Supplementary Online Content

Peciña M, Chen J, Karp JF, Dombrovski AY. Dynamic feedback between antidepressant placebo expectancies and mood. *JAMA Psychiatry*. Published online March 1, 2023. doi:10.1001/jamapsychiatry.2023.0010

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods

© 2023 American Medical Association. All rights reserved.

1. 1. Inclusion Criteria

Enrolled participants were unmedicated, right-handed, fluent in English, and provided written and signed informed consent, approved by the University of Pittsburgh Institutional Review Board. All research subjects had a diagnosis of non-psychotic MDD with or without anxiety disorders using the Mini-International Neuropsychiatric Interview (M.I.N.I) [1]. Participants had at least moderate depression, as determined by a Hamilton Depression Rating Scale (HDRS-17) [2] score of ≥ 16 at screening. All participants were antidepressant medication-free for at least 21 days prior to the collection of imaging data (five weeks for fluoxetine). Only one participant had received an antidepressant during the current episode. Patients had a diagnosis of Major Depressive Episode (21% first episode, 25% recurrent, 54% chronic), with (65%) or without suicidal thoughts, panic disorder (11%), agoraphobia (35%), social anxiety (21%) and generalized anxiety disorder (45%).

1. 2. Exclusion Criteria

Exclusion criteria included pregnancy or breastfeeding; history of psychotic depression, bipolar disorder, schizophrenia or other psychotic-spectrum disorder; meeting M.I.N.I. criteria for substance dependence in the last six months except for nicotine, or substance abuse in the last two months; requiring immediate hospitalization for a psychiatric disorder or an unstable general medical condition; actively suicidal or considered high suicide risk, or having any contraindication for having an MRI.

In addition to excluding individuals with current/recent use of opioids or opioid receptor modulating drugs or other substances/alcohol, participants were excluded if they tested positive to a 12-panel drug test (Alere iCup Dx Drugs of Abuse Test Cup by Alere Toxicology Services – Products Division, Portsmouth, VA 23704) that included the following drugs: Cocaine, THC, Methamphetamine, Opiates, Methadone, Tricyclic Antidepressants, Oxycodone, MDMA, PCP,

© 2023 American Medical Association. All rights reserved.

Amphetamine, Barbiturates, Benzodiazepines. Four participants tested positive for THC but remained in the study after agreeing to remain absent for the duration of the study. Current alcohol abuse and dependence were evaluated using the M.I.N.I. alcohol abuse and dependence items. Subjects were excluded if they had current alcohol abuse and/or dependence. Current nicotine use was not an exclusion criterion. Only five participants (8%) were current smokers.

Results on a subset of these participants (n=20) have been published elsewhere [3]. This study included 20 psychotropic-free patients with MDD with the same inclusion and exclusion criteria, who completed a randomized, double-blind, placebo-controlled crossover study of 1 oral dose of 50 mg of naltrexone or matching placebo immediately before completing 2 sessions of the antidepressant placebo functional magnetic resonance imaging task. Only results from the placebo session were included in this study. Clinical, behavioral, and neuroimaging procedures across both studies were identical.

1. 3. Consent procedures

Participants were deceived about the purpose of the study. The deceptive narrative described an experimental manipulation aimed at investigating the brain effects of a “fast-acting antidepressant” compared to a “conventional antidepressant” while recording “participants’ brain activity” and providing neurofeedback. Participants did not know when they were given the “fast-acting” vs. the “conventional antidepressant”. Instead, they were supposed to differentiate the two of them based on the different levels of positive neurofeedback that followed the antidepressant infusions. Participants were told that, after each infusion, neurofeedback of positive signal would present with acute mood improvement, whereas baseline neurofeedback signal was unlikely to be followed by mood improvement. Participants were also informed that in addition to the infusion periods, there will be periods of “equipment calibration” – the study

© 2023 American Medical Association. All rights reserved.

control condition – where no drug would be administered, but neurofeedback signal will be recorded and displayed in the monitor. No additional information about the calibration periods was provided. Consequently, participants could have interpreted the meaning of the neurofeedback signal during calibration as evidence of spontaneous mood changes or previous drug infusions. Participants were asked to rate their expected and actual change in mood (YES/NO) in response to each infusion and neurofeedback signal, respectively, by using a keypad and their index fingers. The use of authorized deception – common in placebo research [4] – was clearly described in the consent form. Participants were assessed for their credibility of the experiment and debriefed about the deceiving procedures at the end of their participation.

1. 4. Assessment of the credibility of the experiment and debriefing procedures

The task's credibility was assessed in all participants at the end of the experiment by asking the following questions: "From 0 to 100% how often: (1) did the neurofeedback signal reflect your brain activity? (2) was an antidepressant treatment given to you during the infusion periods? and (3) was saline given to you during the "calibration" periods?" Questions 1 and 2 assessed the credibility of the infusions and neurofeedback signal. Subjects who responded 0 to questions 1 and 2 were excluded from the experiment. Question 3 was used to assess the credibility of the control condition and was not used to exclude participants from the experiment. No participants were excluded from our previous studies [3,5], or the current study itself based on the credibility assessment.

The debriefing session was completed immediately after the credibility assessment, where participants received an explanation of the hypothesis tested - the investigation of the neurobiological bases of antidepressant placebo effects. All deceiving procedures, and the reasons why it was necessary to deceive them. Specifically, participants were told that during the scanning session, they had received no antidepressant treatment, but rather saline, and that

© 2023 American Medical Association. All rights reserved.

the neurofeedback presented to them was sham. All participants reported positive reactions to the debriefing procedures.

1. 5. Analysis of Behavioral Data

We estimated multi-level logistic regression models predicting participants' evolving expectancies and mood ratings using R (version 4.1.2) [6], the R Studio *lme4* [7] package. These models estimated the fixed effects of two orthogonal experimental conditions [expectancy condition (antidepressant placebo infusion vs. calibration) and reinforcement condition (high vs. low reinforcement)], and their interaction. Subject intercepts were taken to be random. Significant predictors were identified using the likelihood ratio test (LRT; *car::Anova*[8]. In additional models, we evaluated moderating effects of subject-level variables, including regional BOLD activation. To rule out multicollinearity among predictors, we used the variance inflation factors (VIFs) function from the *car* package[8], to ascertain that all predictors and interactions of interest met a rigorous criterion of VIF < 3[8]. Interaction terms were plotted using the "effect" function in R Studio[9].

1. 6. Model comparison and selection

A basic Rescorla-Wagner reinforcement learning model and its variants modified to embody our hypotheses and alternative accounts were then compared against a null model that assumed no learning:

$$Q_{t+1}(s) = Q_t(s) \qquad \text{(eq. 3)}$$

Individual model log-evidence values were entered into a Bayesian model comparison (BMC) performed using the mbb-vb-toolbox (http://mbb-team.github.io/VBA-toolbox/) [10]. This Bayesian procedure estimates, among other criteria, the exceedance probability (EP) for each model within a set of models, given the data gathered from all participants. EP quantifies the belief that the model is more likely than all the other models of the set. An EP > 95% for one

© 2023 American Medical Association. All rights reserved.

model within a set is therefore typically considered as providing strong evidence in favor of this model being the most likely. This procedure uses an adjustment for the Bayesian omnibus risk (BOR), a measure of statistical risk in group model comparisons quantifying whether chance is likely to explain differences in estimated model frequencies .

Model name	Learning rule	Choice rule	EP
Basic Rescorla-Wagner learning	$Q_t = Q_{t-1} + \alpha (r - Q_{t-1})$		<10^-2
Placebo-biased learning	$Q_t = Q_{t-1} + \alpha_{plac/cal} (r - Q_{t-1})$		<10^-5
Feedback-biased learning	$Q_t = Q_{t-1} + \alpha_{pos/base\_nf} (r - Q_{t-1})$	$p_t = sig(Qt + K) * \beta$	<10^-5
Mood	$Q_t = Q_{t-1} + \alpha (u_t - Q_{t-1})$ where $u_t = r + mood_t$		0.02
Placebo-biased learning and mood	$Q_t = Q_{t-1} + \alpha_{plac/cal} (u_t - Q_{t-1})$ where $u_t = r + mood_t$		0.97
Null	$Q_t = Q_{t-1}$	$p_t = sig(K)$	<10^-5

eFigure 1. Antidepressant Placebo fMRI Task Reinforcement Learning Model-Based Behavioral Results: key equations and exceedance probabilities (EPs) for reinforcement learning models.

1. 7. MRI Data Acquisition

MR images were collected at the University of Pittsburgh Magnetic Resonance Imaging Center on a 32-channel head coil on a 3T Siemens PRISMA scanner. The MPRAGE sequence had repetition time (TR)=2400ms, echo time (TE)=2.22ms, flip angle (FA)=8deg, inversion time (TI)=1000ms, field of view (FOV)=300x320, 208 slices, 0.8mm isotropic (0.4mm space between slices), with GRAPPA factor of 2, and lasted ~6min 38 sec. An axial, whole brain echo planar (EPI) T2*-weighted functional images were collected to measure the blood-oxygen-level dependent (BOLD) response with TR=1000ms, TE=30ms, FA=45º, FOV=95x95, 60 slices, 2.3mm isotropic (no spaces), multiband factor of 5, and 2688 volumes (four ~11-minute runs). Participants were scanned for 90 minutes, and the fMRI task started at approximately minute 15.

© 2023 American Medical Association. All rights reserved.

Anatomical images first underwent gradient unwarping, and then were registered to the MNI152 template using both affine and nonlinear transformations methods implemented in FLIRT (FSL) and FNIRT (FSL), respectively. A mask of the brain was also created by removing the non-brain voxels from the anatomical images using BET (FSL) for later use in functional image co-registration.

