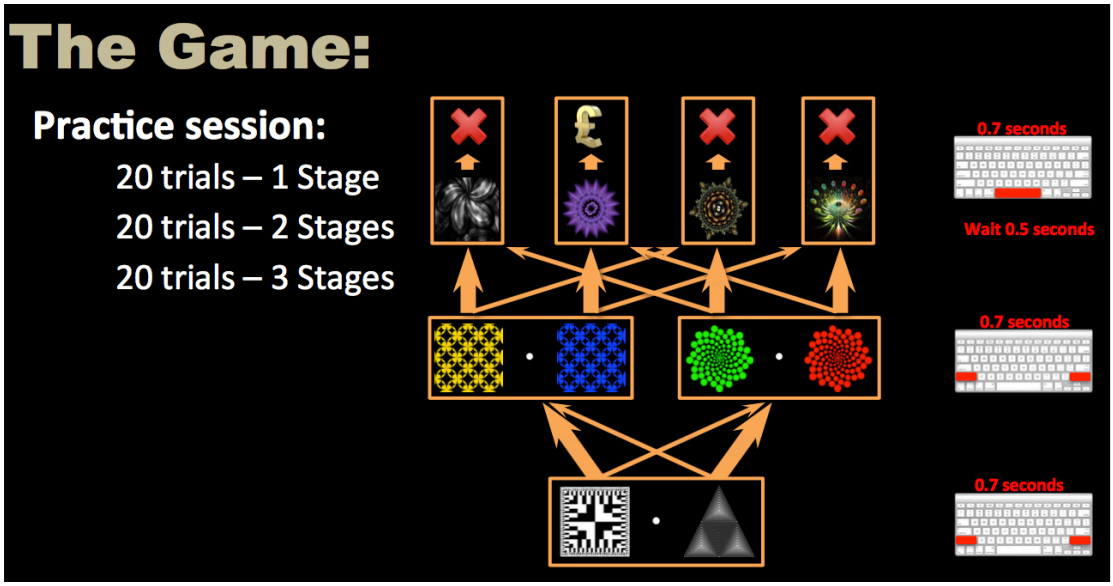
[3] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, UK.

*Correspondence: mehdi@gatsby.ucl.ac.uk

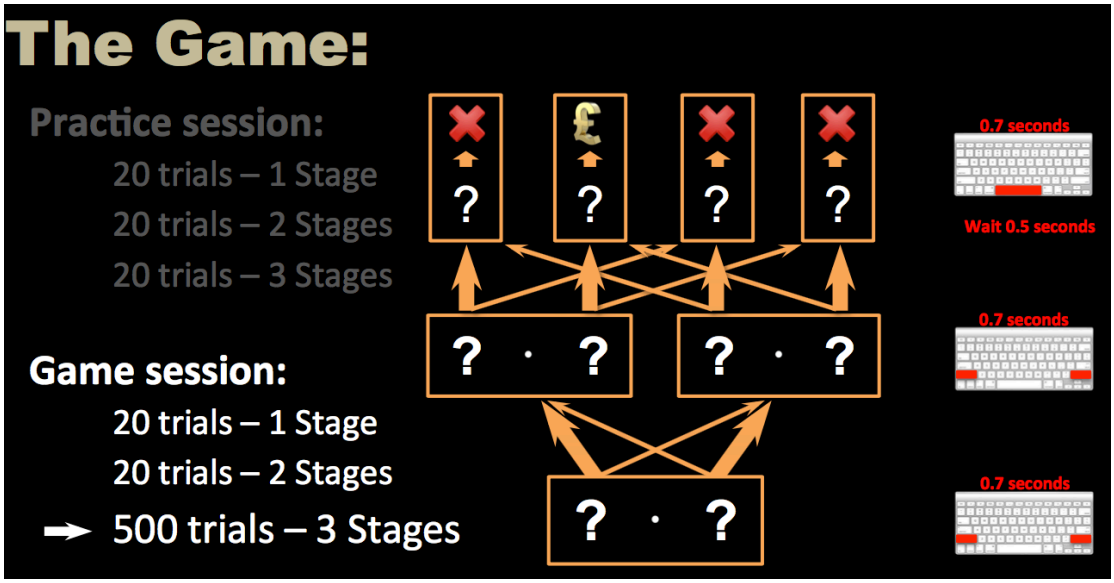
Includes:

- **Supplementary Figures 1-10**
- **Supplemental Information:**
    - Experimental Details
    - Simulations
    - Lagged Logistic Regression Analysis
    - Model-fitting
    - Number of trials

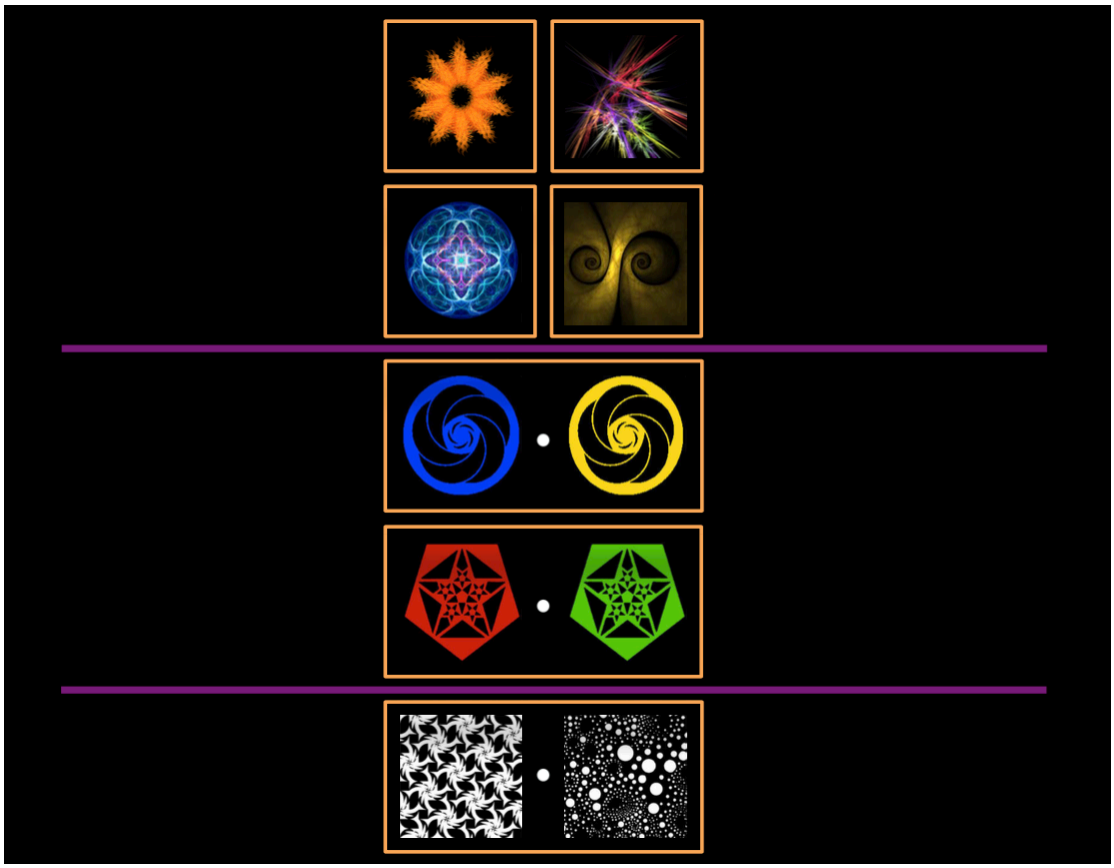
Supplementary figures:



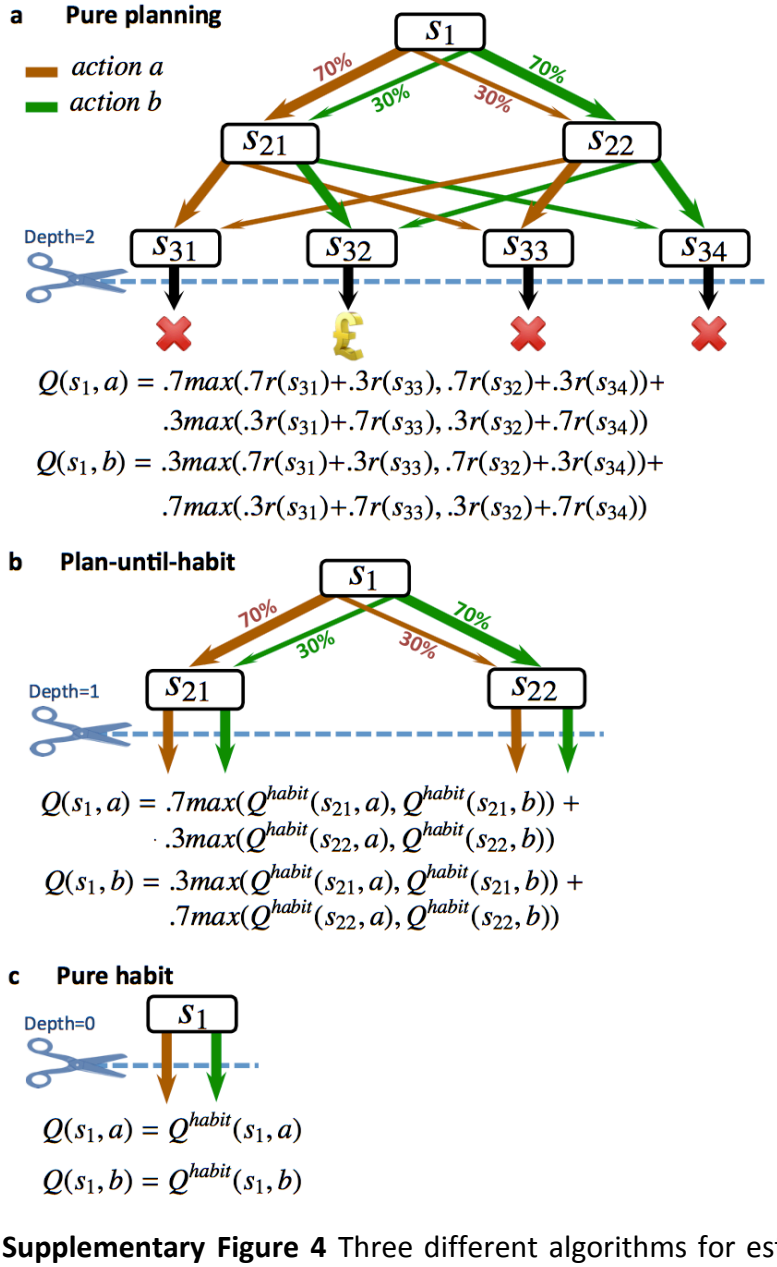
Supplementary Figure 1 During the initial instructions, subjects were shown the tree structure of the practice session they were going to perform.



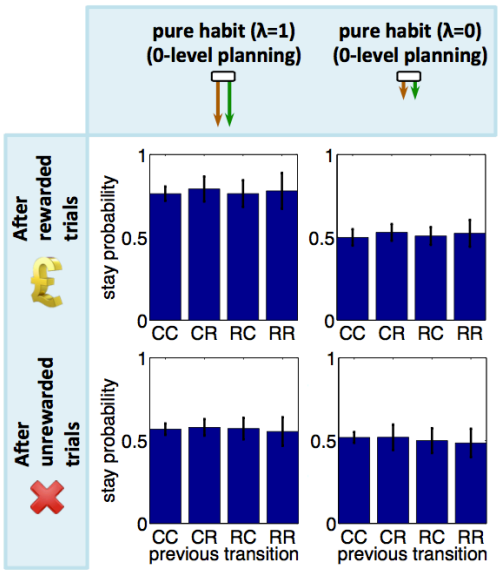
Supplementary Figure 2 During the instructions, subjects were shown the tree structure of the main game session, but not the fractal images corresponding to states and actions.



