Behavioral/Cognitive

Irrelevant Threats Linger and Affect Behavior in High Anxiety

Kristoffer C. Aberg, Ido Toren, and Rony Paz
Department of Brain Sciences, Weizmann Institute of Science, Rehovot, 76100, Israel

Threat-related information attracts attention and disrupts ongoing behavior, and particularly so for more anxious individuals. Yet, it is unknown how and to what extent threat-related information leave lingering influences on behavior (e.g., by impeding ongoing learning processes). Here, human male and female participants ($N = 47$) performed probabilistic reinforcement learning tasks where irrelevant distracting faces (neutral, happy, or fearful) were presented together with relevant monetary feedback. Behavioral modeling was combined with fMRI data ($N = 27$) to explore the neurocomputational bases of learning relevant and irrelevant information. In two separate studies, individuals with high trait anxiety showed increased avoidance of objects previously paired with the combination of neutral monetary feedback and fearful faces (but not neutral or happy faces). Behavioral modeling revealed that high anxiety increased the integration of fearful faces during feedback learning, and fMRI results (regarded as provisional, because of a relatively small sample size) further showed that variance in the prediction error signal, uniquely accounted for by fearful faces, correlated more strongly with activity in the right DLPFC for more anxious individuals. Behavioral and neuronal dissociations indicated that the threat-related distractors did not simply disrupt learning processes. By showing that irrelevant threats exert long-lasting influences on behavior, our results extend previous research that separately showed that anxiety increases learning from aversive feedbacks and distractibility by threat-related information. Our behavioral results, combined with the proposed neurocomputational mechanism, may help explain how increased exposure to irrelevant affective information contributes to the acquisition of maladaptive behaviors in more anxious individuals.

Key words: **anxiety; distractors; dorsolateral PFC; maladaptive; prediction error; reinforcement learning**

Significance Statement

In modern-day society, people are increasingly exposed to various types of irrelevant information (e.g., intruding social media announcements). Yet, the neurocomputational mechanisms influenced by irrelevant information during learning, and their interactions with increasingly distracted personality types are largely unknown. Using a reinforcement learning task, where relevant feedback is presented together with irrelevant distractors (emotional faces), we reveal an interaction between irrelevant threat-related information (fearful faces) and interindividual anxiety levels. fMRI shows provisional evidence for an interaction between anxiety levels and the coupling between activity in the DLPFC and learning signals specifically elicited by fearful faces. Our study reveals how irrelevant threat-related information may become entrenched in the anxious psyche and contribute to long-lasting abnormal behaviors.

Introduction

In modern-day society, people are increasingly exposed to emotionally loaded information that is irrelevant for ongoing and prospective behaviors (e.g., via online news and social media). Moreover, efficient everyday learning requires the ability to ignore peripheral information that is not indicative of, but presented in the vicinity of, actual performance feedback (e.g., intrusive social media notifications). The ability to filter out irrelevant information is therefore important for an individual's everyday function and well-being, even when not experienced first-hand. For example, media exposure to disasters and violence relate to negative psychological outcomes (Holman et al., 2014; Hopwood and Schutte, 2017), and information regarding potential threats,

Received June 17, 2022; revised Nov. 18, 2022; accepted Nov. 24, 2022.

Author contributions: K.C.A. and R.P. designed research; K.C.A. performed research; K.C.A. contributed unpublished reagents/analytic tools; K.C.A., I.T., and R.P. analyzed data; K.C.A., I.T., and R.P. wrote the first draft of the paper; K.C.A., I.T., and R.P. edited the paper; K.C.A., I.T., and R.P. wrote the paper.

K.C.A. was supported by Swiss Society of Friends of the Weizmann Institute Postdoctoral Fellowship Grant, and is the incumbent of the Sam and Frances Belzberg Research Fellow Chair in Memory and Learning. This work was supported by a Joy-Ventures Grant, International Science Foundation 2352/19, and ERC-2016-CoG Grant 724910 to R.P. We thank Dr. Edna Furman-Haran and Fanny Attar for MRI procedures.

The authors declare no competing financial interests.

Correspondence should be addressed to Kristoffer Carl Aberg at kc.aberg@gmail.com.

https://doi.org/10.1523/JNEUROSCI.1186-22.2022

obtained via social interactions, may induce maladaptive behaviors (Atlas, 2019; Lindstrom et al., 2019). Finally, distracted learning has detrimental effects on learning performance in general (for review, see Schmidt, 2020). Surprisingly, the neurocomputational mechanisms influenced by affective irrelevant information during learning, and how these interact with personality types that are more easily distracted by affective information, are largely unknown.

Threat-related distractors attract attention and disrupt ongoing behavior, and particularly so for more anxious individuals (Bishop et al., 2004; Bar-Haim et al., 2007; Cisler and Koster, 2010). While there are obvious adaptive advantages of being more attuned to potential threats (e.g., increased survivability) (Ohman, 1986; Grillon, 2002; Robinson et al., 2012), such a sensitivity may have maladaptive properties if subsequent behaviors are guided by irrelevant threat-related information. More generally, failures to ignore irrelevant aversive feedback information could compromise future decision by assigning inappropriate aversive properties to stimuli and the actions that elicited them.

Three different behavioral hypotheses were considered: (1) the null hypothesis that affective distractors have no impact on the learning; (2) affective distractors disrupt the learning (i.e., learning performance in conditions with affective distractors should be reduced); and (3) affective distractors are integrated during learning (i.e., learning performance, respectively, increases and decreases when affective distractors are congruent/incongruent with the relevant feedback). Because anxious individuals are more distracted by threat-related information, we predicted an interaction between inter-individual anxiety levels and irrelevant threat-related information during learning.

To explore the neuronal correlates, our *a priori* analyses focused on the DLPFC given that it has been implicated in attentional selection, such that the DLPFC is engaged when distractors consist of threat-related stimuli, or stimuli to which participants attended in a previous experimental phase (Fales et al., 2008; Browning et al., 2010). For example, Browning et al. (2010) first trained participants to attend either neutral or fearful faces, and reported increased activity in the DLPFC when the attuned stimulus types were subsequently presented as distractors in a different task. Second, converging evidence suggests that aberrant prediction error encoding in the right DLPFC is involved in the acquisition of irrelevant associations (Corlett et al., 2007, 2016), with the prediction error being the mismatch between an experienced and a predicted outcome (Sutton and Barto, 1998). Accordingly, some studies report that prediction error encoding in the R DLPFC correlated with an individual's tendency to learn associations in conditions that normally prevent the formation of stimulus-outcome associations (Corlett and Fletcher, 2012, 2015). As such, abnormal updating of stimulus-outcome contingencies in the R DLPFC may cause learning about stimuli and events that should normally be ignored, eventually leading to the formation of maladaptive beliefs and behaviors. Following reviewer suggestions, we also performed *post hoc* analyses to elucidate potential roles for the amygdala. This is relevant because the amygdala is activated by emotional distractors (for review, see Carretié, 2014), plays a role in emotional learning (for review, see Phelps, 2006), and has been implicated in encoding prediction errors (Averbeck and Costa, 2017; Aberg et al., 2020b). Additionally, amygdala activation during aversive learning and the presentation of irrelevant distractors has been correlated with differences in anxiety levels (for reviews, see Bishop et al., 2004; Lissek et al., 2005; Bishop, 2007; Aupperle and Paulus, 2010; Duval et al., 2015).