The functional images underwent slice timing and motion correction simultaneously using NiPy's four-dimensional registration algorithm SpaceTimeRealign. Running both simultaneously ensured that motion artifacts would not be reintroduced into the data in later processing. Non-brain voxels in the images were removed by masking low-intensity voxels, calculated from the field map, and using brain extraction function BET (FSL). After intensity-normalizing every voxel to have a mean of 1000, wavelet despiking was performed using the BrainWavelet Toolbox with the spike threshold set to 10. The resulting image was aligned and warped to their anatomical images, resampled to 3mm isotropic voxels, and warped into MNI152 standard space. A 7mm full-width at half maximum kernel was used to smooth the images spatially and a high-pass filter was applied to remove signal slower than 0.008 Hz. Lastly, the images were intensity normalized by rescaling the intensity by 100 divided by voxel mean.

1. 8. MRI Data Acquisition, Preprocessing, Task-based and RL-based MRI Analysis

Acquired images were preprocessed using functions in the following software packages: NiPy[11,12], AFNI[13], BrainWavelet Toolbox[14] and the fMRI software library (FSL[15]). We have previously detailed this preprocessing pipeline elsewhere[16].

Subject-Level Analysis. The model included four event-regressors: infusion event, expectancy ratings event, neurofeedback event and mood ratings event. For the task-based

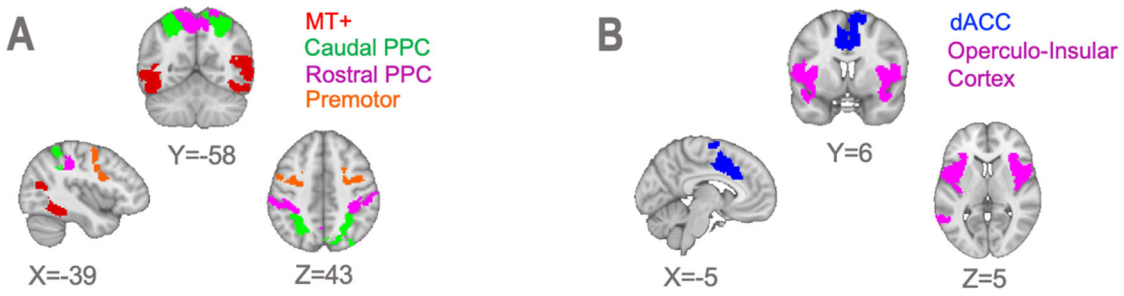
© 2023 American Medical Association. All rights reserved.

analysis we constructed two additional regressors for the expectancy manipulation, coded as 1 or -1 (“antidepressant” infusion cue or “calibration” no-infusion cue, respectively) and the reinforcement manipulation, coded as 1 or - 1 (positive or baseline sham neurofeedback, respectively). The expectancy regressor was aligned to the infusion event and the expectancy ratings event, whereas the reinforcement regressor was aligned to the neurofeedback event and the mood ratings event. For the RL voxel-wise model, we included parametric modulators: an expected value regressor aligned to the infusion and expectancy rating events and a PE regressor aligned to the neurofeedback and mood rating events. We convolved regressors with the HRF and estimated general linear models using FSL FEAT for each run and participant[17].

Group Analysis. We conducted group-level voxel-wise analyses using FSL *randomise* (one-sample t-test)[18] and Threshold Free Cluster Enhancement (TFCE) (1- P > 0.95)[19]. For all models, we convolved box-car regressors with the HRF and used general linear models using FSL FEAT for each run and participant.

The anatomical localization of cortical activation clusters was referenced against the Schaefer and colleagues’ 7-network, 400-node cortical parcellation[20]. Individual regression coefficients (“betas”) from *a priori* anatomical regions of interest (ROIs) in the dorsal attention network (DAN) and the salience network (SN), as defined by Schaefer’s atlas, were extracted for brain-to-behavior analyses. Schaefer and colleagues’[20] 7-network and 17-network parcellations reveal that the human DAN encompasses the temporo-occipital cortex (putative human MT+, abbreviated MT+ below for simplicity), the posterior parietal (caudal and rostral) and frontal premotor regions. The two main nodes of the SN were the anterior cingulate cortex and the operculo-insular cortex.

© 2023 American Medical Association. All rights reserved.