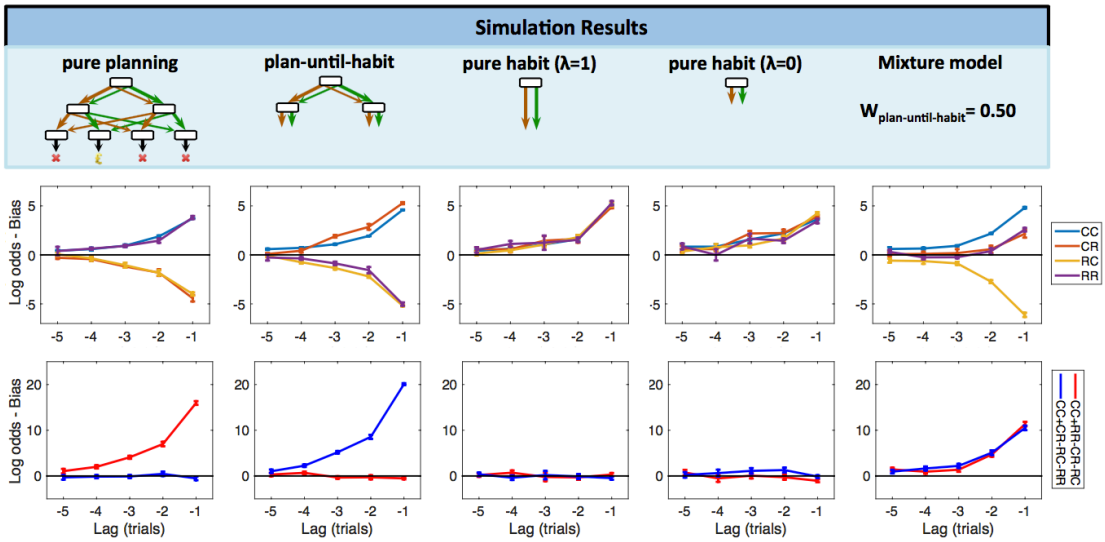
Supplementary Figure 3 Unstructured presentation of the fractal images used during the main game session.



Supplementary Figure 4 Three different algorithms for estimating the value of the two actions at the first stage of the task: (A) Model-based value estimation, (C) model-free value estimation, and (B) integration of model-based and model-free value estimations.

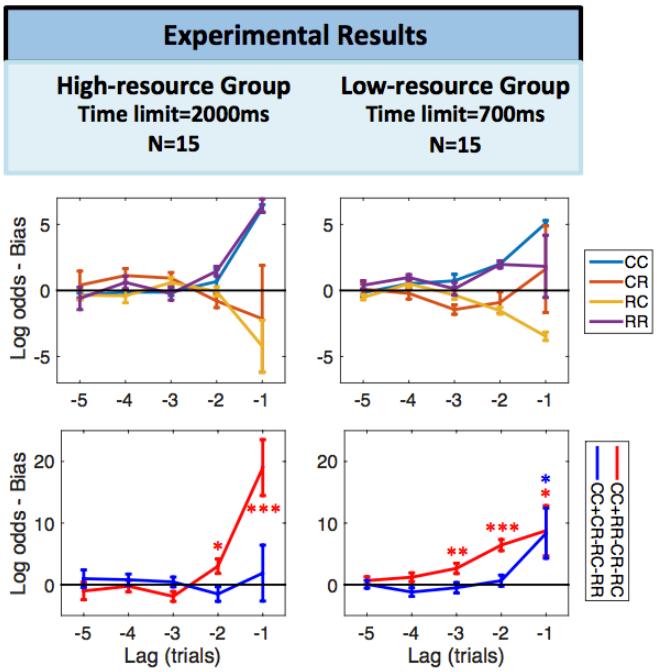


Supplementary Figure 5 Results of Simulating the temporal-difference algorithm with eligibility traces equal to one (left column), or zero (right column) in the task. Both these implementations of a habitual system predict equal stay probability after all transition types (similar to a perseveration bias).



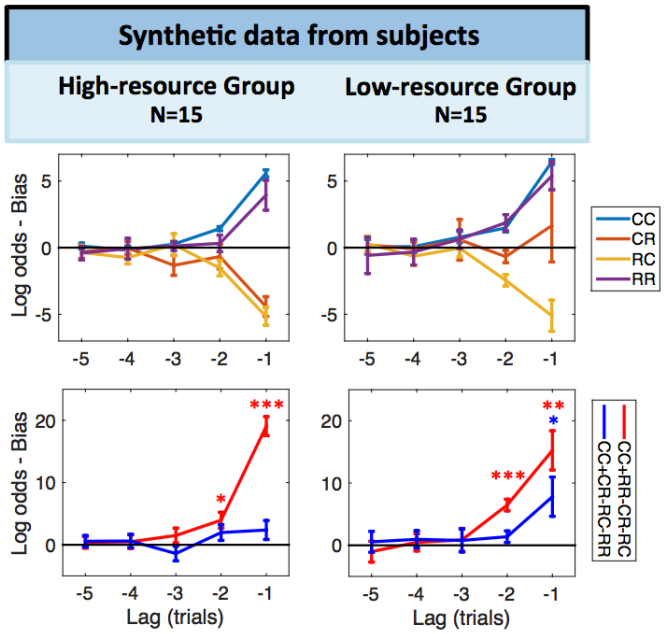
Supplementary Figure 6 Results of the lagged logistic regression analysis over synthetic data generated by simulating a pure MB planner (left column), a plan-until-habit system (second column from left), a MF system with eligibility trace of one (middle column), a MF system with eligibility trace of zero (second column from right), and a mixture of pure planning and plan-until-habit strategies, with equal weights (right column). These results examine the effect on choice probability, of events (i.e., transition types and reward) at several lags relative to the current trial. Regression results (i.e., log-odds) for the first-lag events (*lag=-1*) are consistent with the results of the basic stay-probability analysis (Fig. 3). That is, planning results in a

high probability of repeating the same first-stage action, after CC and RR transitions, but not after CR and RC transitions (assuming the previous trial was rewarded), and so on. However, these predictive weights smoothly faded at increasing lags (*lag<-1*), consistent with the operation of a learning rule.

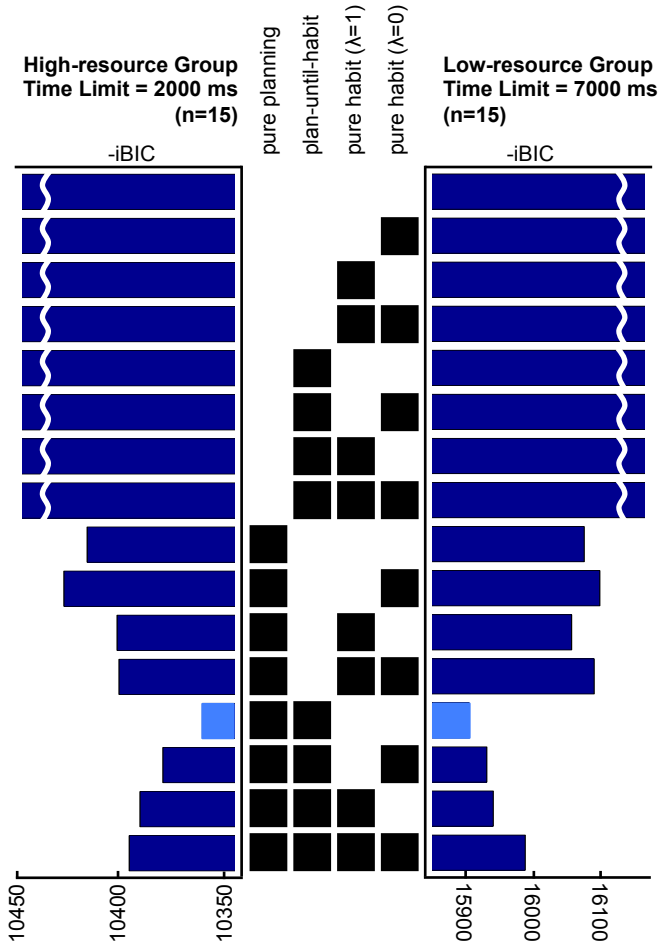


Supplementary Figure 7 Results of the lagged logistic regression analysis over experimental data. At the first lag (*lag=-1*), a significant presence of both pure planning and plan-until-habit strategies was observed in the low-resource group, but only the pure planning strategy in the high-resource group. As in synthetic data, the weights smoothly faded at increasing lags. Surprisingly, the planning effect (red curve in the two bottom panels) fades very fast in the high-resource group, but endures across more lags in the low-resource group. This appears to contract the prediction that time pressure reduces the planning effect. This observation could reflect some intrinsic patterns in the experimental data, or could be simply due to the fact that regression analysis is only an approximation of the dynamic generative model behind the data (e.g., its results are influenced by patterns of non-independence between the regressors at the different lags, which relate in a complex way to both the experimental contingencies and the ways that subjects react to these contingencies. To investigate these possibilities further, we performed the same lagged logistic regression analysis over synthetic data generated by simulating 30 agents (two groups of 15; 350 trials each) where the values of the free parameters for each agent were set equal to the mean of the posterior distribution estimated for that variable from one of the subjects. The results (Fig. S8) showed a similar pattern to those observed in experimental results. This suggests that the

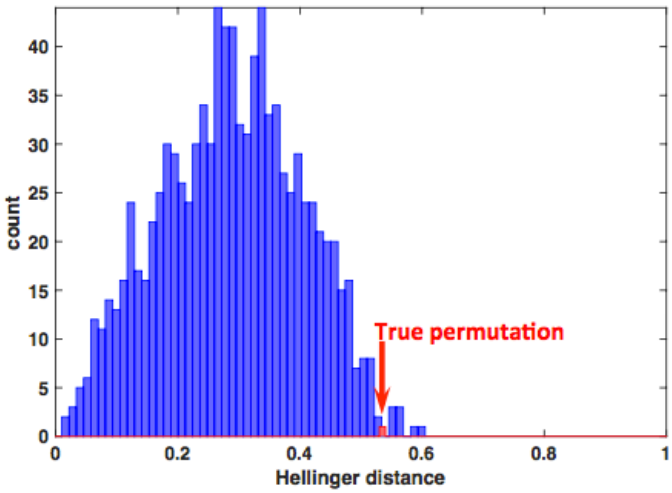
longer-lasting planning effect in the low- recourse as compared to the high-recourse group is an artifact of the simplified assumptions behind regression analysis. There remains one discrepant datapoint – the persistence to lag -3 in the real but not the synthetic data for the low resource group which we leave to future investigation (** p<0.05*; *** p<0.01*; **** p<0.001*).



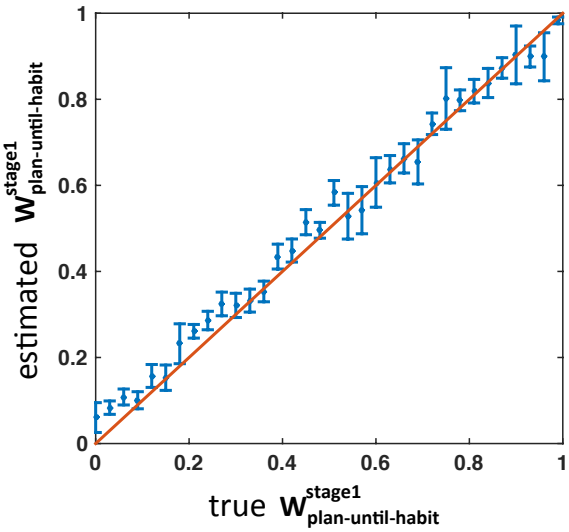
Supplementary Figure 8 Results of the lagged logistic regression analysis over synthetic data generated by simulating the best-fit model to data investigation *(* p<0.05; ** p<0.01; *** p<0.001*).



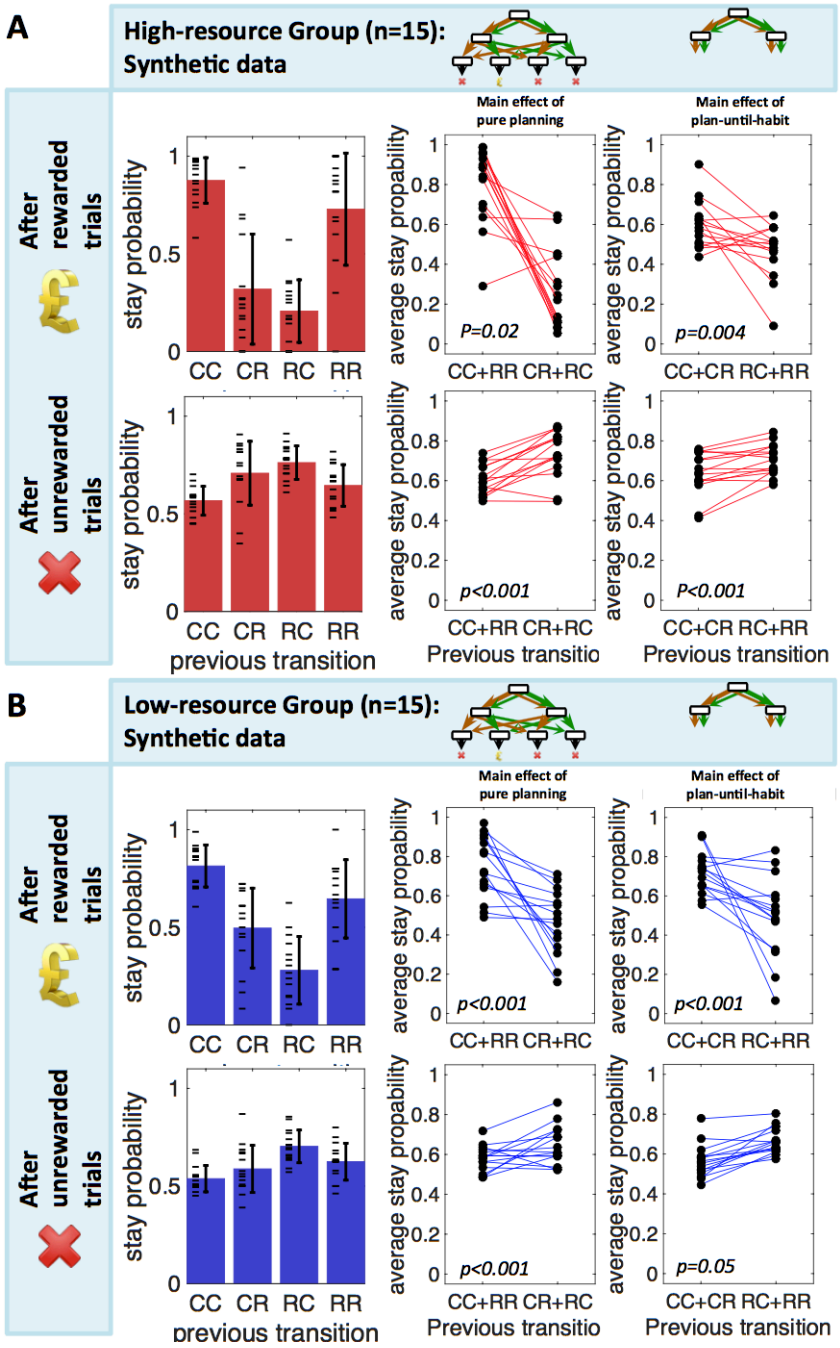
Supplementary Figure 9 iBIC score for different combinations of the four individual models, for the two groups of subjects. The number of free parameters of the models, from top to bottom, are 7, 7, 9, 10, 7, 8, 10, 11, 7, 8, 10, 11, 8, 9, 11, 12.



Supplementary Figure 10 Histogram of Hellinger distance between the distributions over $\boldsymbol{\omega_{1,1}}$ in the two permuted groups of subjects. Each permuted group was a random selection of 15 subjects, out of the pool of the total 30 subjects.



Supplementary Figure 11 Parameter recovery confusion matrix. For every possible value of $0 \leq w^{stage1}_{plan-until-habit} \leq 1$ (33 values between 0 and 1 were tested), a set of synthetic data was produced that consisted of 15 simulated agents, each having 500 trials (as in the experimental resource-limited group). Each synthetic agent was a hybrid model of pure planning and plan-until-habit strategies. When generating the synthetic data, aside from $w^{stage1}_{plan-until-habit}$ that was variable along the 33 sets, the rest of the free parameters of the model were fixed to the parameter values recovered form the experimental resource-limited group (That is, $\alpha_{plan} = 0.8$, $\alpha_{habit(\lambda=0)} = \rho_{habit(\lambda=0)} = 0.55$, $w^{stage2}_{habit} = 0.59$, and $\beta 1 = 8.2$, $\beta 2 = 4.2$). For each tested $w^{stage1}_{plan-until-habit}$, the mean and standard deviation of the recovered $w^{stage1}_{plan-until-habit}$ is plotted.