Materials and Methods

Participants

After having provided written consent according to the ethical regulations of the Weizmann Institute of Science, 51 participants joined the experiment (behavioral pilot study/fMRI study: 20/31). All participants were right-handed, native Hebrew speakers, and without any previous history of psychiatric or neurologic disorders. The study was performed in accordance with the Declaration of Helsinki.

To ensure sufficient power regarding the behavioral effects in the fMRI study, a power analysis was conducted using data from the behavioral pilot study. This analysis showed that 16 participants are required to detect a one-tailed Pearson correlation coefficient of 0.548 (as obtained in the pilot study) with a power (1-β) of 0.8 and error probability (α) of 0.05. However, because 16 participants are not sufficient to detect interindividual differences in fMRI activation, we recruited additional participants to be more in-line with previous fMRI studies that investigated fMRI activation as a function of trait anxiety in learning and decision-making tasks, for example, $n = 31$ (Browning et al., 2015), $n = 32$ (Bijsterbosch et al., 2015), $n = 25$ (Xu et al., 2013), $n = 30$ (Fung et al., 2019), and $n = 28$ (Aberg et al., 2022).

Two participants frequently fell asleep in the MRI scanner (as indicated by frequently missed trials and post-task interviews). One participant did not perform the task satisfactorily (they pressed the same button in all trials of a block), and one participant displayed excessive movement in all three blocks of learning (as indicated by translational movements in a direction larger than the relevant voxel dimension) (Wylie et al., 2014). Therefore, data from 27 participants were included in the subsequent analyses of fMRI data (20 females; average age ± SD: 25.667 ± 4.961), while data from 20 different participants were included in the behavioral pilot study (11 females; average age ± SD: 27.350 ± 4.171). Trait anxiety was estimated using the State-Trait Anxiety Inventory (Spielberger et al., 1983).

Experimental design and statistical analyses

Reinforcement learning task with distracting emotional faces

Task description. In each trial, participants were presented with a pair of objects and selected the object believed to be more likely to provide Correct feedback (see Fig. 1*A*). The best object in each pair provided Correct feedback with a probability of 0.7 (pilot study) or 0.8 (fMRI study), while the other object provided Correct feedback with a 0.3 (pilot study) or 0.2 (fMRI study) probability.

A schematic of a trial progression is shown in Figure 1*B*. If no response was made within 2.5 s after the presentation of the objects, the letters "Too slow" appeared on the screen and one shekel was deducted. The jittered durations were drawn from a truncated exponential distribution (Dale, 1999), with an average duration of 3 s and a maximum duration of 10 s. To prevent difficulties in identifying the numerical feedback, the location of the feedback number on the screen was identical to the location of the preceding fixation cross.

The different feedback types provided in the experiment are shown in Figure 1*C* (pilot study) and Figure 2*A* (fMRI study). To test for learning differences between appetitive and aversive conditions, the Correct and Incorrect feedbacks were, respectively, +1₪ (a gain of one shekel) or 0₪ (no shekel gained) in a Gain condition, while in a Loss condition the Correct and Incorrect feedbacks were, respectively, 0₪ (no shekel lost) or −1₪ (one shekel lost). The accumulated sum of shekels corresponded to a monetary bonus provided at the end of the experiment. To assess the impact of affective distractors on associative learning, the numerical feedbacks were superimposed on fearful, neutral, or happy faces (see Figs. 1*C*, 2*A*). In Affirmative pairs, the facial expression was matched with the feedback type (e.g., a positive face was presented together with Correct feedback), while in Contradictory pairs the contingencies were reversed (e.g., a positive faces was presented with Incorrect feedback). There were slight differences in the different

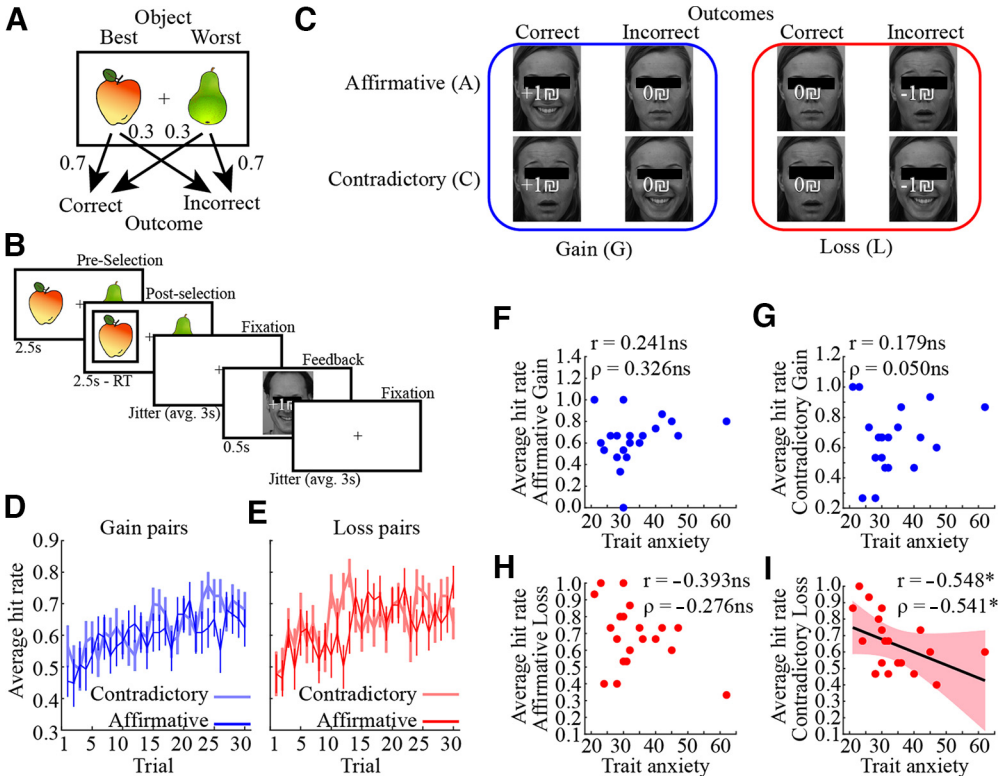

Figure 1. ***A***, Principle of the learning task in the pilot study. In each trial, participants select one object in a pair of objects. The best and worst object in each pair, respectively, provides correct feedback with a probability of 0.7 and 0.3. Each pair is presented 30 times, allowing participants to learn which the best object is by trial and error. ***B***, Illustration of a trial progression. ***C***, Schematic of the outcomes provided in each pair type in the pilot study. In total, four different pair types were presented: Contradictory Loss, Affirmative Loss, Contradictory Gain, and Affirmative Gain. ***D***, ***E***, The average change in performance across participants for the different pair types in Gain (***D***) and Loss (***E***) conditions. A hit is defined as the selection of the best object in a pair. The error bars show the standard error of the mean. ***F–I***, Correlations between Trait anxiety and the average learning performance for Trials 16-20 in Affirmative and Contradictory pair types. Trait anxiety correlated significantly with learning performance in Contradictory Loss pairs only (***I***). *$p < 0.05$; ns not significant ($p > 0.05$). r = Pearson's correlation coefficient. ρ = Spearman's rank-order correlation coefficient.

feedback types presented in the pilot and in the fMRI study. Specifically, in the fMRI study, only emotional faces were presented in the Affirmative and Contradictory conditions because we wanted to add a control condition (Neutral pairs) with only neutral faces to provide a baseline of learning performance without affective distractors.

The learning task was divided into three separate blocks, each consisting of four (pilot) or six (fMRI study) different types of pairs: Affirmative Gain, Affirmative Loss, Contradictory Gain, and Contradictory Loss (as well as Neutral Gain and Neutral Loss for the fMRI study). In total, 12 of 18 different pairs of objects were used, and participants performed 120 trials per block (30 trials per pair in the pilot, and 20 trials per pair in the fMRI study) for a total of 360 trials. Participants were allowed a break between each block. Pairs were presented in an interleaved fashion, such that each pair was presented once before any other pair was repeated. Moreover, the object pairs were randomly assigned to a condition for each participant, and each object was presented equally many times to the left and to the right. Finally, no facial identity was repeated until all facial identities has been presented. For more information regarding the stimuli, see Stimulus selection, below.

To get familiarized with the task and the different facial identities, all participants performed one block of the task outside the scanner. Here, two pairs of objects were presented in one Loss and one Gain pair for a total of 40 trials. These two objects were not used for the main task. Critically, to ensure that all participants understood the goal of the task, they were explicitly instructed that they should try to collect as many shekels as possible and that the faces, including their emotional expression, were irrelevant for performing the task well.

Statistical analyses. Learning performance was defined as the average proportion of selections of the best object in each pair for Trials 16-20 (as well as Trials 26-30 for the pilot study). Correlations with trait anxiety were conducted using Pearson's correlation coefficient, as well as Spearman's rank-order correlation. Tests were one-tailed when testing directed predictions (i.e., positive or negative correlations), while two-tailed tests were used when no direction was predicted. The Bonferroni correction for multiple comparisons was applied where required.

Categorization task

To provide a functional localization of the right DLPFC and the amygdala, participants performed a categorization task before the learning task. This task was inspired by a previous task in which participants categorized attended neutral and fearful faces, and which showed increased DLPFC activation for neutral (vs fearful) faces for participants that had been previously been attuned to fearful faces compared with participants that had been attuned to neutral faces (see Fig. 4*B*, Face Attended condition in the study by Browning et al., 2010). Furthermore, the amygdala is robustly engaged by faces, suggesting it should be activated more strongly by faces than by numbers (Todorov, 2012). We selected a task in which participants attended the faces, rather than presenting them as distractors, because we wanted to prevent any task-related perceived difficulty to confound the results. For example, a differential brain activity between distracting fearful and neutral faces could be wrongly attributed to increased task difficulty caused by, for example, a disruptive attentional bias toward fearful faces. Because it is hard to disentangle these processes, we took advantage of previous reports of differential brain activation when faces were in attentional focus.

Task description. Stimuli were classified as negative, neutral, or positive (see Fig. 4*A*). In each trial, one stimulus was presented from one of six different stimulus types, which could be either a number (−1, 0, or +1) or an emotional face (fearful, neutral, or happy). The classification was performed in the absence of feedback and no specific instructions about the "correct" classification was provided. The six different stimulus types were presented pseudorandomly interleaved in 15 blocks of six

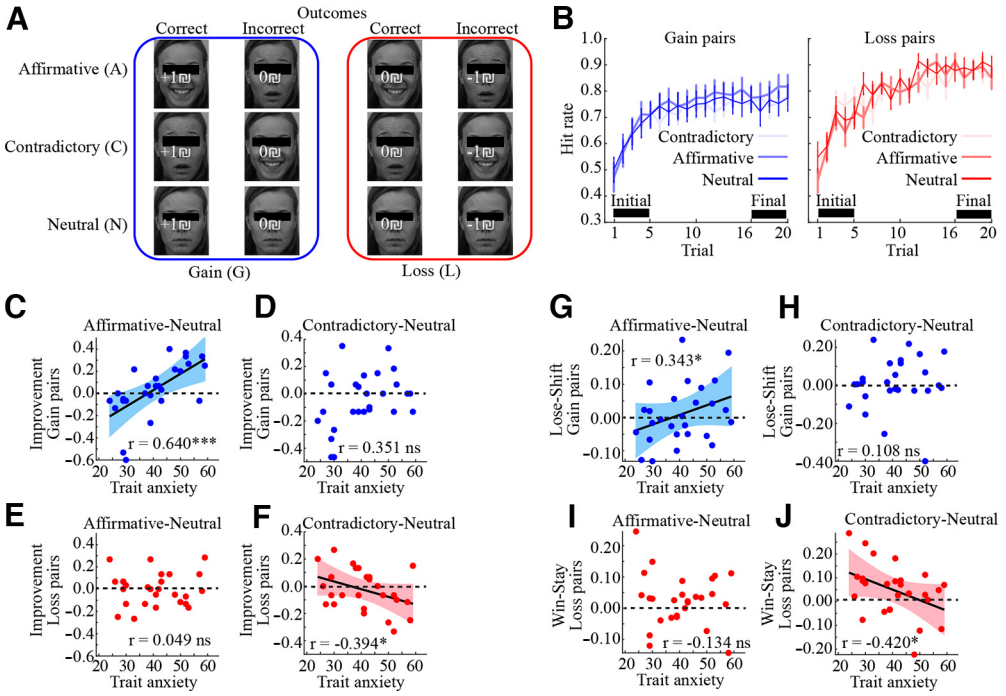

Figure 2. ***A***, Schematic of the outcomes provided in each pair type in the fMRI study. Neutral pairs acted as control conditions by presenting Neutral faces for both Correct and Incorrect outcomes. In total, six different pair types were presented: Contradictory Loss, Affirmative Loss, Neutral Loss, Contradictory Gain, Affirmative Gain, and Neutral Gain. ***B***, The average change in performance across participants for the different pair types in Gain and Loss conditions. A hit is defined as the selection of the best object in a pair. Learning was statistically assessed via the average performance in Trials 16-20. The error bars show the standard error of the mean. ***C–F***, Learning in Affirmative and Contradictory pair types relative the Neutral control condition. Trait anxiety significantly improved/impaired learning in Affirmative Gain pairs (***C***)/Contradictory Loss pairs (***F***). ***G–J***, Win-stay and Lose-shift decisions in Affirmative and Contradictory pair types relative their Neutral counterparts. Trait anxiety significantly increased behavioral switching in Affirmative Gain pairs (***G***) and in Contradictory Loss pairs (***J***). *$p < 0.05$; ***$p < 0.001$; ns not significant ($p > 0.05$).

trials each, where one stimulus from each category was presented in each block. The emotional and the neutral faces were exactly those used for the learning task (see Stimulus selection). A schematic of trial progressions is shown in Figure 4*A*. The intertrial interval durations were drawn from a truncated exponential distribution (Dale, 1999), with an average duration of 3 s and a maximum duration of 10 s. In total, 90 trials were performed (15 trials for each stimulus type). Each facial identity was presented once in each of the fearful, neutral, and happy categories. Of note, the position of the numbers and the faces directly overlapped with the positions of the same stimuli used in the learning task.

Data analysis. The R DLPFC was defined by contrasting fearful and neutral faces (Browning et al., 2010), while the amygdala was defined by contrasting faces and numbers. After defining the ROI on a group level, the average activity within the identified R DLPFC and amygdala clusters was correlated with trait anxiety scores. Because the ROI selections were blind to trait anxiety scores, this procedure conforms to recommendations on how to correlate fMRI data with interindividual factors (Vul et al., 2009).

