SUPPLEMENTAL METHODS

Probabilistic reversal learning task: design

In our version of the task, each participant completed 80 trials, 40 before and 40 after a single reversal. We believe that this design allows one to meaningfully apply reinforcement learning models for the following reasons. Longer tasks with multiple reversals offer superior detection power for neuroimaging studies (1, 2). However, on multiple-reversal tasks participants often adopt a strategy-based approach (e.g. switch after 3 errors), a phenomenon noted in earlier studies (1, 2), making reinforcement learning models less applicable. Thus, we feel that a task with a single reversal, where participants lack a definite reversal expectation, is best suited for modeling individual differences in trial-by-trial learning. Additionally, our version of the task was not 'adaptive': all participants completed the same number of trials regardless of their performance. Further, all experienced the same trial-by-trial stimulus-reinforcement contingencies.

Model selection

We aimed to identify a model that would both fit the participants' behavior and adequately represent relevant individual differences in its parameters, distinguishing between participants who perform qualitatively differently on the probabilistic reversal learning task. Importantly, in order to test our main hypothesis, the model was required to represent attention to reinforcement history vs. the last trial, as distinct from attention to valence (rewards vs. punishments).

Reinforcement learning model

Optimal expected values of choosing stimulus 1 vs. stimulus 2 using a prior and a prediction error were computed using the Rescorla-Wagner rule (3):

(1) For rewarded trials:

$$E_{\mathrm{s1t}} = memory * E_{\mathrm{s1t\text{-}1}} + learning.rate_{rewards} * (reward_{s1t} - E_{\mathrm{s1t\text{-}1}})$$

For punished trials:

$$E_{\mathrm{s1t}} = memory * E_{\mathrm{s1t\text{-}1}} + learning.rate_{punishments} * (punishment_{s1t} - E_{\mathrm{s1t\text{-}1}})$$

where $E_{s1\ t-1}$ is the expected value of the stimulus from the previous trial, and the difference between the experienced outcome and the previously expected outcome (prior), and $(reward_{s1t} - E_{s1t-1})$ or $(punishment_{s1t} - E_{s1t-1})$, is the prediction error. The **memory** reflects how much one's choice is determined by the reward and punishment history on all previous trials. Low memory results in rapid switches of choice in response to reward/punishment on the last trial. The **learning rate from rewards** and **learning rate from punishments** reflected the impact that reward or punishment on trial $t\text{-}1$ had on the subject's choice on trial t.

Rescorla-Wagner reinforcement learning models are known to learn more slowly than humans or non-human primates (4), partly because they are ignorant of the structure of the environment. By contrast, human subjects performing this task received instructions that one stimulus would be 'correct' while the other would be 'wrong', and thus were able to exploit their knowledge of task structure. Trying to overcome this limitation at least in part, we incorporated the knowledge that stimulus values are reciprocal into our modified Rescorla-Wagner model. The expected value for the non-chosen stimulus (*s2* in this case) was updated reciprocally to the expected value of the chosen stimulus, reflecting the instructions:

(2) $$E_{s2t} = -E_{s1t}$$

In this case, a full update is applied to the unselected stimulus. Since it has been proposed that agents may apply only a partial update to the unselected stimulus (5), we have also tested a **partial double update model**, described below.

The probability of choosing the stimulus based on its expected value was calculated as:

(3) $$P_{s1t} = sigmoid([10 - Exploration] * E_{s1t})$$

where $sigmoid(z)=1/[1+exp(-z)]$. The **exploration** parameter reflects the randomness of choice, varying between 0 and 10. At low exploration parameter values, the choice is mostly based on the reinforcement history represented in $E_{s1\,t}$ (exploitation). When the exploration parameter reaches the value of 10, the agent's choice completely ignores the reinforcement history, becoming completely stochastic. Exploration on probabilistic learning tasks in humans may reflect purely random choice or certain false assumptions about the environment (e.g. that the reinforced stimulus alternates every two trials or is always presented on the left). In general, exploration can be adaptive in a dynamic environment where reinforcement contingencies change unpredictably.

We fitted our model to the subjects' behavioral data using a non-linear gradient descent simplex function (6) which incorporates hard constraints into a Nedler-Mead 'amoeba' optimization algorithm, implemented in MATLAB 7.6.0 (The MathWorks, Inc.). To find the best set of parameters to represent the pattern of subject's choices, we maximized the likelihood function $l(parameters \mid y)$ for each subject, where y is the subject's set of choices. Towards this end, for t^{th} trial, we calculated the probability that the model with a given set of parameters would select the option actually chosen by the subject, $P(choice_t \mid parameters)$. Thus, for all trials:

(4) $$l(parameters \mid y) = \prod_t P(choice_t \mid parameters)$$

Because of numerical precision constraints when calculating the product of very small numbers in MATLAB, we maximized the log-likelihood as recommended by Wallisch and colleagues (7):

(5) $$\log[l(parameters \mid y)] = \sum_t \log[P(choice_t \mid parameters)]$$

Partial double update model

We have tested an alternative model, in which a reward or punishment associated with the *selected* stimulus results in only a partial update of the value of the *unselected* stimulus (5). For example, when punished for selecting $s1$, one may not be completely certain that he/she would have been rewarded for selecting $s2$. Compared to the full double update model described above, this model uses an additional free parameter ε. It reflects the degree to which the unselected stimulus $s2$ is updated as a result of prediction error for the selected stimulus $s1$:

$$E_{s2t} = E_{s2t-1} - \varepsilon * update_{s1t} \tag{2a}$$

where $update_{s1t}$ is $learning.rate_{rewards}*(reward_{s1t} - E_{s1t-1})$ for rewarded trials or $learning.rate_{punish.}*(punishment_{s1t} - E_{s1t-1})$ for punished trials. The probability of choosing the stimulus was calculated as a function of the difference between its expected value and the expected value of the other stimulus:

$$P_{s1t} = sigmoid([10 - Exploration]*[E_{s1t} - E_{s2t}]) \tag{3a}$$

Otherwise, the partial double update model is identical to the full double update model described above, and includes five free parameters: memory, learning rate from rewards, learning rate from punishments, exploration, and ε.

Simplified 3-parameter full double update model

We have also tested a simplified model, which included learning rate from rewards, learning rate from punishments, and exploration as free parameters and fixed the memory parameter at 1. Otherwise, the simplified model was identical to the full double update model above.

Model comparison

To compare the goodness of the best fit between the models, we first calculated the Bayesian information criterion (BIC), which penalizes models with a greater number of free parameters:

$$BIC = -2 \log l + k * \log n \tag{6}$$

where l is the likelihood for the best set of parameters, k is the number of free parameters, and n is the number of trials. To determine the extent to which model parameters captured individual differences between participants, we also examined correlations between these parameters and the following behavioral indices: perseverative errors, switches, and probabilistic switches (switches following non-contingent punishment).

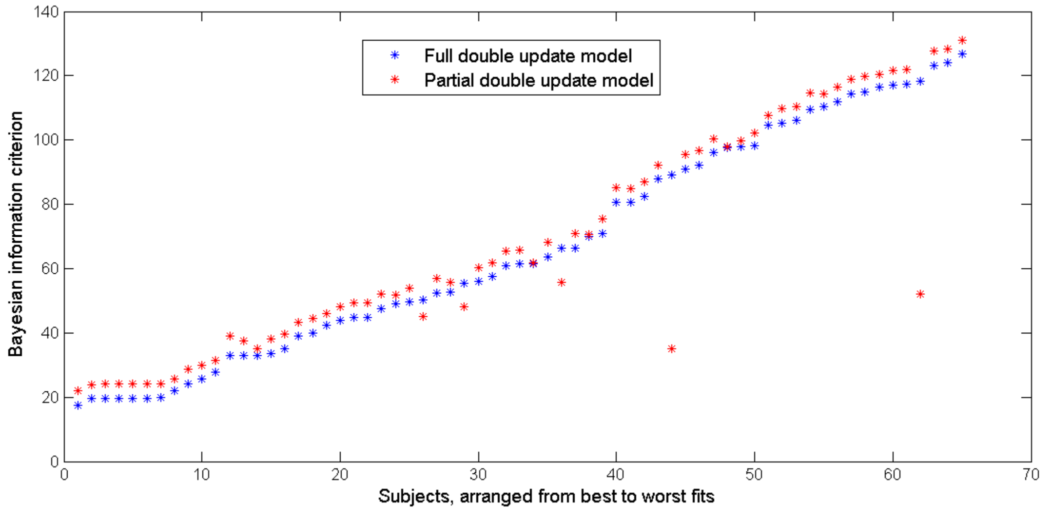
RESULTS

Probabilistic reversal learning: model comparisons

Fits for the **full double update model** were insignificantly better for healthy controls (mean $\log l$=20.0), non-suicidal depressed (mean $\log l$=25.9), and for suicide ideators (mean $\log l$=21.6) than for suicide attempters (mean $\log l$=30.0; F(3,61)=1.0, p=0.40, post-hoc: NS).

Allowing for a **partial double update** did not improve model fits. Overall, fits were similar for both models (Figure S1), and mean Bayesian information criterion (BIC) was higher for the partial double update model (68.7) than for the full double update model (67.1). Thus, we retained the more parsimonious full double update model.

FIGURE S1. Full double update vs. partial double update reinforcement learning models: best fit across subjects



Lower Bayesian information criterion values indicate better fit. Allowing for a partial double update to the unselected stimulus did not improve model fits. Overall, both models fit equally well, and mean Bayesian information criterion was higher for the partial double update model (68.7) than for the full double update model (67.1).

Fits for the **simplified 3-parameter full double update model** were also similar to those for the 4-parameter full double update model, with a slight advantage for the 3-parameter model when considering the number or parameters (mean BIC: 63.9 vs. 67.1).

However, the four parameters of the full double update model with the memory parameter captured the two extreme behaviors on the task (excessive switching and perseveration) better than the three parameters of the simpler model. As we show below, this was due to the fact that in the 4-parameter model excessive switching was captured by the memory parameter and perseveration, by the learning rate from punishments, while in the 3-parameter model both behaviors were captured by a single parameter, the learning rate from punishments. Indeed, the 4-parameter model showed correlations of memory with switches (r=-.58, p<.001) and probabilistic switches (r=-0.70, p<.001) and of learning rate from punishments with perseverative errors (r=-38, p=.002; Fig. 3b, c, e). Meanwhile, in the 3-parameter model, the learning rate from punishments was still correlated with perseverative errors (r=-.34, p=.008) but only weakly correlated with switches (r=.13, NS) and probabilistic switches (r=.26, p=.045). Thus, in the 3-parameter model, the learning rate from punishments captured excessive switching only to a limited extent.

Finally, and perhaps most importantly, the three parameters of the simplified model did not represent the temporal dimension of the task independently from the valence dimension. Thus, such a model would not be suited for testing our main hypothesis. These considerations lead us to retain the 4-parameter model with a memory parameter for subsequent group contrasts.

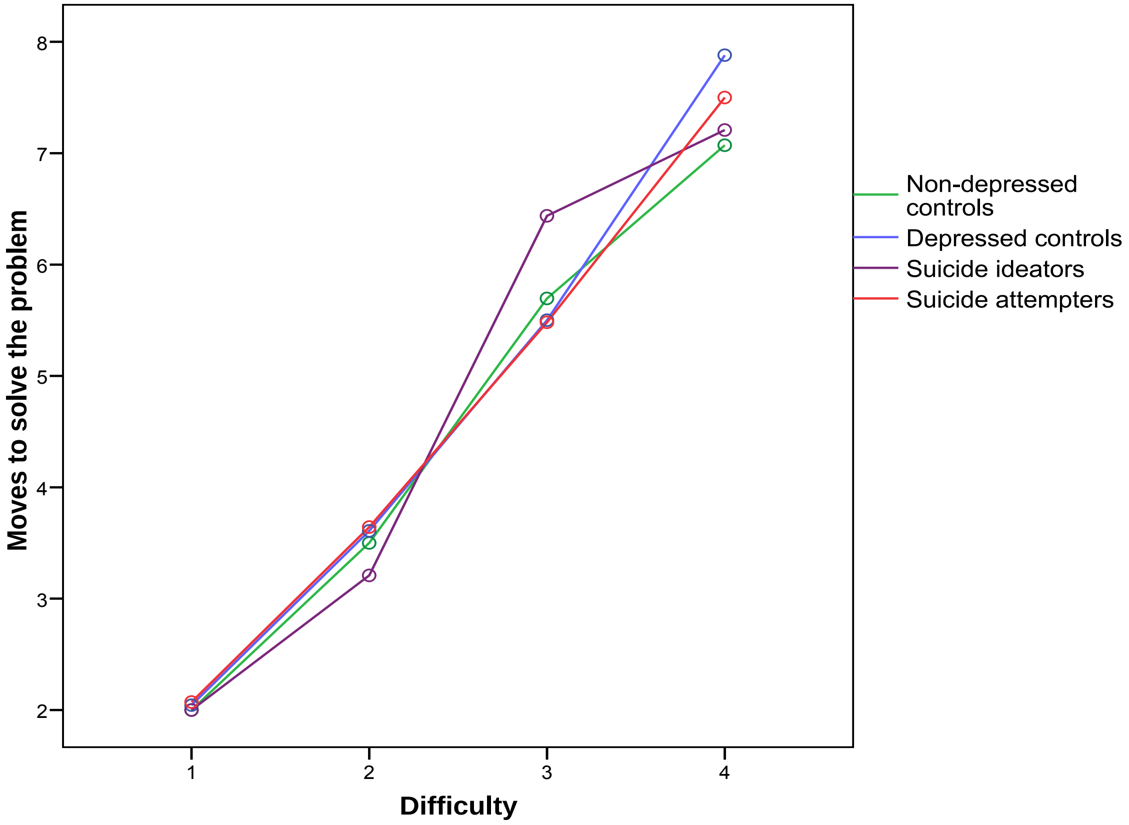
Probabilistic reversal learning: learning rate from rewards and exploration

