SUPPORTING INFORMATION

METHODS

Participants

The general cognitive abilities of the sample were assessed using the Identical Picture Test (Ekstrom, French, Harman, & Dermen, 1976) as a marker for cognitive speed and the Colored Progressive Matrices (CPM; Raven, Bulheller, & Häcker, 2002) as a measure of fluid intelligence. The cognitive profile of our sample was reflected in these two measures (CPM: mean raw score = 32.23, SD = 3.363; Identical Picture Test: mean percent correct = .670, SD = .087).

Data analysis

EEG Data. Additional artifacts were rejected based on a maximum admissible voltage step (50 µV), and by a maximum admissible difference between 2 values on a segment (200 µV). For five participants, the data from one to three malfunctioning electrodes (FC1, FC4, P2, P4) were replaced via spherical spline interpolation (Perrin, 1990).

Model-generated choices of “model players”

The choices of the “model players” presented to the participants were generated using a Q learning algorithm (e.g., Sutton & Barto 1998; cf. Burke et al., 2010). Specifically, after a reward *r*, the value *Q* of action *a* in the next trial was calculated according to the delta updating rule:

$$Q_a(t+1) = Q_a(t) + \alpha\,[r(t) - Q_a(t)]$$

where *α* is the learning rate, *r(t)* the reward obtained after performing action *a* and *t* indexes the current trial. The probability of performing action *a* was computed using a softmax function (O'Doherty, 2004):

$$P(a) = \frac{e^{Q_a(t)/\beta}}{\sum_{c \in A} e^{Q_c(t)/\beta}}$$

where *P*(*a*) is the probability of choosing action *a*, *A* is the set of all possible actions and *β* is the temperature parameter that controls the competition between possible choices. The computer-controlled behavior of the model players was associated with the same percentage of probabilistic positive or negative outcomes (80% gains for the good, 20% for the bad choice), like the participants experienced during the individual learning conditions. To ensure comparability between conditions, the mean of the rewards obtained by the model were constrained to small deviations from each condition's true mean with a 2.5% maximum deviation (between 77.5% and 82.5% upon choosing the good option and between 17.5% and 22.5% upon choosing the bad option). The Q-values were set to zero at the beginning of the task and continuously updated on subsequent trials. The *β* and *α* parameters were estimated based on data of 30 subjects that were acquired in a prior pilot testing. Figure 1 C shows the mean learning curve resulting from 10^5-simulated runs with the same model.

REFERENCES

Ekstrom, R. B., French, J. W, Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain:

insights from neuroimaging. *Current opinion in neurobiology*, *14*(6), 769–776.

Perrin, F. (1990). Correction. *Electroencephalography and clinical Neurophysiology*, *76*(6), 565–565.

Raven, J. C., Bulheller, S., & Häcker, H. (2002). *CPM. Coloured Progressive Matrices.* Wien: Hogrefe Austria GmbH.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachussets.