UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Modeling Reward Learning Under Placebo Expectancies: A Q-Learning Approach

Permalink

https://escholarship.org/uc/item/6cm876t2

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors

Augustat, Nick
Chuang, Li-Ching
Panitz, Christian
et al.

Publication Date

2022

Peer reviewed

Modeling Reward Learning Under Placebo Expectancies: A Q-Learning Approach

Nick Augustat (nick.augustat@staff.uni-marburg.de)
Li-Ching Chuang (chuangl@staff.uni-marburg.de)
Department of Psychology, University of Marburg, Germany

Christian Panitz (christian.panitz@staff.uni-marburg.de)
Department of Psychology, University of Marburg and University of Leipzig, Germany
Center for the Study of Emotion and Attention, University of Florida, Gainesville, FL, USA

Christopher Stolz (christopher.stolz@lin-magdeburg.de)
Leibniz Institute for Neurobiology, Magdeburg, Germany
Department of Psychology, University of Marburg, Germany

Erik Malte Müller* (erik.mueller@staff.uni-marburg.de)
Dominik Endres* (endresd@staff.uni-marburg.de)
Department of Psychology, University of Marburg, Germany; *equal contribution

Abstract

Although expectancy effects induced by placebo treatment are reported to attenuate depressive symptoms in the long run, mechanisms underlying situational dynamics are not well understood. Improved reward learning has been discussed as a candidate mediator for effects of positive expectancies on more positive mood. Here, we fitted a series of Q-learning models to measure the effect of sham antidepressant treatment vs. open-label placebo in a probabilistic reinforcement learning task. Treatment effects were observed mainly in those Q-learning models justified by the task structure. Additionally, interindividual variability remained the largest origin of unexplained variance in predictive match across models. These findings provide further support for the role of expectancies in reward learning. They also highlight the need for unraveling individual differences in cognitive mechanisms that account for differences in reward learning, and obtaining reliable estimates for them.

Keywords: placebo; expectation; reinforcement learning; Q-learning; computational modeling

Introduction

Reinforcement learning (RL), the process in which humans or animals learn to make decisions in order to gain rewards, is thought to be of significance in the development of depression (Huys, Daw & Dayan, 2015) and particularly anhedonia as a core symptom of depression (Pizzagalli, 2014). RL is closely linked to dopamine (DA) activities (Dabney et al., 2020). Moreover, blunted DA signaling within reward-associated pathways has been recently shown to characterize depressive disorder (Belujon & Grace, 2017).

Placebo effects in the treatment of depression have been well documented (Petrie & Rief, 2019). Antidepressant placebo responses may accordingly be driven by positive expectations towards a successful treatment outcome, which contribute greatly to reducing depressive symptoms in clinical interventions. Further, based on the notion that placebo effects are triggered by the expectation of clinical benefit, i.e. expectation of reward, a tight link between the placebo effect and reward mechanisms has been highlighted in the context of other disorders such as Parkinson's disease (de la Fuente-Fernández, 2009).

RL is commonly assessed by means of behavioral or computational parameters, such as the count of collected rewards (Schmidt et al., 2014), or algorithms incorporating learning from prediction errors (Turi et al., 2017). The latter is often implemented by Q-learning as a model-free algorithm that slowly integrates trial-wise reward feedback in order to map a reward value on a number of states. It incorporates learning from prediction errors via learning rates, which reflect how strong a learner adjusts its reward expectations depending on new feedback. Q-learning has been postulated to serve as an efficient ground for capturing ganglio-basal reinforcement learning processes by means of latent parameters (Frank et al., 2007). A task that has widely been used to assess reinforcement learning this way is the Frank task (Frank, Seeberger & O'Reilly, 2004), in which participants have to learn from probabilistic feedback to identify the most rewarding stimulus within different stimuli pairs. With respect to Q-learning, this task conventionally involves the generation and updating of one reward value per stimulus. However, participants might only learn to distinguish between "good" and "bad" stimuli irrespective of individual reward probabilities, or misleadingly assume multiple reward values per stimulus depending on prior trial features. Differences in such strategies would result in a varying number of learned representations leading to more nuanced RL parameter estimates and thus, we were interested in how different assumptions regarding reward value generation could contribute to model participants' RL behavior more accurately.

Taken together, previous findings indicate that positive expectations may enhance reward-based decision making, and the Frank task constitutes an approved task for elucidating this relationship. The goal of our study was

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* ©2022 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

therefore to examine if differences in induced expectations covary with participants' ability to learn from reward in the Frank task. More specifically, we expected that RL learning rates would be enhanced in the experimental group which received antidepressant expectation, and explored to what extent differences in the learning rate for gain (α_{gain}) and state representations could contribute to a better explanation of participants' decision behavior.

Methods

Sample and Design

For this study, we re-analyzed data of 55 (7 male) healthy, non-depressed university students, who participated for course credits. The mean age was 21.1 years (*SD* = 2.4). Of the original 56 participants, one participant was excluded from analysis due to missing task data. The study was approved by the Local Ethics Committee. Data and analysis scripts were made publicly available[1].

In a psychophysiological experiment, participants were told that they took part in an open-label treatment study, in which they would receive either an antidepressant drug (50mg sulpiride; positive expectation), or an inactive substance (placebo; neutral expectation). In fact, all participants received a placebo, but the expectancy group allocation was randomized. As an incentive, the three best performing participants received monetary rewards. The participants finished a Frank task, and subsequently underwent a mood induction procedure, which is not part of the current study.

Paradigm

We analyzed the training phase of a Frank task (cf. Frank, Woroch & Curran, 2005) which we had adapted to electrocardiography by extending the inter trial interval. Three letter pairs (Japanese hiragana characters) with different reward probabilities (0.8:0.2, 0.7:0.3, and 0.6:0.4) were used as stimulus material. In each trial, participants had to select either of the letters from a pair by pressing one of two buttons. Stimuli pairs were presented until button press (but max. 2000ms), subsequently followed by a black screen and win or loss feedback screen (green circle or red cross) for 1000ms each. Afterwards, a resting screen with a fixation cross appeared for 4000-5000ms. Depending on individual task performance, which was determined by reaching a criterion of correct decisions made per stimulus pair (at least 65%, 60% and 50% for the pairs with the highest to the lowest difference of reward probabilities, respectively), the task comprised two to four reinforcement learning blocks with 20 location-balanced repetitions of all stimuli pairs resulting in 60 trials per block and at least 120 trials per participant as in the study conducted by Frank et al. (2005). After reaching this performance criterion, participants continued with the test phase, which required a sufficient knowledge of reward probabilities and was not analyzed in the present study. Within the first six trials, each pair was presented two times and the reward probabilities were rounded to one and zero.

Task instructions were presented on a screen. Participants were informed that they would be presented stimuli pairs and that they would have to identify the more rewarding stimulus of each pair by trial-and-error in order to maximize their reward (points). The instructions stated that the participant should keep choosing the more rewarding stimulus, although it could be punished at times by losing points in case of negative feedback. At this time, the participants were not aware of the different proportions of reward contingencies between different stimuli pairs.

Q-learning

We explored a series of Q-learning models differing with regard to the number of Q-states. The aim was to investigate if modeling an inappropriate task-structure would explain individual differences in predictive Q-learning model fits. In order to assess the relationship between the number of Q-states, model performance and estimated RL parameters, we tested simplified models against a task-appropriate standard model in the first place. These simpler models allowed for an aggregate representation of reward within both more and less rewarding stimuli. After observing poor average fits for the simplified compared to the standard model, the aim was to assess if participants might, on the contrary, have erroneously learned a task structure that was overly complex by taking prior extrinsic/intrinsic trial outcomes into account. As such, preceding reward and trial accuracy a, which is equal to one if the more probably rewarded stimulus of a pair was chosen in a particular trial (and equal to zero otherwise), were used to index Q-values differently resulting in two up to four reward representations per stimulus. Lastly, the models also varied with regard to the involvement of a group effect on the RL parameters as well as concerning the number of Q-values used during RL.

E6 and N6 Our standard Q-learning models E6 and N6 included six Q-states and differed with regard to the inclusion of a group effect. Here, "E" (for "effect") refers to the model with a group effect and "N" to no group effect. Models were fitted by sampling. For standard Q-value updating, we used a dual learning rate model (see eqn. 1) with $Q_i \in [0,1]$, $i \in \{1,2,\ldots,6\}$, $r \in \{0,1\}$ and $\alpha \in [0,1]$.Here, the learning rate parameter α was computed separately for reward (gain, α_G) and loss (α_L) based on the reward prediction error $r(t) - Q_i(t)$, which is the difference between received and expected reward. Decision probability for selecting stimulus A over stimulus B was computed using a softmax function (see eqn. 2), where $P_A(t) \in [0,1]$ and $\beta \geq 0$. Q_A and Q_B denote the Q-values of either stimulus in trial t. Decision noise is reflected in β, i.e. large (necessarily positive) values for β imply strong decision noise.

[1] http://dx.doi.org/10.17192/fdr/89

$$Q_i(t+1) = Q_i(t) + \alpha_G[r(t) - Q_i(t)]_+ + \alpha_L[r(t) - Q_i(t)]_- \quad (1)$$

$$P_A(t) = \frac{e^{\frac{Q_A(t)}{\beta}}}{e^{\frac{Q_A(t)}{\beta}} + e^{\frac{Q_B(t)}{\beta}}} \quad (2)$$

RL parameters θ, i.e. α (gain and loss) and β, of participant j were computed with the constraint of a normal distributed group-level parameter μ_θ and group-independent parameter variance σ_θ. If a group-effect ("E") model was fitted, μ_θ was the group-level parameter of the placebo condition, and a group effect term $\delta_\theta * Sulpiride_i$ was added with $Sulpiride_i \in \{0,1\}$:

$$logit(\alpha_j) \sim Normal(\mu_\alpha + \delta_\alpha * Sulpiride_j, \sigma_\alpha)$$

$$log(\beta_j) \sim Normal(\mu_\beta + \delta_\beta * Sulpiride_j, \sigma_\beta)$$

$$\mu_\theta \sim Normal(0{,}100)$$

$$\sigma_\theta \sim Uniform(0{,}100)$$

$$\delta_\theta \sim Normal(0{,}1)$$

Initially, models E6 and N6 were computed. In a second step, we tried to reduce model complexity by reducing the number of Q-states allowed for parameter estimation with the goal to analyze if a simpler model would perform equally in this task.

E2 and N2 We defined a simplified model, which consisted of only two Q-states for more (reward probability 0.8 to 0.6) and less frequently (reward probability 0.4 to 0.2) rewarded stimuli, respectively. In other words, this model assumes that the learner assigns the same Q-value to all stimuli that are expected to yield more (or less) reward than the alternative stimulus in a given trial. Since we do not have direct access to the learner's beliefs, we use trial accuracy as a proxy in the model. Thus, trial accuracy *a* indexed the Q-state i:

$$Q_i(t): i = i(a), a \in \{0,1\} \Rightarrow i \in \{1,2\}$$

E6B and E6B1 At the level of six concomitant Q-states, we also asked if participants might stick to their initial selections by introducing a moderate (E6B) to strong (E6B1) initial bias rewarding the first selection of each pair p:

$$Q_i(t_p = 1): Q_i = 0.5, p \in \{1,2,3\}$$

$$Q_i(t_p = 1): Q_i = 1, p \in \{1,2,3\}$$

Thus, when the three stimuli pairs were presented for the first time, the first Q-value of the selected stimuli was fixed to $Q_i = 0.5$ in the E6B model, and to $Q_i = 1$ in the E6B1 model.

E12S, E12R and E24 Furthermore, we wanted to investigate whether participants had adopted an overly complex representation of the task structure by indexing Q-values by the outcome of the previous trial in addition to the stimulus identity. Such a model would ultimately result in a poor predictive match of the E6/N6 model, since the latter could not capture reward history information in the Q-states. Hence, we designed models with two (E12) or four (E24) possible Q-states per stimulus as a consequence of accuracy and reward of the previous trial. In the E12 models, accuracy a (E12S) or reward r (E12R) indexed the Q-state i:

$$Q_i(t): i = i(p_t, a_{t-1}) \Rightarrow i \in \{1,2,\ldots,12\}$$

$$Q_i(t): i = i(p_t, a_{t-1}) \Rightarrow i \in \{1,2,\ldots,12\}$$

In the E24 model, accuracy and reward gated the trial-wise Q-value updating simultaneously resulting in 24 Q-states:

$$Q_i(t): i = i(p_t, a_{t-1}, r_{t-1}) \Rightarrow i \in \{1,2,\ldots,24\}$$

Since the decision likelihood of trial *t* depended on a valid trial *t-1*, initial trials without preceding trials were skipped for parameter estimation and simulation of models with more than six Q-states.

Parameter Estimation and Predictive Match

As a measure for model performance, we used the predictive match of stimulus choices on the data set used for fitting, i.e. the match between participants' and models' selection in all individual trials. We define the *predictive match* as a measure of agreement between model and participant regarding trial accuracy, i.e. the percentage of trials per participant, in which model and participant selected the same stimulus, and therefore achieved the same accuracy within a trial. We fitted all models in parallel using the RStan (Stan Development Team, 2021) package in R (version 4.1.2; R Core Team, 2021) on an AMD Threadripper 2990WX. As a likelihood function for model fitting, we used the model decision probability (see eqn. 2) evaluated at participants' actual choices. Markov chain Monte Carlo sampling was performed using four chains with 6000 warm-up iterations and additional 6000 iterations of sampling. The target average acceptance probability was set to .99 and the maximum tree depth was restricted to 15 units. Analysis scripts were provided by Turi et al. (2017)[2] and adapted to the purpose of our study.

Single-trial decisions were predicted through re-running Q-learning with the individual posterior means of the sampled RL parameters $\overline{\theta_j}$ on individual trial-by-trial data used for model fitting. Since the model decisions rely on random uniform sampling (0 or 1) in case of equal decision likelihoods between two stimuli of the same pair, we iterated the predictive match computation 25-times and took the mean as a predictive estimate.

[2] https://github.com/ihrke/2016-placebo-tdcs-study

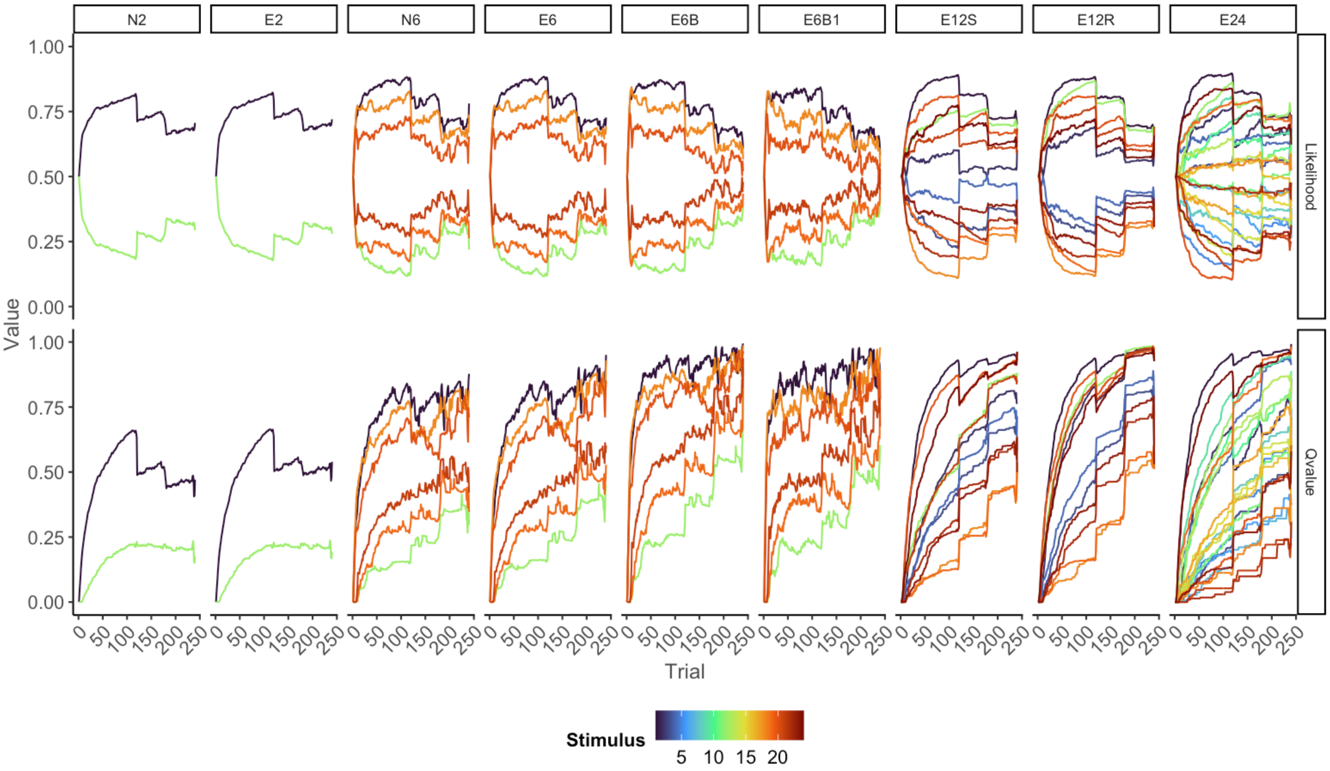

Figure 1: The learning curves illustrate the likelihood (top row) of selecting a stimulus (colored lines) as well as the expected value of reward (Q-value; bottom row) of each stimulus over the course of the training phase of the Frank task (x-axis). The higher the value (y-axis), the more likely a stimulus is chosen and the higher is the subjective reward probability. Stimulus means the indexed (discrete) Q-state. Note that in case of models with two Q-states, each graph represents the subjective reward probability of either "good" or "bad" stimulus category, whereas in the models with more than six Q-states, each stimulus consists of multiple subjective reward probabilities depending on previous intrinsic and/or extrinsic trial outcomes.

Statistical Analysis and Results

For statistical analysis, R (version 4.1.0; R Core Team, 2021) was used. All reported confidence intervals (CI) represent the 95%-CI.

Predictive Match

A hierarchical linear mixed-effect regression was performed to capture the variance of predictive match explained by model differences while accounting for individual differences. The predictive match was regressed onto model type as fixed effect and participants as random effect. The mixed-effect model was compared against an intercept-only random effects model using the Likelihood Ratio Test (LRT). For this purpose, the mixed models were estimated using maximum likelihood. Model E6 served as reference. The LRT revealed a significant contribution of model type to the model fit ($\chi^2(8) = 66.24$, $p < .001$). On average, all models comprising more or less than six Q-states performed worse than the E6 model. Notably, 94.5% of the variance was accounted for by individual differences, whereas the model type explained 0.7% of the variance in predictive match.

Learning curves and predictive matches for all models are depicted in Fig.1 and Fig.2 (top row), respectively. In our sample, average predictive matches of the top three models E6 (β_0 = 75.6, CI = [71.86, 79.30]), N6 (β = -0.05, CI = [-1.19, 1.08]) and E6B (β = -0.07, CI = [-1.21, 1.07]) were nearly identical. E6B1 (β = -0.95, CI = [-2.09, 0.18]) and E12S (β = -1.65, CI = [-2.78, -0.51]) achieved a match of 74.6% and 73.9%, respectively. N2 (β = -1.91, CI = [-3.04, -0.77]) and E2 (β = -1.91, CI = [-3.05, -0.78]) performed equal (73.7%). The worst match was obtained for E24 (72.7%) and E12R (72.4%). Given that the predictive matches exhibit large variability, we took a closer look at the best performing model for each participant. Although deviations from the standard model resulted in lower mean predictive matches, both biased models (E6B and E6B1), the more complex E12S model as well as the simplified N2 model showed a higher or least same number of cases in which each model performed the best compared to all other models (Fig.2; middle row). The E2 model showed the highest mean predictive match in the cases where it outperformed all other models for a participant (Fig.2; bottom row). Furthermore, these cases were relatively frequent, see middle row of Fig.2.

We further asked to what extent the individuals' best explaining models could contribute to an overall improvement in predictive accuracy. When averaging over the highest predictive matches per participant, the E6 model

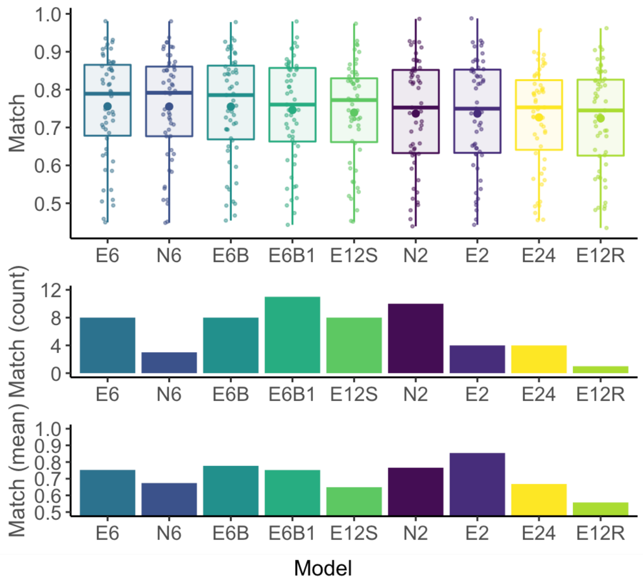

Figure 2: Predictive model performance. Mean predictive matches per participant (horizontally jittered dots) and model (bold dots) are illustrated in the top row. The middle row shows the number of cases (i.e. the count) in which a particular model outperformed all other models. The bottom row depicts the mean predictive match per model in these cases.

could be outperformed by 1.9% with a mean predictive match of 77.5%. Notably, when contrasting the decision outcomes of models and participants in terms of collected reward, the models consistently outperformed participants irrespective of the adjustments that were made to the Q-learning formula. Descriptively, participants were rewarded in 58.5% of their performed trials, whereas on average, model predictions on the same data achieved reward in 64% (E24) up to 69% (N2) of the trials, therefore signaling overestimation of the predicted reward.

Learning Rate

A linear mixed-effect regression with model type and expectancy group as interacting fixed effects and participants as random effect was performed to assess how the size of the expectancy effect regarding α_{gain} differs across model types. Group and model type were referenced to the placebo condition and E6 model, respectively. Residual maximum likelihood estimation was applied to the mixed model.

The results are illustrated in Tab.1. Density curves for model-dependent expectancy effects on α_{gain} are shown in Fig.3 (top). A total of 10.1% of the overall variance in α_{gain} can be explained by individual differences, whereas 67.8% is explained by the fixed effects. Overlap of the 95% confidence intervals of the marginal means for the model types with significant group interaction terms (E2: [-3.58, -2.93] and [-3.48, -2.76]; E12S: [-1.66, -1.01] and [-1.18, -0.46]; E24: [-1.20, -0.55] and [-0.79, -0.08]) indicate no expectancy effects on α_{gain} in these model types, whereas the other model types are predicted to show the same strength in group effect according to the regression results depicted in Tab.2. Regarding the size of α_{gain}, the coefficients involving both models with initial bias (E6B and E6B1) translate to a numerically higher overall α_{gain}, whereas, in contrast, the simplified model (E2) features a numerically lower overall α_{gain} as depicted in Fig.3 (bottom).

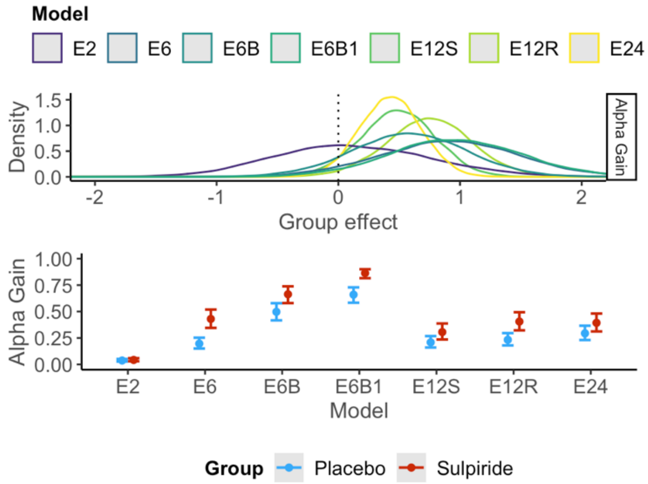

Figure 3: Distributions of sampled group effects per model on logit-scale for α_{gain} in the top figure. The bottom figure shows the regression-predicted α_{gain} transformed to original scale per model type and expectancy group.

Remarkably, the three most rewarded participants were best described by models with only two Q-states. The individual reward count relative to the trial count was the highest in participants performing best under the E2 model (63.3%), followed by E6 (60.2%), E6B (60.0%), E6B1 (58.0%), E24 (55.8%), E12S (55.8%), E12R (47.3%).

Discussion

Our aim was to test for antidepressant expectancy effects on a computationally obtained estimate for reward learning and to explore different Q-learning model structures in order to improve predictive model performance. For this purpose, participants were allocated to two groups that differed regarding verbal instructions on treatment efficacy (open-label placebo vs. sham antidepressant). We observed a robust expectancy effect on the α_{gain} mainly across models involving six Q-states, and a rather small account of the conducted model adjustments in explaining variance of the predictive match. Given that E6 is the model with the highest predictive match on average, and given that this model is the best match to the task structure, the average participant behaved like an ideal learner, and the presence of a positive antidepressant treatment expectation was observed. Even though the average participant was ideal in terms of learning

Table 1: Results of linear mixed-effect regression for logit-space α_{gain}. Kenward-Roger approximation was used for p-value computation.

	$\alpha_{gain}(logit)$		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-1.41	-1.74 – -1.08	**<0.001**
E2	-1.84	-2.23 – -1.46	**<0.001**
E6B	1.40	1.02 – 1.78	**<0.001**
E6B1	2.07	1.69 – 2.45	**<0.001**
E12S	0.08	-0.30 – 0.46	0.692
E12R	0.22	-0.16 – 0.60	0.263
E24	0.53	0.15 – 0.91	**0.006**
Sulpiride	1.13	0.64 – 1.61	**<0.001**
E2 * Sulpiride	-0.99	-1.56 – -0.43	**0.001**
E6B * Sulpiride	-0.44	-1.00 – 0.13	0.129
E6B1 * Sulpiride	0.05	-0.52 – 0.62	0.860
E12S * Sulpiride	-0.61	-1.18 – -0.05	**0.034**
E12R * Sulpiride	-0.32	-0.89 – 0.25	0.268
E24 * Sulpiride	-0.69	-1.25 – -0.12	**0.018**
Random Effects			
σ^2	0.57		
$\tau_{00\ subj}$	0.26		
ICC	0.31		
N $_{subj}$	55		
Observations	385		
Marginal R^2 / Conditional R^2	0.678 / 0.779		

the task structure, there was a notable fraction of participants that gained more reward, and were better described by a model that was much simpler than the correct model for this task. We observed the lowest marginal means for α_{gain} in the simplified model (E2) and the highest in the biased models (E6B/E6B1). Expectancy effects were only observed in models with a sufficient amount of α_{gain}, indicating that expectancy manipulation only worked in those participants showing at least some RL capabilities. To our surprise, participants performing under the E2 model received the most rewards relative to their counts of performed trials. This indicates a low α_{gain} that is constant throughout the task to be highly adaptive to the Frank task. Contrary to the simple learners, high learning rates in participants with biased initial decisions may point at the necessity of quickly relearning reward values after realizing that the initial reward expectations were inappropriate. Models with time-based RL parameters considering contextual changes and exploring further unintended task learning structures would advance the understanding (cf. Eckstein, Wilbrecht & Collins, 2021).

In the same regard, as the predictive match varied strongly between participants, the appropriateness of plain vanilla Q-learning as a description of observable behavior clearly differed between individuals. This raises the need for integrating different sources of reward value generation into reinforcement learning approaches for cognitive modeling. Some of the performed model adjustments could map such relationships as for example the biased models. Decisions within a probabilistic RL task may be biased by task-independent weights, as for instance task instructions (Doll et al., 2009). In our study, participants were advised to stay with the more rewarding stimulus, although it may have not been rewarded at times. Therefore, in a reasonable number of participants, such an instruction could have been represented as initial Q-value bias. Future approaches could aim at parametrizing the initial bias probabilistically such that instead of working with fixed values as in our study, the bias would be estimated using sampling methods.

The treatment effect observed in the current study is in line with previous findings of enhanced α_{gain} via expectancy-driven placebo intervention (Turi et al., 2017). Analogous to this, we used verbal instructions to modulate expectations, but without controlling for baseline performance. More positive expectations are also thought to enhance DA-mediated RL (de la Fuente-Fernández, 2009). Taken together with the proposed blunted RL in anhedonia, treatments with the aim of facilitating anhedonia may want to consider enhancing striatal DA activity by the use of contextual factors similar to expectancy enhancement in order to trigger striatal DA activity and thereby improving RL as the potential origin of symptom development. This could comprise instructions, whereas the direct neural target of such instructions remains unclear. A possible mechanism might rely on the induction of uncertainty (Tobler, Fiorillo & Schultz, 2005), which is thought to evoke tonic DA activation and was proposed to facilitate learning (Monosov, 2020).

Taken together, the present results shed light on converging evidence for the strength of expectation effects on reward learning, and suggest that the role of DA in RL mechanisms via induction of antidepressant expectation holds additional potential for improving a part of depressive conditions. The link between expectancy effect and reward processing could especially be of importance for a better understanding and development of clinical interventions. Furthermore, we conclude that a large proportion of interindividual variability in model-free reward learning could not be accounted for by restricting or extending the number of reward representations allowed per stimulus, which shifts the focus to other promising sources of reward value generation.

Acknowledgements

This work was supported by the DFG SFB-TRR 289 „Treatment Expectation“, project A07 (project number 422744262), and “The Adaptive Mind”, funded by the Excellence Program of the Hessian Ministry for Science and the Arts.

References

Belujon, P., & Grace, A. A. (2017). Dopamine System Dysregulation in Major Depressive Disorders. *International Journal of Neuropsychopharmacology*, *20*(12), 1036–1046. https://doi.org/10.1093/ijnp/pyx056

Cooper, J. A., Arulpragasam, A. R., & Treadway, M. T. (2018). Anhedonia in depression: biological mechanisms and computational models. *Current Opinion in Behavioral Sciences*, *22*, 128–135. https://doi.org/10.1016/j.cobeha.2018.01.024

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

de la Fuente-Fernández, R. (2009). The placebo-reward hypothesis: dopamine and the placebo effect. *Parkinsonism & Related Disorders*, *15*, S72–S74. https://doi.org/10.1016/S1353-8020(09)70785-0

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. https://doi.org/10.1016/j.brainres.2009.07.007

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104

Frank, M. J., Seeberger, L. C., & O’Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, *47*(4), 495–501. https://doi.org/10.1016/j.neuron.2005.06.020

Huys, Q. J. M., Daw, N. D., & Dayan, P. (2015). Depression: A Decision-Theoretic Analysis. *Annual Review of Neuroscience*, *38*(1), 1–23. https://doi.org/10.1146/annurev-neuro-071714-033928

Monosov, I. E. (2020). How Outcome Uncertainty Mediates Attention, Learning, and Decision-Making. *Trends in Neurosciences*, *43*(10), 795–809. https://doi.org/10.1016/j.tins.2020.06.009

Petrie, K. J., & Rief, W. (2019). Psychobiological Mechanisms of Placebo and Nocebo Effects: Pathways to Improve Treatments and Reduce Side Effects. *Annual Review of Psychology*, *70*(1), 599–625. https://doi.org/10.1146/annurev-psych-010418-102907

Pizzagalli, D. A. (2014). Depression, Stress, and Anhedonia: Toward a Synthesis and Integrated Model. *Annual Review of Clinical Psychology*, *10*(1), 393–423. https://doi.org/10.1146/annurev-clinpsy-050212-185606

R Core Team. (2021). *R: A language and environment for statistical computing* (4.1.0). R Foundation for Statistical Computing. https://www.r-project.org/

Schmidt, L., Braun, E. K., Wager, T. D., & Shohamy, D. (2014). Mind matters: placebo enhances reward learning in Parkinson’s disease. *Nature Neuroscience*, *17*(12), 1793–1797. https://doi.org/10.1038/nn.3842

Stan Development Team. (2021). *RStan: the R interface to Stan* (R package version 2.21.3).

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, *307*(5715), 1642–1645. https://doi.org/10.1126/science.1105370

Turi, Z., Mittner, M., Paulus, W., & Antal, A. (2017). Placebo Intervention Enhances Reward Learning in Healthy Individuals. *Scientific Reports*, *7*(1), 41028. https://doi.org/10.1038/srep41028