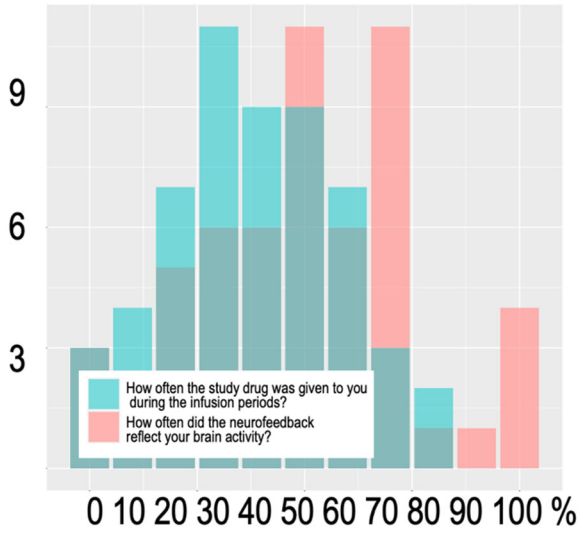
eFigure 2. Dorsal Attention Network (DAN) and Salience Network (SN) Regions of Interest (ROI)

Brain-behavior analysis. To understand the relevance of these brain responses to behavioral antidepressant placebo effects we examined the moderation effects of the regions of interest described above in the prediction of expectancy and mood ratings during the task. These analyses were performed to investigate which regional responses are required for task behavior (in this case the expectancy and learning components of the antidepressant placebo effect) and which responses are epiphenomenal. We expected the former, and not the latter, to scale with individual differences in behavior, providing additional correlational evidence for their role. While this evidence is not sufficient for demonstrating a causal role, we consider it necessary to rule out epiphenomenal brain responses. Importantly, these analyses are statistically independent of our preceding analyses of behavior and BOLD, because our model-free analyses of BOLD are agnostic of individuals' behavior and model-based analyses of BOLD use model-predicted signals generated at the group mean task parameters and not at the individual model parameters. In summary, while positive brain-behavior relationships are only suggestive and not sufficient to demonstrate a causal role of these networks, the lack of such relationships would have led us to view these regional responses as likely epiphenomenal.

© 2023 American Medical Association. All rights reserved.

eResults

2. 1. Credibility Assessments



Overall bar graphs of participants' credibility ratings are displayed in eFigure 2. Credibility reports were missing in 4 individuals.

eFigure 3. Credibility Questionnaire's Histograms

2. 2. Model-free behavioral analyses: manipulation effects

We estimated an additional mixed-effects model predicting mood ratings where instead of the expectancy condition we used individuals' trial-wise expectancy ratings. As expected, the fit of this model (II) was considerably better compared to the model based on expectancy manipulations (I): (Model with expectancy condition - AIC: 6400.9, Model with expectancy ratings - AIC: 6254.4, Observations = 6486, df = 6481), suggesting that momentary fluctuations in expectancy ratings predict mood ratings above and beyond the task condition effects.

2. 3. Model-free behavioral analyses: depression moderation effects

Our findings regarding the modulation effect of depression severity replicated previous findings in a smaller sample [1]. As described earlier, the effects of the expectancy manipulation on expectancy ratings were reduced in more severely depressed individuals, as reflected in a negative depression severity by expectancy condition interaction, when using both the HDRS and the MADRS. The effects of the reinforcement manipulation on mood ratings were also reduced in more severely depressed individuals, but only when using the MADRS and not the

© 2023 American Medical Association. All rights reserved.

HDRS. The self-reported QIDS-16 scale did not discriminate between those who would respond to the task conditions.

MODEL	A. Expectancy Ratings					B. Mood Ratings I				
	χ2	**Estimate**	**S.E.**	**z value**	**p**	**χ2**	**Estimate**	**S.E.**	**z value**	**p**
Expectancy Condition	139.67	1.01	0.08	11.80	**<0.001**	4.25	0.19	0.08	2.06	**0.03**
Reinforcement Condition	7.12	0.23	0.08	2.66	**0.01**	273.64	1.42	0.08	16.50	**<0.001**
Expectancy * Reinforcement	14.47	0.46	0.12	3.80	**<0.001**	4.23	0.27	0.12	2.06	**0.03**
Expectancy Rating						39.46	0.59	0.09	6.28	**<0.001**
Reinforcement Cond.						182.37	1.21	0.09	13.50	**<0.001**
Expectancy Rating * Reinforcement						17.60	0.55	0.13	4.19	**<0.001**
	χ2	**Estimate**	**S.E.**	**z value**	**p***	**χ2**	**Estimate**	**S.E.**	**z value**	**p***
HDRS	3.29	-0.47	0.26	-1.81	0.07	6.52	-0.60	0.23	-2.55	**<0.01**
Expectancy*HDRS	16.27	-0.38	0.09	-4.03	**<0.001**	0.41	-0.06	0.09	-0.64	0.51
Reinforcement *HDRS	2.16	-0.14	0.09	-1.47	0.14	0.67	-0.07	0.09	-0.82	0.41
Expectancy * Reinforcement	3.07	0.23	0.13	1.75	0.08	0.74	0.11	0.12	0.86	0.38
MADRS	3.45	-0.49	0.27	-1.85	0.06	3.49	-0.45	0.24	-1.87	0.06
Expectancy*MADRS	6.95	-0.24	0.09	-2.64	**<0.01**	0.68	-0.08	0.09	-0.82	0.41
Reinforcement *MADRS	0.13	0.03	0.09	0.35	0.72	6.18	-0.23	0.09	-2.49	**<0.01**
Expectancy * Reinforcement * MADRS	5.95	0.31	0.13	2.44	0.02	2.18	0.19	0.13	1.48	0.14
QIDS-16SR	5.51	-0.63	0.27	-2.35	0.02	7.21	-0.66	0.25	-2.68	**<0.01**
Expectancy*QIDS-16SR	0.16	0.04	0.09	0.40	0.69	1.08	0.09	0.09	1.04	0.30
Reinforcement *QIDS-16SR	1.60	0.12	0.09	1.26	0.20	2.36	0.14	0.09	1.54	0.12
Expectancy * Reinforcement * QIDS-16SR	0.36	0.08	0.13	0.60	0.55	0.14	0.05	0.13	0.38	0.71

*Adjusted p= 0.05/3scales = 0.016

eTable 1. Mixed-Effects Models for the Prediction of Expectancy and Mood Ratings and Their Modulation by Depression Severity

2. 4. Model-free DAN ROIs modulation of Expectancy and Mood Ratings.

To understand the relevance of these brain responses to antidepressant placebo effects, we entered mean coefficients from clusters responsive to expectancy and reinforcement cues into separate LME models predicting participants' ratings along with the task conditions. Greater BOLD responses in the MT+ and the premotor cortex during the expectancy condition were associated with a greater effect of the task conditions on expectancy ratings, as reflected by the

© 2023 American Medical Association. All rights reserved.

positive 2-way interaction. Greater BOLD responses in the MT+ and the rostral PPC during the high reinforcement condition were associated with higher mood ratings (see eTable 2).

DAN response to infusion cues	A. Expectancy Ratings					DAN response to reinforcement cues	B. Mood Ratings				
	χ2	Estimate	S.E.	z value	p*		χ2	Estimate	S.E.	z value	p*
MT+	11.08	-40.07	12.04	-3.32	**<0.001**	MT+	2.53	-18.24	11.47	-1.59	0.11
Expectancy*MT+	49.19	33.45	4.77	7.01	**<0.001**	Expectancy*MT+	1.81	7.96	5.91	1.35	0.178
Reinforcement*MT+	34.11	28.97	4.96	5.84	**<0.001**	Reinforcement*MT+	18.53	24.97	5.80	4.31	**0.001**
Expectancy*Reinforcement*MT+	28.48	-35.16	6.59	-5.34	**<0.001**	Expectancy*Reinforcement*MT+	3.52	-14.68	7.83	-1.88	0.06
PPC caudal	0.79	-13.52	15.24	-0.89	0.38	PPC caudal	0.72	-6.66	7.87	-0.85	0.40
Expectancy*PPC caudal	1.88	7.06	5.15	1.37	0.17	Expectancy*PPC caudal	3.76	8.99	4.64	1.94	0.05
Reinforcement*PPC caudal	6.01	12.93	5.27	2.45	0.01	Reinforcement*PPC caudal	6.57	11.30	4.41	2.56	**0.01**
Expectancy*Reinforcement*PPC caudal	0.00	-0.17	7.07	-0.03	0.98	Expectancy*Reinforcement*PPC caudal	1.62	-8.06	6.34	-1.27	0.20
PPC rostral	2.66	-12.22	7.49	-1.63	0.10	PPC rostral	0.68	-8.39	10.18	-0.83	0.41
Expectancy*PPC rostral	3.31	8.23	4.52	1.82	0.07	Expectancy*PPC rostral	1.30	5.14	4.50	1.14	0.254
Reinforcement*PPC rostral	1.23	-4.83	4.36	-1.11	0.27	Reinforcement*PPC rostral	8.08	12.90	4.54	2.84	**<0.001**
Expectancy*Reinforcement*PPC rostral	2.80	9.55	5.71	1.68	0.09	Expectancy*Reinforcement*PPC rostral	0.38	-3.41	5.51	-0.62	0.54
Premotor	9.90	-37.62	11.96	-3.15	**<0.001**	Premotor	1.09	-19.64	18.81	-1.05	0.30
Expectancy*Premotor	18.72	27.64	6.39	4.33	**<0.001**	Expectancy*Premotor	0.26	3.41	6.64	0.51	0.61
Reinforcement*Premotor	15.34	22.80	5.82	3.92	**<0.001**	Reinforcement*Premotor	1.59	8.19	6.50	1.26	0.21
Expectancy*Reinforcement*Premotor	0.42	-4.73	7.29	-0.65	0.52	Expectancy*Reinforcement*Premotor	0.61	-7.07	9.05	-0.78	0.43

*Adjusted p= $0.05/4_{regions} = 0.0125$

eTable 2. Mixed-Effects Models Examining Manipulation Effects on Expectancy and Mood Ratings and Their Moderation by Model-Free Neural Responses During the Antidepressant Placebo fMRI Task

2. 5. Model-free DAN ROIs correlations with depression severity

Correlation coefficients between expectancy (A) and reinforcement (B) DAN BOLD responses and depression severity were mostly negligible, with the exception of the correlation between MT+ and the QIDS-16SR (r=0.31, p=0.01), which did not reach significance after Bonferroni-correction (adjusted p= $0.05/12_{regions*scales} = 0.004$).

	Correlation Coef. (r)		
Prior expectancy brain responses	**HDRS**	**MADRS**	**QIDS-16SR**
MT+	0.19	0.19	0.31
PPC caudal	0.22	0.12	0.14
PPC rostral	0.17	0.08	0.01
Premotor	0.10	0.09	0.18
Reinforcement responses	**HDRS**	**MADRS**	**QIDS-16SR**
MT+	0.04	0.08	-0.14
PPC caudal	-0.14	-0.09	-0.21
PPC rostral	0.04	0.09	-0.14
Premotor	0.08	0.10	-0.06
*Adjusted p= $0.05/4_{regions}*3_{scales} = 0.0125$ - None survived correction.			

eTable 3. Correlation Between Prior Expectancy (A) and Reinforcement (B) DAN BOLD Responses and Depression Severity

2. 6. Model-predicted SN and DAN modulation of Expectancy and Mood Ratings.

© 2023 American Medical Association. All rights reserved.

SN responses to learned expectancies positively interacted with the high-reinforcement condition predicting expectancy ratings (dACC: χ^2= 5.94, p=0.015; eTable 4). DAN responses to prediction errors positively interacted with the task condition predicting expectancy ratings, as reflected by a 3-way interaction (MT+: χ^2= 6.86, p=0.002; PPC caudal: χ^2= 30.23, p<0.001, PPC rostral: 20.30, p<0.001; Premotor: χ^2= 13.96, p=0.001 eTable 4).

SN response to learned expectancies	A. Expectancy Ratings					DAN response to prediction error	B. Mood Ratings				
	χ2	Estimate	S.E.	z value	p*		χ2	Estimate	S.E.	z value	p*
Cingulate C.	2.40	-7.85	5.07	-1.55	0.122	MT+	3.19	-19.69	11.02	-1.79	0.074
Expectancy*Cingulate C.	2.91	3.40	2.00	1.71	0.088	Expectancy*MT+	7.37	-10.60	3.59	-2.95	**0.007**
Reinforcement*Cingulate C.	5.94	5.05	2.07	2.44	**0.015**	Reinforcement*MT+	1.11	3.86	3.15	1.22	0.292
Expectancy*Reinforcement*Cingulate C.	5.99	-6.70	2.77	-2.45	**0.014**	Expectancy*Reinforcement*MT+	6.86	12.86	4.91	2.62	**0.002**
Insula	0.02	-0.89	7.12	-0.12	0.901	PPC caudal	12.38	-20.95	5.96	-3.52	**<0.001**
Expectancy*Insula	0.41	-1.73	2.70	-0.64	0.523	Expectancy*PPC caudal	17.30	-13.39	3.22	-4.16	**<0.001**
Reinforcement*Insula	5.01	6.51	2.91	2.24	**0.025**	Reinforcement*PPC caudal	1.29	-3.31	2.92	-1.13	0.257
Expectancy*Reinforcement*Insula	4.55	-8.54	4.01	-2.37	0.033	Expectancy*Reinforcement*PPC caudal	30.23	21.75	3.96	5.50	**<0.001**
*Adjusted p= $0.05/2_{regions}$ = 0.025						PPC rostral	4.12	-15.14	7.46	-2.03	0.042
						Expectancy*PPC rostral	16.96	-15.26	3.47	-4.40	**<0.001**
						Reinforcement*PPC rostral	0.04	0.68	3.20	0.21	0.832
						Expectancy*Reinforcement*PPC rostral	20.30	20.33	4.29	4.74	**<0.001**
						Premotor	0.86	-6.02	6.50	-0.93	0.354
						Expectancy*Premotor	12.21	-12.87	3.68	-3.50	**<0.001**
						Reinforcement*Premotor	1.36	-4.28	3.67	-1.17	0.243
						Expectancy*Reinforcement*Premotor	13.96	17.75	4.75	3.73	**0.001**
						*Adjusted p= $0.05/4_{regions}$ = 0.012					

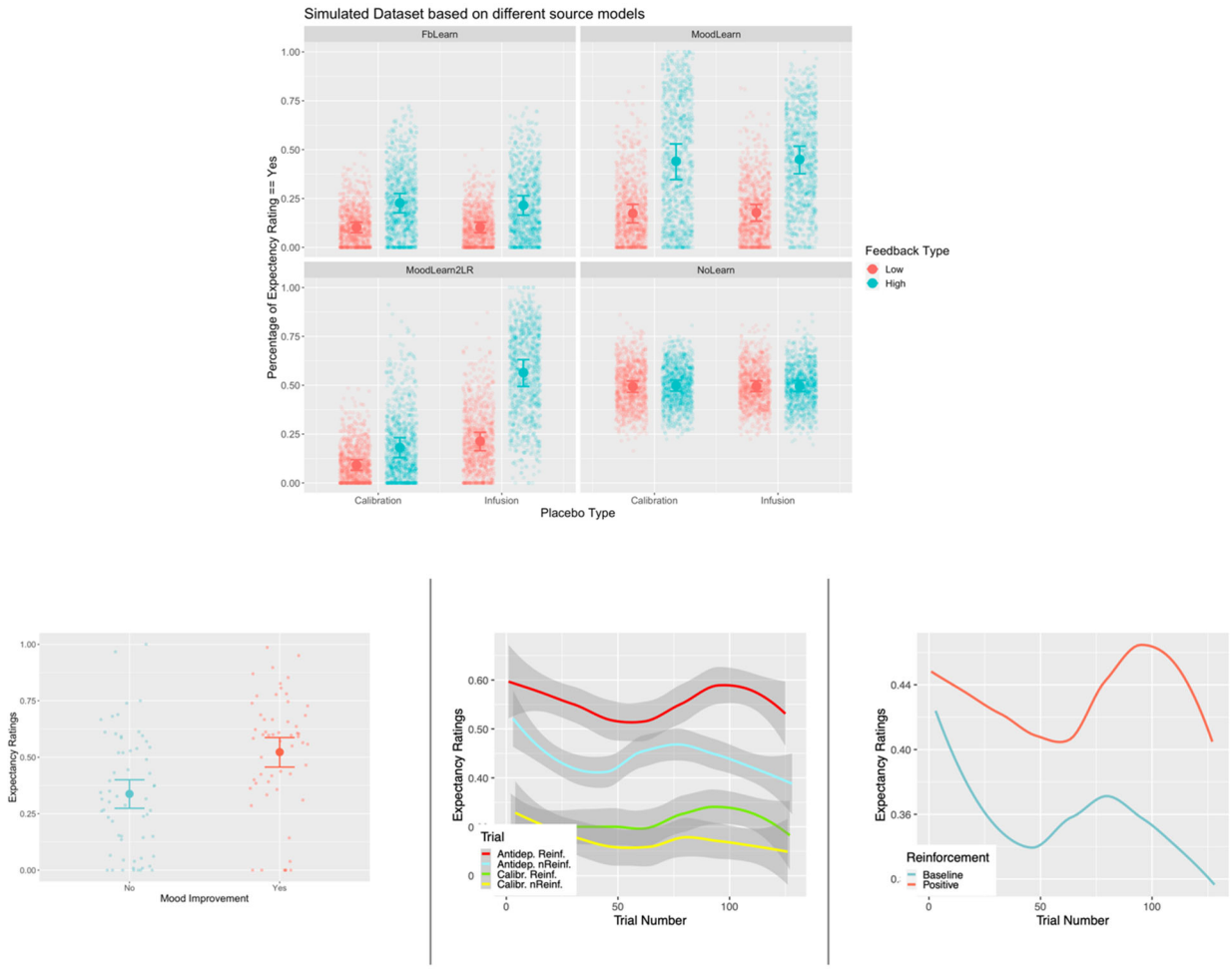
eTable 4. Mixed-Effects Models Examining Manipulation Effects on Expectancy and Mood Ratings and Their Moderation by RL-Based Neural Responses During the Antidepressant Placebo fMRI Task

2. 7. Qualitative model posterior checks vis-à-vis subjects' behavior; learning curves by condition

To examine whether alternative RL models qualitatively recapitulate subjects' behavior, we performed posterior predictive checks (eFigure 3, top). Only the model with biased learning for placebo cues and mood reinforcement, which dominated our model comparison, did so.

We also provide model-free tests of our model's predictions using participants' actual behavior: reinforcement of placebo expectancies by mood (eFigure 3, bottom left), relative ranking of stimuli throughout learning (eFigure 3, bottom center), and differential learning curves based on neurofeedback (eFigure 3, right).

© 2023 American Medical Association. All rights reserved.



eFigure 4. Top: Model-Predicted Expectancies Across 4 Key Models. Bottom Left: Our Mood Model Makes a Strong Prediction That if Mood Improved at a Preceding Presentation of the Same Stimulus (Trial t – k), the Expectancy is Heightened at the Current Presentation of This Stimulus k Trials Later at Trial t. Bottom Center: Average Expectancy Ratings by Stimulus Type (Learning Curves). Bottom Right: Average Expectancy Ratings by Reinforcement (Baseline vs Positive Neurofeedback)

2. 8. Model Recovery and confusion matrix

Our inferences about behavioral processes are based on model comparison. However, model comparisons can be misleading if models cannot be reliably identified. To address the question of identifiability (model recovery), we now report the confusion matrix (eTable 5) confirming high model uniqueness and lack of excessive model flexibility. Behavior produced by a given generative model was best explained by the same model (uniqueness) and our dominant model did not accommodate behavior produced by other generative models (lack of excessive flexibility

© 2023 American Medical Association. All rights reserved.

Recovery / Source	Null	Basic Learning	Mood	Mood Learning
Null	0.0000 | 0.0003	**0.5685 | 0.4347**	0.0000 | 0.1364	0.4315 | 0.4296
Basic Learning	0.0000 | 0.0003	**0.9760 | 0.3804**	0.0002 | 0.2925	0.0238 | 0.3279
Mood	0.0000 | 0.0003	0.0000 | 0.0003	**1.0000 | 0.7077**	0.0000 | 0.2929
Mood Learning	0.0000 | 0.0003	0.0000 | 0.0030	0.0000 | 0.3977	**1.0000 | 0.6028**

eTable 5. Simulation-Based Model Recovery: the Confusion Matrix

For each source (original) model, a simulated set of behavioral responses is generated (N=1000). In the next, recovery step, these responses are fit using the original model and key alternatives. Each cell shows how well the recovering model (columns) fits the source model (rows) using criteria produced by Bayesian model comparison: Exceedance Probability (EP) | estimated model frequency. Higher probabilities and frequencies indicate better fit. Values corresponding to the best fit for each source model are bolded.

© 2023 American Medical Association. All rights reserved.

eReferences

1. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of clinical psychiatry*. 1998;59 Suppl 20:22-33;quiz 34-57.

2. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.

3. Peciña M, Chen J, Lyew T, Karp JF, Dombrovski AY. μ Opioid Antagonist Naltrexone Partially Abolishes the Antidepressant Placebo Effect and Reduces Orbitofrontal Cortex Encoding of Reinforcement. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2021;6(10):1002-1012. doi:10.1016/j.bpsc.2021.02.009

4. Miller FG, Wendler D, Swartzman LC. Deception in research on the placebo effect. *PLoS Med*. 2005;2(9):e262. doi:10.1371/journal.pmed.0020262

5. Pecina M, Heffernan J, Wilson J, Zubieta JK, Dombrovski AY. Prefrontal expectancy and reinforcement-driven antidepressant placebo effects. *Transl Psychiatry*. 2018;8(1):222. doi:10.1038/s41398-018-0263-y

6. RStudio | Open source & professional software for data science teams. Accessed August 29, 2022. https://www.rstudio.com/

7. Bates D, Maechler M, Bolker [aut B, et al. lme4: Linear Mixed-Effects Models using “Eigen” and S4. Published online July 8, 2022. Accessed August 29, 2022. https://CRAN.R-project.org/package=lme4

8. Fox J, Weisberg S, Price B, et al. car: Companion to Applied Regression. Published online June 15, 2022. Accessed August 29, 2022. https://CRAN.R-project.org/package=car

9. Fox J. Effect Displays in *R* for Generalised Linear Models. *J Stat Soft*. 2003;8(15). doi:10.18637/jss.v008.i15

10. Daunizeau J, Adam V, Rigoux L. VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Comput Biol*. 2014;10(1):e1003441. doi:10.1371/journal.pcbi.1003441

11. Gorgolewski K, Burns CD, Madison C, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*. 2011;5. doi:10.3389/fninf.2011.00013

12. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies — Revisited. *NeuroImage*. 2014;84:971-985. doi:10.1016/j.neuroimage.2013.08.065

13. Cox RW. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*. 1996;29(3):162-173. doi:10.1006/cbmr.1996.0014

© 2023 American Medical Association. All rights reserved.

14. Patel AX, Kundu P, Rubinov M, et al. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *Neuroimage*. 2014;95:287-304. doi:10.1016/j.neuroimage.2014.03.012

15. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *NeuroImage*. 2012;62(2):782-790. doi:10.1016/j.neuroimage.2011.09.015

16. Vanyukov PM, Hallquist MN, Delgado M, Szanto K, Dombrovski AY. Neurocomputational mechanisms of adaptive learning in social exchanges. *Cogn Affect Behav Neurosci*. 2019;19(4):985-997. doi:10.3758/s13415-019-00697-0

17. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*. 2001;14(6):1370-1386. doi:10.1006/nimg.2001.0931

18. Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. *Neuroimage*. 2014;92:381-397. doi:10.1016/j.neuroimage.2014.01.060

19. Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*. 2009;44(1):83-98. doi:10.1016/j.neuroimage.2008.03.061

20. Schaefer A, Kong R, Gordon EM, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex*. 2018;28(9):3095-3114. doi:10.1093/cercor/bhx179

© 2023 American Medical Association. All rights reserved.