While group differences were seen in memory and learning rate from punishments (Results, Computational Model Analyses, Fig. 3), **learning rate from rewards** (mean(SD), non-depressed controls (C): 0.42(0.37), depressed non-suicidal (D): 0.44(0.39), suicide ideators(SI): 0.33(0.36), suicide attempters (SA): 0.41(0.38); F(3,61)=0.21, p=0.89) and **exploration** (C: 3.4(3.0), D: 3.3(3.5), SI: 2.7(2.7), SA: 4.3(3.3); F(3,61)=0.59, p=0.62) did not vary across groups.

Forward planning and spatial working memory

The Stockings of Cambridge (SoC) test requires participants to rearrange colored balls in vertical columns to match a desired final arrangement in a specified minimum number of moves. Participants are told to plan their sequence of moves before starting to move the balls shown on the monitor. The time to plan the sequence of moves and the total number of moves made to solve the problem are recorded. Groups did not differ in the number of problems solved in minimum moves (‘perfect solutions’; F[3,61]=1.0, p=0.39, partial η^2=0.05). Figure S2 illustrates SoC performance in the four groups: participants used more moves to solve problems of greater difficulty (F[1,59]=821, p<0.001), however there was no effect of group (F[3,59]=0.33, p=0.80) and no group by difficulty interaction (F[9,59]=1.70, p=0.13). Similarly, the groups did not differ in initial and subsequent deliberation times, stratified by problem difficulty (F[3,59]<1.6, p>0.19).

FIGURE S2. Forward planning and spatial working memory: Stockings of Cambridge test.



On the Stockings of Cambridge test, participants used more moves to solve problems of greater difficulty (F[1,59]=821, p<0.001), however groups did not differ in the number of moves needed to solve the problem (F[3,59]=0.33, p=0.80).

Additional sensitivity analyses: effects of current substance use, age at first suicide attempt, gender, severity of depression, and medication exposure

Since 6/15 suicide attempters and none of the participants in other groups had current substance use disorders, we performed a sensitivity analysis excluding these 6 participants: suicide attempters were still least likely to pass the reversal stage (χ^2=10.2, p=0.017, N=52, N=59; suicide attempters vs. non-suicidal depressed: χ^2=5.0, p=0.026, N=33). Likewise, we performed an additional sensitivity analysis, limiting the group of attempters to only 9/15 who first attempted suicide after age 60. The differences in passing the reversal stage among groups (χ^2=10.2, p=0.017, N=52) and between suicide attempters and non-suicidal depressed elders (χ^2=5.0, p=0.026, N=33) persisted. Of note, 5/6 early-onset attempters vs. only 1/9 late-onset attempters had current substance use disorders (χ^2=7.8, p=0.005, N=15), hence the identical statistics in the two analyses above. Males and females were equally likely to pass the reversal stage (χ^2=1.85, p=0.17). Severity of depression measured by the 16-item Hamilton Depression Rating scale was not related to performance in the reversal stage among the three depressed groups (binary logistic regression: Wald[1]=0.033, p=0.86). Similarly, performance in the reversal stage was not related to the cumulative strength of antidepressant treatment (r=–

0.095, p=0.564) or to sedative (χ^2=1.34, p=0.25), anticholinergic (χ^2=0.42, p=0.84), or opioid (χ^2=0.11, p=0.74) exposure.

References

1. Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. Journal of Neuroscience. 2006; 26:8360-8367.

2. Cools R, Clark L, Owen AM, Robbins TW. Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. J Neurosci. 2002; 22:4563-7.

3. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, editors. Classical Conditioning II: Appleton-Century-Crofts; 1972. p. 64-99.

4. Samejima K, Doya K. Multiple representations of belief states and action values in corticobasal ganglia loops. Ann N Y Acad Sci. 2007; 1104:213-28.

5. Matsumoto M, Matsumoto K, Abe H, Tanaka K. Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci. 2007; 10:647-656.

6. Bajzer Z, Penzar I. SIMPS. Natick, MA: The MathWorks, Inc. ; 1998. p. SIMPS (StrategySimplex)-Constrained minimizer.

7. Wallisch P, Lusignan M, Benayoun M, Baker TI, Dickey AS, Hatsopoulos NG. Matlab for Neuroscientists: An Introduction to Scientific Computing in Matlab. Burlington, MA; San Diego, CA; London, UK: Academic Press; 2009.