Supplementary Figure 12 Stay-probability patterns on synthetic data generated by simulating the best-fit model to data. We simulated 30 agents (two groups of 15) where the values of the free parameters for each agent were set equal to the mean of the posterior distribution estimated for that variable from one of the subjects. Posteriors were obtained from the best fit of the models to each group, using hierarchical Bayesian model-fitting. These plots should be compared with those in Fig. 4 in the main paper.

Supplemental Information:

Experimental details:

The whole process, including receiving instructions, performing practice and test sessions, and final payment took between 55 and 65 minutes for each subject.

All subjects received oral instructions (which were read to them from a pre-written manuscript), accompanied by slides. As illustrated in **Supplementary Fig. 1**, they were shown the tree structure of a game they played as a practice session.

This session consisted of three separate phases. In the first phase, subjects had 20 trials of only making the first-stage choice and arriving at the second stage of the task, without experiencing the further stages. In the second phase, subjects had 20 trials of only making the first- and second-stage choices and arriving at one of the four terminal states, without experiencing whether or not reward was present in that state. These two phases were intended to familiarize subjects with the complicated structure of the task gradually, without requiring them to collecting rewarding points. In the third phase of the practice session, subjects experienced the whole game for a further 20 trials, and collected points.

At the first and the second stages of each trial, subject had to press left or right "shift keys" to choose the left or the right fractal shown on the screen, within the time-limit that they were instructed. At the third stage of each trial, after the fractal representing the terminal state appeared, they had to wait for 500 msec and then a white square appeared around the fractal. They then had to press the "space keys" within their instructed time-limit in order to see whether or not that terminal state contained a rewarding point. Subjects were instructed that reacting later than the time-limit would result in missing the relevant trial, which implied losing an opportunity to collect a point. In that case, they received a message indicating they responded too late. In either case, the next trial started after an inter-trial interval of 500 msec.

At each stage of the task, only the fractal(s) representing the subject's current state appeared on the screen. At the first and the second stages of the task where two choices (fractals) were available, the positions (left vs. right) of the fractals associated with the two actions and thus, the corresponding motor-level actions, were chosen randomly across trials (pseudo-counterbalancing).

After performing the practice session, subjects received further instructions about the main session of the experiment. As demonstrated in **Supplementary Fig. 2**, the subjects were not shown the fractal images corresponding to states and choices of the main session. They were instructed that the rules of the game were the same as during the practice session, but that they would experience a whole new set of images.

This session also started with two phases of 20 trials each, exactly as in the training trials, to allow the structure of the tree at the first and the second stages to be gradually acquired. Following that, subjects in the high-resource and the low-resource groups performed 350 or 500 trials, respectively. These were the trials used for all the analyses in the paper. During this last phase, subjects received a break of at least 30 seconds (or longer, if they wanted), after each 50 trials.

During the main session, the fractals representing the two second-stage states, as well as the fractals representing the four terminal states were randomly counterbalanced across subjects. Also, as in the practice session, the positions (left vs. right) of the fractals associated with the two actions at the first and the second stages of the task were chosen randomly across trials (pseudo-counterbalancing).

In order to familiarize subjects with the fractal images used in the main session, they were shown a non-structured presentation of the fractals (**Supplementary Fig. 3**) at the end of the instructions. The transition structure among those state-action pairs, however, was left to subjects to be discovered by experience.

Furthermore, to make it easier for subjects to understand the task, the probability of common vs. rare transitions at both levels was equal to p=0.8 and p=0.2, respectively, during the practice session. These probabilities were p=0.7 and p=0.3 in the main session of the task. Furthermore, we manipulated the pseudo-randomness of the common and rare transitions, in order to avoid perceived non-stationarity in the task structure by subjects. That is, no two consecutive rare transitions were allowed for any state-action pair (during both training and test sessions). Also, in realizing the probabilities p=0.7 and p=0.3 for common and rare transitions, there were always exactly 14 common and 6 rare transition in each block of 20 trials.

Simulations

Four different pure models were simulated in the environment of the task, with parameters equivalent to those used in the test session of the experiment (That is, *p=0.7* and *p=0.3* for common and rare transitions, respectively; reward staying put in a terminal state for $\lfloor X \rfloor_+$ trials, where $X \sim N(\mu = 5, \sigma^2 = 2)$ and $\lfloor . \rfloor_+$ is the positive floor operator).

The first model was a pure MB system (pure planning). The Q-values of the two top-stage actions were computed according to:

$$Q_t^{plan}(s_1, a) = .7max\big(.7r_t(s_{31}) + .3r_t(s_{33}), .7r_t(s_{32}) + .3r_t(s_{34})\big) + .3max\big(.3r_t(s_{31}) + .7r_t(s_{33}), .3r_t(s_{32}) + .7r_t(s_{34})\big) \quad (7$$

$$Q_t^{plan}(s_1, b) = .3max\big(.7r_t(s_{31}) + .3r_t(s_{33}), .7r_t(s_{32}) + .3r_t(s_{34})\big) + .7max\big(.3r_t(s_{31}) + .7r_t(s_{33}), .3r_t(s_{32}) + .7r_t(s_{34})\big)$$

The rewarding value $r_t(.)$ of the four terminal states were initialized to 0.25, and were updated according to:

$$for\ all\ s_3 \in \{s_{31}, s_{32}, s_{33}, s_{34}\}: \quad (8$$

$$r_{t+1}(s_3) = \begin{cases} r_t(s_3) + \alpha_{plan}\big(R_t - r_t(s_3)\big) & if\ s_3 = s_{3,t} \\ r_t(s_3) - \alpha_{plan}\big(R_t - r_t(s_3)\big) * \dfrac{\big(1 - R_t - r_t(s_3)\big)}{\sum_{s_3 \neq s_{3,t}}\big(1 - R_t - r_t(s_3)\big)} & otherwise \end{cases}$$

where R_t was equal to one or zero, depending on whether or not a reward was received on the experienced terminal states, $s_{3,t}$. Furthermore, α_{plan} was the learning rate for the MB system. This updating rule changes the rewarding value of the visited terminal state in the same direction as the experienced reward (i.e., increases if rewarded, decreases if not rewarded), and updates the rewarding value of the other three terminal states in the opposite direction. This later counterfactual part is justified by the fact that subjects know that the reward is always in one, and only one, of the terminal states. Thus, if it was experienced in the visited terminal, the value of the other states should decrease, and vice versa.

The second model was a plan-until-habit system. The Q-values of the two top-stage actions were computed according to:

$$Q_t^{plan-until-habit}(s_1, a) = .7max\left(Q_t^{habit(\lambda=0)}(s_{21}, a), Q_t^{habit(\lambda=0)}(s_{21}, b)\right) + .3max\left(Q_t^{habit(\lambda=0)}(s_{22}, a), Q_t^{habit(\lambda=0)}(s_{22}, b)\right) \quad (9$$

$$Q_t^{plan-until-habit}(s_1, a) = .3max\left(Q_t^{habit(\lambda=0)}(s_{21}, a), Q_t^{habit(\lambda=0)}(s_{21}, b)\right) + .7max\left(Q_t^{habit(\lambda=0)}(s_{22}, a), Q_t^{habit(\lambda=0)}(s_{22}, b)\right)$$

Where $Q_t^{habit(\lambda=0)}(s, a)$ denotes the Q-value of state-action pair (s, a) cached in a MF system with an eligibility trace of zero ($\lambda = 0$). These Q-values for the state-action pairs at the second stage of the task were updated according to:

$$for\ all\ s_2 \in \{s_{21}, s_{22}\}\ and\ a_2 \in \{a, b\}: \quad (10$$

$$Q_{t+1}^{habit(\lambda=0)}(s_2, a_2) =$$

$$\begin{cases} Q_t^{habit(\lambda=0)}(s_2, a_2) + \alpha_{habit(\lambda=0)}(R_t - Q_t^{habit(\lambda=0)}(s_2, a_2)) & if\ s_2 = s_{2,t}\ and\ a_2 = a \\ Q_t^{habit(\lambda=0)}(s_2, a_2) + \rho_{habit(\lambda=0)}(0 - Q_t^{habit(\lambda=0)}(s_2, a_2)) & otherwise \end{cases}$$

Where $s_{2,t}$ and $a_{2,t}$ were the state experienced and action taken in the second stage, respectively. Furthermore, $\alpha_{habit(\lambda=0)}$ was the learning rate for the MF system with zero eligibility trace, and $\rho_{habit(\lambda=0)}$ was the rate of the diffusion of Q-values to zero for the MF system with zero eligibility trace.

The third system was a MF system with eligibility trace of one. The Q-values of the two top-stage actions were initialized to zero, and were updated according to:

$$for\ all\ a_1 \in \{a, b\}:$$

$$Q_{t+1}^{habit}(s_1, a_1) = \begin{cases} Q_t^{habit(\lambda=1)}(s_1, a_1) + \alpha_{habit(\lambda=1)}(R_t - Q_t^{habit(\lambda=1)}(s_1, a_1)) & if\ a_1 = a_{1,t} \\ \rho_{habit(\lambda=1)} Q_t^{habit(\lambda=1)}(s_1, a_1) & otherwise \end{cases} \quad (11$$

Where $\alpha_{habit(\lambda=1)}$ was the learning rate for the MF system with eligibility trace one, and $\rho_{habit(\lambda=1)}$ was the rate of the diffusion of Q-values to zero for that system.

The fourth system was a MF system with eligibility trace of zero. The Q-values of the two top-stage actions were initialized to zero, and were updated according to:

$$for\ all\ a_1 \in \{a, b\}:$$

$$Q_{t+1}^{habit(\lambda=0)}(s_1, a_1) = \begin{cases} Q_t^{habit(\lambda=0)}(s_1, a_1) + \alpha_{habit(\lambda=0)}\delta_t & if\ a_1 = a_{1,t} \\ \rho_{habit(\lambda=0)} Q_t^{habit(\lambda=0)}(s_1, a_1) & otherwise \end{cases} \quad (12$$

Where δ_t is a prediction error, computed as following:

$$\delta_t = \max\left(Q_t^{habit(\lambda=0)}(s_{2,t}, a), Q_t^{habit(\lambda=0)}(s_{2,t}, b)\right) \quad (13$$

For simulating each model, a softmax action-selection rule was used:

$$p_t(a) = \frac{e^{\beta Q_t(s_1, a)}}{\sum_{a_1 \in \{a,b\}} e^{\beta Q_t(s_1, a_1)}}$$

$$p_t(b) = \frac{e^{\beta Q_t(s_1, b)}}{\sum_{a_1 \in \{a,b\}} e^{\beta Q_t(s_1, a_1)}} \quad (14$$

We also simulated mixture models, where the Q-values of the two top-stage actions were weighted averages of the Q-values of pure planning and plan-until-habit systems:

$$for\ all\ a_1 \in \{a, b\}: \quad (15$$

$$Q_t^{mix}(s_1, a_1) = \omega_1 Q_t^{plan-until-habit}(s_1, a_1) + (1 - \omega_1) Q_t^{plan}(s_1, a_1)$$

For computing the Q-values of the actions at the second stage of the task, we always used a mixture models where the Q-values were a weighted average of the MB and MF system. Since at this stage, the MB system can only take a depth of 1 for planning, and the MF system can only adopt an eligibility trace of zero, the Q-values could be computed as:

$for\ all\ a_2 \in \{a, b\}:$

$$Q_t^{mix}(s_{2,t}, a_2) = \omega_2 Q_t^{habit(\lambda=0)}(s_{2,t}, a_2) + (1 - \omega_2) Q_t^{plan}(s_{2,t}, a_2) \quad (16$$

Where $Q_t^{habit(\lambda=0)}(s_{2,t}, a_2)$ comes from equation 10, and $Q_t^{plan}(s_{2,t}, a_2)$ can be computed according to:

$for\ all\ a_2 \in \{a, b\}:$

$$\begin{aligned} Q_t^{plan}(s_{21}, a_1) &= .7r_t(s_{31}) + .3r_t(s_{33}) \\ Q_t^{plan}(s_{21}, a_2) &= .7r_t(s_{32}) + .3r_t(s_{34}) \\ Q_t^{plan}(s_{22}, a_1) &= .3r_t(s_{31}) + .7r_t(s_{33}) \\ Q_t^{plan}(s_{22}, a_2) &= .3r_t(s_{32}) + .7r_t(s_{34}) \end{aligned} \quad (17$$

Lagged Logistic Regression Analysis:

As an extension of the classical stay-probability analysis presented above, we also examined the effect on choice probability, of events (i.e., transition types and reward) at several lags relative to the current trial. We used mixed-effect lagged logistic regression analysis on synthetic data generated by simulating the planning, plan-until-habit, and habitual strategies separately (as in **Supplementary Fig. 3a**), as well as a mixture of the planning and plan-until-habit strategies. Regression results (i.e., log-odds) for the first-lag events corroborated the stay-probability patterns inferred from the statistical analysis above (**Supplementary Fig. 6**). The predictive weights smoothly faded at increasing lags, consistent with the operation of a learning rule. When applied to the experimental data, the regression analysis showed, at the first lag, a significant presence of both pure planning and plan-until-habit strategies in the low-resource group, but only the pure planning strategy in the high-resource group (**Supplementary Fig. 7**). As in synthetic data, the weights smoothly faded at increasing lags.

In the lagged Logistic Regression Analysis, we used choice on trial *t* as the dependent variable, and the independent variables were CC, CR, RC, and RR, during the trials *t-1*, *t-2*, *t-3*, *t-4*, and *t-5*. Each variable took values +0.5 or -0.5, if the transition types

at that specific lagged trial was of the corresponding type, and if the actions that were chosen on that trial where actions *a* or *b* (as in **Supplementary Fig. 4**), respectively, and the trial was rewarded. The variable took the value zero, otherwise.

We used Matlab *nlmefitsa(.)* function which fits a nonlinear mixed-effects regression model to data. **Supplementary Fig. 6** shows the results of the lagged logistic regression analysis over synthetic data. **Supplementary Fig. 7** shows the results of the lagged logistic regression analysis over experimental data.

Model-fitting:

The rates of exploration were allowed to take different values for the first- vs. the second-stage choices to capture the potential difference in the nature of exploratory behavior between the two stages. This was motivated by the fact that the Q-values are more extreme at the second stage, as compared to the first stage. In this case, having equal average actual entropy of the choices at the two stages would require a higher exploration rate (lower β) in the second as compared to the first stage. We indeed observed that in both groups, the mean value of the estimated β parameter for the second stage is half or third that of the first stage ($\beta 1 = 8.2, \beta 2 = 4.2$, for the low-resource group; $\beta 1 = 15.1, \beta 2 = 5.3$, for the high-resource group).

As seen above, the β parameters are larger for the high-resource, as compared to the low-resource, group. This could be simply because time pressure increases randomness. This, however, raises a concern that the observed difference in the depth of planning between the two groups could have been affected by this difference in the β parameters between the two groups. In other words, different β parameters could have resulted in observing different planning depths, due to parameter mis-estimation. To test this possibility, we generated synthetic data for two groups of agents (15 agents in each group, 350 trials each for each agent). Whereas the depth of planning was the same in both groups, the β parameters were different. The parameter values were as follows:

Groups 1: $\omega_1 = \omega_2 = 0.5,\ \beta 1 = 8,\ \ \beta 2 = 4.$

Groups 2: $\omega_1 = \omega_2 = 0.5,\ \beta 1 = 16,\ \beta 2 = 8.$

Using the model-fitting procedure used for human subjects, the recovered parameters for these two synthetic groups were as follows:

Groups 1:

$$w = 0.51 \pm 0.01 \qquad \beta 1 = 7.94 \pm 0.24 \qquad \beta 2 = 3.86 \pm 0.13$$

Groups 2:

$$w = 0.53 \pm 0.02 \quad \beta 1 = 15.62 \pm 0.33 \quad \beta 2 = 7.97 \pm 0.25$$

Correct recovery of parameters in this example shows that the difference in the depth of planning between the two groups is not explained by different levels of randomness in choice.

Number of trials:

Since level of fatigue can affect the relative contribution of planning and habit systems, we roughly equated the total time of the sessions equivalent between the low- and high-resource groups. Therefore, low- and high-resource subject were provided with 500 and 350 trials, respectively. However, the amount of experience with the task could also affect the relative contribution of different strategies. To investigate this factor, we repeated the same analyses reported in the manuscript, but now only on the first 350 trials for the low-resource group (thus, the same number of trials in both groups). All statistical tests on the stay-probability patterns produced significance levels equivalent to those reported in Fig 4b. Therefore, the presence of both planning and plan-until-habit systems was significant in the first 350 trials of the low-resource subjects, after both rewarded and non-rewarded trials. For comparing between the two models, we again compared the $w_{plan-until-habit}^{stage1}$ of the best fit model to the first 350 trials in the two groups, using permutation test. $w_{plan-until-habit}^{stage1}$ was significantly different between the two groups ($p = 0.038$).