Stimulus selection

Objects

Eighteen different pairs of objects were created from a colored version of the Snoddgrass and Vanderbilt object dataset, and only familiar objects were selected, as determined by a familiarity rating > 4.0 (Rossion and Pourtois, 2004). All pairs of objects used in the reinforcement learning experiment are presented in Table 1.

Faces

Fearful, neutral, and happy faces from 15 different identities (7 males and 8 females) were selected from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundquist et al., 1998). To ensure that the different facial expressions could be easily identified, only facial identities with a high degree of correspondence between the expressed and the rated emotion were selected. Specifically, only identities with a correct identification > 85% for all of the three facial expressions (Neutral, Fearful, and Happy) were selected (Calvo and Lundqvist, 2008). This resulted in 7 male and 8 female identities (AF01, AF02, AF09, AF16, AF19, AF20, AF29, AF31, AM08, AM10, AM11, AM13, AM17, AM31, AM35). All face stimuli were normalized by rotating and changing the size of each face in accordance with a template image that ensured that the relative locations of the eyes and the tip of the nose were aligned across identities and facial expressions. Finally, the faces were cropped using a rectangular mask which allowed part of the hair to be included in the image.

Table 1. Object pairs used in the reinforcement learning task

Pair	Object descriptions	Object numbers
1	Pen, pencil	167, 168
2	Glasses, book	105, 30
3	Chair, table	53, 226
4	Candle, light bulb	44, 138
5	Key, door	128, 76
6	Tree, flower	241, 91
7	Belt, pants	26, 162
8	Carrot, onion	48, 157
9	Apple, pear	6, 166
10	Cat, dog	49, 73
11	Car, bus	47, 39
12	Lamp, light switch	132, 139
13	Water glass, wine glass	104, 258
14	Shoe, socks	204, 211
15	Telephone, television	227, 228
16	Moon, sun	146, 222
17	Pot, pan	179, 101
18	Fork, spoon	97, 215

Behavioral modeling

Q-learning

Following standard reinforcement learning theory, each object *i* in a pair was assigned an expected value *Qi* which represents the expected

outcome if that object is selected in a trial. *Qi* is updated when object *i* has been selected and there is a mismatch between the expected outcome (*Qi*) and the actual feedback received (φ), i.e., the so-called prediction error (δ). The update of Qi is regulated by a learning rate α as follows:

$$Q(t+1)_i = Q(t)_i + \alpha \cdot \delta(t)_i$$
$$\delta(t)_i = \varphi - Q(t)_i$$

The probability of selecting object *i* in a given trial *t* can be estimated by a soft-max choice probability function (Sutton and Barto, 1998) as follows:

$$p(t)_i = e^{Q(t)_i \cdot \beta} / \left(e^{Q(t)_i \cdot \beta} + e^{Q(t)_j \cdot \beta} \right)$$

The β parameter estimates the trade-off between exploration and exploration/randomness of choice.

Modeling the influence of distractor type

To include distractor types in the model, it was presumed that emotional faces alter the subjective value of the received feedback. For example, happy faces may increase the subjective value of any feedback type, or fearful faces could specifically reduce the subjective value of neutral feedback, etc. To test these notions, the subjective value of the feedback term ϕ was fitted separately for different types of feedback.

In the 12ϕ model, 12 different ϕ's were fitted: one ϕ for each type of face for +1₪ feedback (3 ϕ's), −1₪ feedback (3 ϕ's), 0₪ feedback in Gain pairs (3 ϕ's), and 0₪ feedback in Loss pairs (3 ϕ's).

In the 9ϕ model, 9 ϕ's were fitted: one ϕ for each type of face for +1₪ feedback (3 ϕ's), −1₪ feedback (3 ϕ's), and for 0₪ feedback across Gain and Loss pairs (3 ϕ's).

In the 6ϕ_0 model, 8 ϕ's were fitted: one ϕ collapsed across faces for +1₪ feedback (1 ϕ) and −1₪ feedback (1 ϕ), and one ϕ for each type of face separately for 0₪ feedback in Gain (3 ϕ's) and Loss pairs (3 ϕ's).

In the 3ϕ_0 model, 5 ϕ's were fitted: one ϕ collapsed across faces for +1₪ feedback (1 ϕ) and −1₪ feedback (1 ϕ), and one ϕ for each type of face for 0₪ collapsed across Gain and Loss pairs (3 ϕ's).

In the ϕ_{0FF}, ϕ_{0NH} model, 4 ϕ's were fitted: one ϕ collapsed across faces for +1₪ feedback (1 ϕ) and −1₪ feedback (1 ϕ), one ϕ_{FF} for fearful faces paired with 0₪ feedback, and one ϕ_{NH} for neutral/happy faces paired with 0₪ feedback.

In the ϕ_{0FFG}, ϕ_{0FFL}, ϕ_{0NH} model, 5 ϕ's were fitted: one ϕ collapsed across faces for +1₪ feedback (1 ϕ) and −1₪ feedback (1 ϕ), one ϕ_{FFG} for fearful faces paired with 0₪ feedback in Gain pairs, one ϕ_{FFL} for fearful faces paired with 0₪ feedback in Loss pairs, and one ϕ_{NH} for neutral/happy faces paired with 0₪ feedback.

In the ϕ_{+1},ϕ_{-1} model, 5 ϕ's were fitted: one ϕ collapsed across faces for 0₪ feedback (1 ϕ), one ϕ for happy faces paired with +1₪ feedback (1 ϕ), one ϕ for happy faces paired with −1₪ feedback (1 ϕ), one ϕ for fearful/neutral faces paired with +1₪ feedback (1 ϕ), and one ϕ for fearful/neutral faces paired with −1₪ feedback (1 ϕ).

We also tested another set of models which fit separate subjective values for each numerical feedback (−1, 0, +1) independent of face type. The impact of irrelevant affect is then added via constant "bias" terms (ε's).

In the 3ϕ, 3ε model, 3 ϕ's were fitted: one ϕ for each numerical feedback type (−1, 0, +1; 3 ϕ's), and one ε for each emotional face type (fearful, neutral, happy; 3 ε's).

In the 3ϕ, ε_{FF}, ε_{NH} model, 3 ϕ's were fitted: one ϕ for each numerical feedback type (−1, 0, +1; 3 ϕ's), one ε_{FF} for fearful faces (1 ε), and one ε_{NH} for neutral/happy faces combined (1 ε).

Finally, in the 3ϕ, ε_{0FF}, ε_{0NH} model, 3 ϕ's were fitted: one ϕ for each numerical feedback type (−1, 0, +1; 3 ϕ's), one ε_{0FF} for fearful faces paired with 0₪ feedback (1 ε), and one ε_{0NH} for neutral/happy faces combined and paired with 0₪ feedback (1 ε).

Model fitting and model selection procedures

For each model, the free parameters were fitted individually to each participant's learning behavior by minimizing the negative log-likelihood estimate as follows:

$$LLE = -\ln\left(\prod_1^n p(t)_i \right)$$

Given *n* trials, $p(t)_i$ is the soft-max choice probability of selecting object *i* in trial *t*. To avoid local minima, each fit was repeated 10,000 times with different random starting points for each free parameter. All model fits were compared by calculating the Bayesian Information Criterion (BIC; Schwarz, 1978), which penalizes model fits based on their complexity as follows:

$$BIC = 2 * LLEm + k * \ln(n)$$

LLEm is the minimal log-likelihood estimate, *k* is the number of fitted parameters, and *n* is the total number of trials. The most parsimonious model is the model with the lowest BIC.

To further validate the selection of the most parsimonious model, a protected exceedance probability for each model being the best model was calculated using a Bayesian model selection procedure (Rigoux et al., 2014).

Model simulations

Two different model-simulations of behavior were performed to validate the most parsimonious model.

First, a model-derived probability for selecting the best object in each trial was calculated using each participant's fitted parameters and the history of previous actions and outcomes (Palminteri et al., 2017). To confirm that these model-simulated behaviors reproduce the observed effects of interest, we calculated the same correlations between trait anxiety and learning performance in the different conditions.

Second, to determine whether specific computational parameters drive the observed effects of interest, another set of simulations were performed. These simulations first set all fitted parameters to their average value across participants. Next, the value of the parameter of interest is gradually changed to see whether there are associated changes in the simulated behavioral effect of interest. Performance improvements in all conditions were simulated, and 1000 simulations were conducted for each data point.

MRI data

Image acquisition

MRI images were acquired using a 3T whole-body MRI scanner (Prisma, Siemens) with a 20-channel head coil. Standard structural images were acquired with a T1-weighted 3D sequence (MPRAGE, TR/inversion delay time (TI)/TE = 2300/900/2.32 ms, flip angle = 8 degrees, voxel dimensions = 0.9 mm isotropic, 192 slices). Functional images were acquired with a susceptibility weighted EPI sequence (TR = 2000, TE = 30 ms, flip angle = 75 degrees, voxel dimensions = 3 × 3 × 3.5 mm, 32 slices). The phase-encoding direction was anterior-posterior, the slice order was all even (2-32) followed by all odd (1-31), with a 0% distance factor. No acceleration technique was applied. The MRI scanner was stopped between each block of the learning task (each block lasted ~15 min), while the functional localizer task lasted ~7 min.

Preprocessing

fMRI data were preprocessed and then analyzed using the GLM for event-related designs in SPM12 (Welcome Department of Imaging Neuroscience, London, UK; http://www.fil.ion.ucl.ac.uk/spm). During preprocessing, all functional volumes were realigned to the mean image (with auto-masking applied), coregistered to the structural T1 image, corrected for slice timing, resampled to 2 × 2 × 2 mm voxel size (upsampling of the voxel size to these dimensions has been suggested to

increase the sensitivity of fMRI analyses) (Hopfinger et al., 2000), normalized to the MNI EPI-template, and smoothed using a 6 mm FWHM Gaussian kernel. Please observe that the resampling of voxels is mainly relevant for ROI identification in the functional localizer task.

First-level analyses

General procedure. At the first level, individual event types (e.g., feedbacks, stimuli, or button presses; depending on task, see below) were modeled by a standard synthetic HRF. A 24-parameter model was used to regress out head motion effects from the realigned data (i.e., six head motion parameters, six head motion parameters calculated as the difference between time points *t* and *t* – *1*, and the 12 corresponding squared items) (Friston et al., 1996). Statistical analyses were performed on a voxel-wise basis across the whole brain.

First-level analysis of the categorization task. An event-related design was created with two different event types (stimulus onset and response onset) for each of the six stimulus types (the numbers −1, 0, and +1, and fearful faces, neutral faces, and happy faces). In total, 12 different event types were created, together with a regressor of no interest which included the onset of trials in which no response was made.

ROI

To test the *a priori* hypothesis regarding an involvement of the DLPFC in the present study, an initial R DLPFC mask was created by intersecting the union of Broadmann areas 9 and 46 with the middle frontal gyrus in the right hemisphere. The resulting ROI was then dilated by a factor of 1. All of these steps were performed using the WFU PickAtlas toolbox which also provided predefined ROIs for Broadmann areas 9, 46, and the middle frontal gyrus (Tzourio-Mazoyer et al., 2002; Maldjian et al., 2003, 2004). For the *post hoc* analysis regarding amygdala involvement, an initial amygdala mask was obtained by including all available amygdala subregions provided by the SPM Anatomy toolbox (Eickhoff et al., 2005).

Statistical analyses

To localize the R DLPFC, we contrasted the BOLD signal evoked by neutral and fearful faces, while the amygdala was localized by contrasting BOLD signal evoked by faces and numbers. Significant differential activations within the initial R DLPFC and amygdala masks were tested via *t* tests implemented in SPM using an initial search threshold of $p = 0.001$, and small volume correction (SVC) using a threshold of $p < 0.05$ family-wise error rate (FWE) to correct for multiple comparisons. For display purposes and follow-up analyses (e.g., correlation with individual anxiety levels), β parameter estimates were extracted and averaged from all voxels within significant clusters of activation.

First-level analysis of the learning task. An event-related fMRI design was created with three different event types (stimulus onset, response onset, and feedback onset) for each of four trial types (Gain Correct feedback, Gain Incorrect feedback, Loss Correct feedback, and Loss Incorrect feedback). In addition to these 12 event types for each of three blocks, trials in which no response was made during the picture display were included as a regressor of no interest. To isolate the contribution of the distractors to the prediction error signal, the prediction error term for the selected model (δ Full) was separated into two parts (for similar procedures, see Wittmann et al., 2008; Eldar and Niv, 2015). In brief, a "basic" prediction error term (δ Basic) accounted for variance in the prediction error signal when there is no differential modulation by distractor type (i.e., the values of parameters of interest are set to be equal). Next, a prediction error "boost" term (δ Boost) was created to account for variance above and beyond variance the δ Basic term; the δ Basic term was subtracted from δ Full in each trial *t*, i.e., δ Boost(t) = δ Full(t) – δ Basic(t). To study the fMRI correlates of the two prediction error types δ Basic and δ Boost, their respective values were added as parametric modulators to the feedback onsets. Critically, to elucidate unique variance explained by δ Boost, the values of δ Boost werer orthogonalized with respect to the values of δ Basic (Mumford et al., 2015).

Table 2. Correlations between trait anxiety and learning performance in the pilot study[a]

	Trials 16-20		Trials 26-30	
Condition	*r*	*p*	*r*	*p*
Affirmative Gain	0.241	0.305	0.160	0.501
Affirmative Loss	−0.393	0.086	−0.090	0.705
Contradictory Gain	0.179	0.451	0.288	0.219
Contradictory Loss	−0.548	0.013	−0.438	0.053

[a]*r*: two-tailed uncorrected Pearson's correlation coefficient.

Table 3. Correlations between trait anxiety and learning performance in the fMRI study[a]

	Trials 16-20	
Condition	*r*	*p*
Affirmative Gain	0.640	<0.001
Affirmative Loss	0.049	0.810
Contradictory Gain	0.351	0.073
Contradictory Loss	−0.394	0.042

[a]Data were normalized based on the neutral condition. *r*: two-tailed uncorrected Pearson's correlation coefficient.

Table 4. Correlations between trait anxiety and the proportion of win-stay/lose-shift responses following neutral feedback in the fMRI study[a]

	Win-stay		Lose-shift	
Condition	*r*	*p*	*r*	*p*
Affirmative Gain	x	x	0.343	0.080
Affirmative Loss	x	x	0.108	0.591
Contradictory Gain	−0.134	0.505	x	x
Contradictory Loss	−0.420	0.029	x	x

[a]Data were normalized based on the neutral condition. *r*: two-tailed uncorrected Pearson's correlation coefficient. x indicates that no such action was available for the neutral feedback in this condition.

ROIs

The R DLPFC and amygdala ROIs identified in the separate categorization task.

Statistical analyses

Correlations between prediction errors and BOLD signal in the ROIs were tested using an ROI approach where the average β parameter estimates for each type of prediction error (δ_{Basic}, δ_{Boost}) were extracted from all voxels within the ROIs. These β parameters were then entered into two separate repeated-measures ANOVAs (i.e., one for each prediction error type) with factors Gain/Loss (Gain, Loss pairs) and Feedback (Correct, Incorrect), and Trait anxiety as continuous covariate. Follow-up analyses were conducted using paired *t* tests and Pearson correlations.

Results

Behavior

Behavioral pilot study

An initial pilot study was conducted with 20 participants to explore interactions between trait anxiety and affective distractors during learning. Learning performance was assessed as the average proportion of correct choices in Trials 16-20 and in Trials 26-30. Furthermore, we tested the relationship between anxiety and learning performance separately in each of the four conditions (i.e., Affirmative Loss, Affirmative Gain, Contradictory Loss, and Contradictory Gain). The average learning curve for each condition is shown in Figure 1*D*,*E*. Trait anxiety scores correlated negatively with the average performance only in the Contradictory Loss condition (Trials 16-20: Pearson's $r = -0.548$,

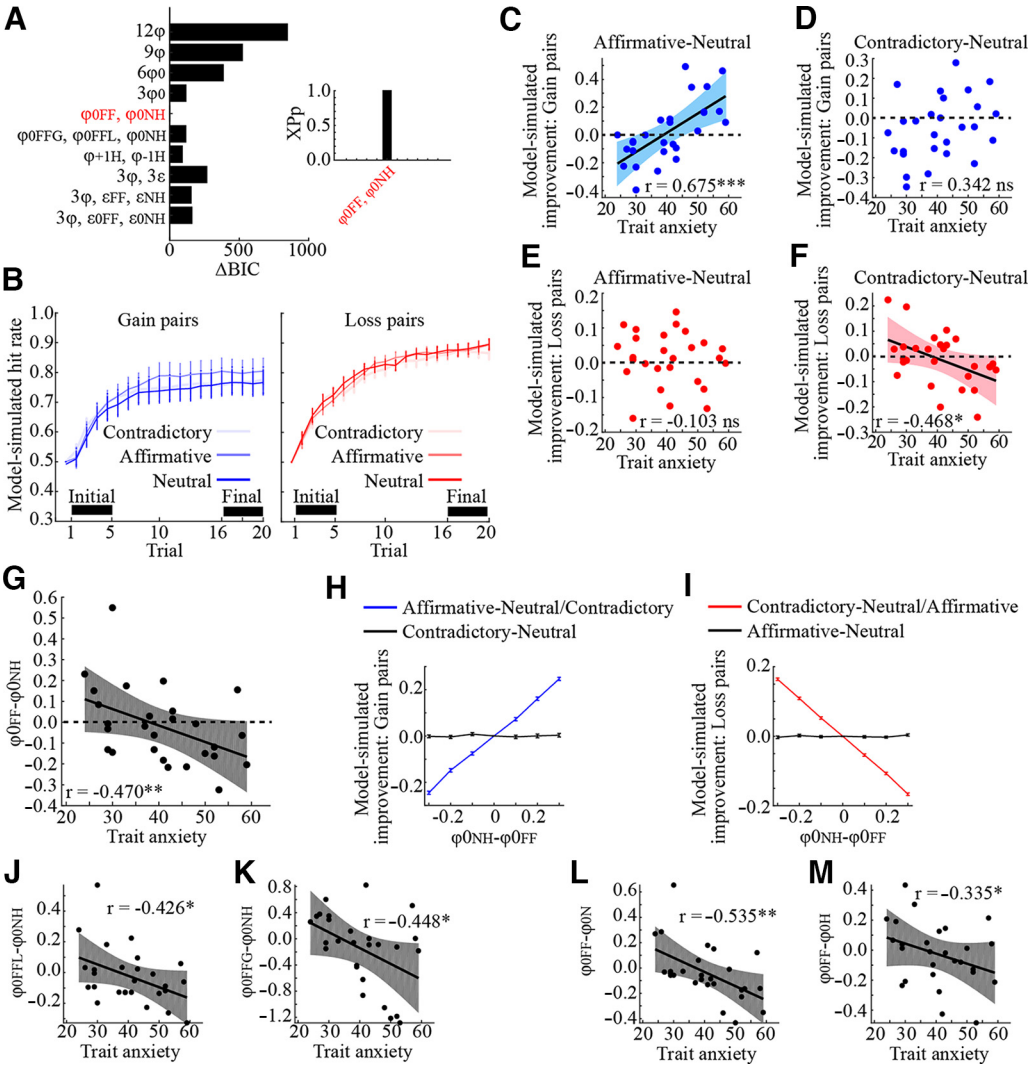

Figure 3. ***A***, Difference in BIC relative the most parsimonious model (highlighted in red). Inset, The protected exceedance probability (XPp) for these models. The most parsimonious model was the most likely model, as evidenced by an exceedance probability of 1.0. ***B***, The average model-simulated change in performance for the different conditions in Gain and Loss pairs. The error bars show the standard error of the mean. ***C–F***, Model-simulated learning in Affirmative and Contradictory pair types relative their Neutral control conditions. Trait anxiety significantly improved/impaired learning in Affirmative Gain pairs (***C***)/Contradictory Loss pairs (***F***). ***G***, Trait anxiety correlated negatively with the difference in the model-fitted subjective values of the neutral 0₪ outcome paired with fearful faces (ϕ_{0FF}) versus neutral/happy faces (ϕ_{0NH}). ***H***, ***I***, Model-simulated performance improvements for gradual changes in the difference between ϕ_{0FF} and ϕ_{0NH}. Smaller values of ϕ_{0FF} (vs ϕ_{0NH}) improve performance in Affirmative Gain pairs (***H***), but impair performance in Contradictory Loss pairs (***I***). For illustration purposes, the performance improvements for when ϕ_{0FF} is equal to ϕ_{0NH} are subtracted from all data points, the *x* axis shows ϕ_{0NH}-ϕ_{0FF}, and the separate lines for Neutral and Contradictory pairs (relative Affirmative pairs) in ***H*** were merged into one line, and similarly were the lines for Neutral and Affirmative pairs (relative Contradictory pairs) in ***I***. ***J***, ***K***, For a model that estimates separate values for the neutral 0₪ outcome paired with fearful faces in Gain (ϕ_{0FFH}) and Loss (ϕ_{0FFL}) conditions, trait anxiety correlated negatively with the difference in the fitted subjective value between the neutral 0₪feedback paired with fearful faces in both Loss (***J***) and Gain (***K***) pairs compared with neutral/happy faces. ***L***, ***M***, For a model that estimates separate values for the neutral 0₪ outcome paired with neutral (ϕ_{0N}) or happy (ϕ_{0H}) faces, trait anxiety correlated negatively with the difference in the fitted subjective value between the neutral 0₪feedback paired with fearful faces compared with both neutral (***L***) and happy (***M***) faces. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

$p = 0.0125$, Trials 26-30: Pearson's $r = -0.438$, $p = 0.053$, two-tailed tests), but not in any other condition (all p values > 0.08; see Table 2). To replicate these results, we conducted a follow-up study where participants also underwent fMRI scanning to provide initial insights into the neurocomputational correlates of the behavioral effects.

fMRI study

The behavioral paradigm of the fMRI study was similar to the one used in the pilot study, with the main addition of a control condition used to normalize learning performance by subtracting the learning performance in the absence of affective distractors (i.e., with neutral faces; Fig. 2*A*). Learning curves are shown in Figure 2*B*, and normalized average learning performances are shown in Figure 2*C–F*.

First, we replicated the main result of the pilot study, namely, a negative correlation between trait anxiety and learning performance in the Contradictory (vs Neutral) Loss condition (Fig. 2*F*; $r = -0.394$, $p = 0.021$, one-tailed Pearson correlation). Because anxiety increases the tendency to display behavioral switching following aversive feedbacks (Aberg and Paz, 2022), we tested whether anxious participants displayed a reduced proportion of win-stay decisions for the Correct feedback in the Contradictory Loss condition (i.e., because the fearful faces were

Table 5. Model fits and parameters[a]

	Models									
Parameters	12ϕ	9ϕ	$6\phi_0$	$3\phi_0$	ϕ_{0FF}, ϕ_{0NH}	ϕ_{0FFG}, ϕ_{0FFL}, ϕ_{0NH}	ϕ_{+1H}, ϕ_{-1H}	3ϕ, 3ε	3ϕ, ε_{FF}, ε_{NH}	3ϕ, ε_{0FF}, ε_{0NH}
Negative LLE	110.8 (7.8)	113.6 (7.8)	113.9 (7.8)	117.8 (7.8)	118.7 (7.8)	117.8 (7.7)	117.3 (7.7)	117.7(7.8)	118.5(7.8)	118.6 (7.8)
BIC	303.9 (15.6)	291.9 (15.6)	286.8 (15.5)	276.9 (15.5)	272.8 (15.6)	276.8 (15.5)	275.9(15.5)	282.4 (15.6)	278.3 (15.5)	278.5 (15.5)
α	0.22 (0.04)	0.21 (0.04)	0.25 (0.04)	0.23 (0.04)	0.21 (0.03)	0.24 (0.04)	0.23 (0.04)	0.23 (0.04)	0.24 (0.04)	0.24 (0.04)
β	0.07 (0.01)	0.07 (0.02)	0.10 (0.02)	0.13 (0.02)	0.12 (0.01)	0.12 (0.02)	0.10 (0.02)	0.19 (0.04)	0.19 (0.03)	0.14 (0.03)
ϕ_{-1}			−0.25 (0.06)	−0.27 (0.07)	−0.23 (0.06)	−0.21 (0.06)		−0.77 (0.05)	−0.78 (0.05)	−0.26 (0.07)
ϕ_0							0.21 (0.06)	0.05 (0.08)	0.03 (0.08)	0.26 (0.12)
ϕ_{+1}			0.70 (0.06)	0.81 (0.06)	0.87 (0.04)	0.76 (0.07)		0.80 (0.06)	0.74 (0.06)	0.82 (0.06)
ϕ_{0NH}					0.37 (0.06)					
ϕ_{-1FFL}	−0.40 (0.08)	−0.23 (0.07)								
ϕ_{-1NL}	−0.25 (0.08)	−0.15 (0.06)								
ϕ_{-1HL}	−0.26 (0.08)	−0.25 (0.07)					−0.25 (0.08)			
ϕ_{0FFL}	0.15 (0.03)		0.21 (0.05)			0.27 (0.07)				
ϕ_{0NL}	0.20 (0.06)		0.26 (0.06)							
ϕ_{0HL}	0.16 (0.07)		0.22 (0.05)							
ϕ_{+1FFG}	0.59 (0.06)	0.50 (0.06)								
ϕ_{+1NG}	0.69 (0.06)	0.59 (0.07)								
ϕ_{+1HG}	0.62 (0.06)	0.53 (0.07)					0.71 (0.07)			
ϕ_{0FFG}	0.08 (0.09)		0.15 (0.09)			0.14 (0.12)				
ϕ_{0NG}	0.10 (0.09)		0.14 (0.09)							
ϕ_{0HG}	0.26 (0.08)		0.30 (0.09)							
ϕ_{0FF}		0.09 (0.05)		0.29 (0.07)	0.35 (0.06)					
ϕ_{0N}		0.15 (0.04)		0.33 (0.06)		0.29 (0.06)				
ϕ_{0H}		0.14 (0.05)		0.32 (0.07)						
$\phi_{-1FF,NL}$							−0.18 (0.07)			
$\phi_{+1FF,NL}$							0.59 (0.07)			
ε_{FF}								0.40 (0.09)	0.45 (0.06)	
ε_N								0.48 (0.08)		
ε_H								0.44 (0.09)		
ε_{NH}									0.49 (0.06)	
ε_{0FF}										0.04 (0.12)
ε_{0NH}										0.07 (0.11)

[a]Data are mean (SEM). LLE, log-likelihood estimate; α, learning rate. β determines the trade-off between exploration and exploitation. ϕ_X is the subjective value for feedback combination X. For example, ϕ_{+1} is the subjective value for −1₪ feedback, ϕ_{+1} is the subjective value for +1₪ feedback, ϕ_{0FF} is the subjective value for the feedback combining 0₪ and fearful faces, and ϕ_{0NH} is the subjective value for 0₪ + happy or fearful face. ε_X is the bias added for feedback combination X. For example, ε_{0FF} is the bias term for neutral 0₪ feedback presented together with fearful faces, and ε_H is the bias term for happy faces.

paired with the neutral 0₪ feedback). As predicted, the proportion of win-stay decisions correlated negatively with trait anxiety in Contradictory (vs Neutral) Loss pairs (Fig. 2*J*; $r = -0.420$, $p = 0.015$, one-tailed Pearson correlation). A similar trend was observed in the Contradictory Loss condition of the pilot study ($r = -0.359$, $p = 0.060$, one-tailed Pearson correlation).

In contrast to the pilot study, we observed a positive correlation between anxiety and learning performance in Affirmative (vs Neutral) Gain pairs (Fig. 2*C*; $r = 0.640$, $p = 0.001$, two-tailed Pearson correlation, p value was corrected for three unplanned comparisons). Notably, in the fMRI study (but not the pilot study), the neutral 0₪ feedback in Affirmative Gain pairs was presented together with a fearful face, therefore providing another opportunity to test whether fearful faces increase the averseness of the neutral 0₪ feedback. Indeed, trait anxiety correlated positively with the proportion of lose-shift decisions in Affirmative (vs Neutral) Gain pairs ($r = 0.343$, $p = 0.040$, one-tailed Pearson correlation; Fig. 2*G*).

Finally, trait anxiety did not correlate significantly with learning performance in the remaining two conditions (Fig. 2*D*,*E*, all uncorrected p values > 0.05, two-tailed Pearson correlations; Table 3), nor with the behavioral switching for neutral 0₪ feedbacks paired with happy faces (Fig. 2*H*,*I*; all uncorrected p values > 0.05, two-tailed Pearson correlations; Table 4).

In summary, the behavioral results from the pilot and the fMRI study suggest that distracting fearful faces increases the averseness of the neutral 0₪ feedback for more anxious individuals. This was demonstrated by increased behavioral switching following this feedback combination, both when it signaled a Correct and when it signaled an Incorrect outcome, which, respectively, caused reduced and improved learning performance.

Behavioral modeling

To explain how anxiety interacts with the distractors, several different behavioral models were designed. To support the aforementioned behavioral result, different subjective feedback values ϕ were fitted for different feedback combinations (for details about the different models and the model-fitting procedures, see Materials and Methods).

A fixed-effect analysis showed an overall lower BIC for the ϕ_{0FF}, ϕ_{0NH} model (Fig. 3*A*), indicating a better fit to behavior on average. Additionally, a random-effects analysis indicated a protected exceedance probability of 1.0 for the same model (Fig. 3*A*, inset), a result that suggests that the selected model is the most likely model to generate the observed behavior (Stephan et al., 2009). Together, these two complementary ways of comparing model indicate the ϕ_{0FF}, ϕ_{0NH} model as being the most parsimonious model.

The selected ϕ_{0FF}, ϕ_{0NH} model contains six free parameters: one learning rate α, one randomness of choice/exploration parameter β, and four feedback parameters (ϕ_{+1}, ϕ_{-1}, ϕ_{0FF}, and ϕ_{0NH}). To clarify, ϕ_{+1} and ϕ_{-1}, respectively,

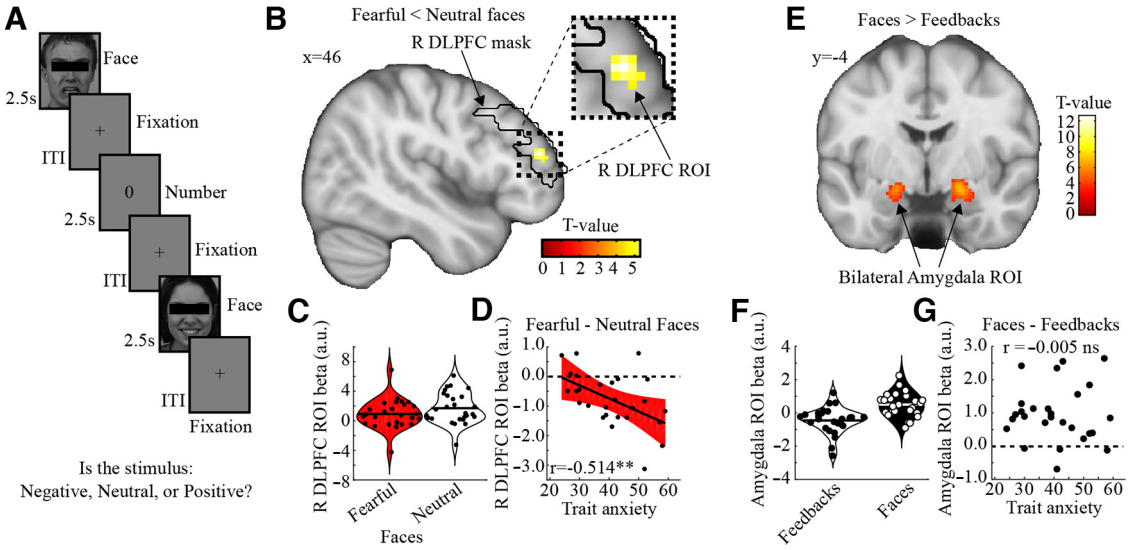

Figure 4. ***A***, Schematic of the functional localizer task. In each trial, a number (−1, 0, +1) or face (Fearful, Neutral, Happy) was presented for 2.5 s. Participants indicated whether the stimulus was perceived as negative, neutral, or positive. No feedback was presented, and participants were not given any particular instructions regarding how stimuli should be categorized. ***B***, The contrast between Neutral and Fearful faces revealed a region in the *a priori* R DLPFC mask that was significantly more activated by Neutral versus Fearful faces. ***C***, For visualization purposes, the average β parameter estimates for Neutral and Fearful faces were extracted for all voxels within the R DLPFC cluster shown in ***B***. ***D***, Trait anxiety correlated negatively with the contrast between Fearful and Neutral faces for the R DLPFC cluster. ***E***, The contrast between Faces and Numbers revealed bilateral regions in the *a priori* amygdala mask that was significantly more activated by Faces versus Numbers. ***F***, For visualization purposes, the average β parameter estimates for Faces and Numbers were extracted for all voxels within the bilateral amygdala cluster shown in ***E***. ***G***, Trait anxiety did not correlate significantly with the contrast between Faces and Numbers for the amygdala cluster.

estimate the subjective value of +1₪ and −1₪ feedbacks and are independent of the distracting faces, while ϕ_{0FF} and ϕ_{0NH}, respectively, estimates the subjective value of the neutral 0₪ feedback paired with fearful (ϕ_{0FF}) and neutral/happy (ϕ_{0NH}) faces. Average fitted model parameters for all models are displayed in Table 5.

To validate the model, and to conform to recent recommendations that effects of interest need to be recovered using model-simulated performance data (Palminteri et al., 2017), the performance of the selected model was simulated using each participant's fitted model parameters. For visualization purposes, the fitted learning curves of the selected model are shown in Figure 3*B*. More importantly, the model successfully reproduced the behavioral effects of interest (Fig. 3*C–F*; compare Fig. 2*C–F*).

One possible explanation for the behavioral results is that the interaction between fearful faces and anxiety reduces the subjective value of the neutral 0₪ feedback. Corroborating this notion, trait anxiety correlated negatively with the difference in the fitted subjective values of the 0₪ feedback paired with fearful and neutral/happy faces (ϕ_{0FF}-ϕ_{0NH}, $r = -0.467$, $p = 0.007$, one-tailed Pearson correlation; Fig. 3*G*). These parameters do not correlate significantly with trait anxiety individually (ϕ_{0FF}: $r = -0.242$, $p = 0.223$; ϕ_{0NH}: $r = 0.051$, $p = 0.802$, two-tailed Pearson correlations).

Additional model simulations were performed to ensure that the impact of the interaction between distractor type and anxiety on learning can actually be attributed to the differential subjective values of ϕ_{0FF} and ϕ_{0NH}. In these simulations, all model parameters are initially set to the average values of the fitted parameters across participants' values (i.e., $\alpha = 0.25$, $\phi_{+1} = -0.25$, $\phi_{0FF} = 0.35$, $\phi_{0NH} = 0.35$, $\phi_{+1} = 0.80$, and $\beta = 0.15$). The values are held constant, except for the values of ϕ_{0FF} and ϕ_{0NH}, which are gradually decreased and increased, respectively, to simulate the modulation by trait anxiety (Fig. 3*G*). Simulated performance improvements are calculated for all conditions, and visualized as a comparison between conditions (for further details, see Materials and Methods). As would be expected, decreases in the difference between ϕ_{0FF} and ϕ_{0NH} improved performance in Affirmative Gain pairs (relative Contradictory and Neutral Gain pairs; Fig. 3*H*) while reducing performance in Contradictory Loss pairs (relative Affirmative and Neutral Loss pairs; Fig. 3*I*).

Table 6. Repeated-measures ANOVA[a]

Predictor	Sum of Squares	df	Mean of Squares	*F*	*p*	η_p^2
(Intercept)	3.449	1	3.449	18.393	< 0.001	
TrAnx	0.386	1	0.386	2.056	0.164	0.076
Error	4.688	25	0.1875			
GaLo	0.0005	1	0.0005	0.001	0.972	<0.001
TrAnx × GaLo	0.003	1	0.003	0.008	0.927	<0.001
Error(GaLo)	10.02	25	0.401			
Feedback	0.024	1	0.024	0.185	0.671	0.007
TrAnx × Feedback	0.227	1	0.227	1.776	0.195	0.066
Error(Feedback)	3.196	25	0.128			
GaLo × Feedback	0.26061	1	0.261	1.357	0.255	0.051
TrAnx × GaLo × Feedback	0.001	1	0.001	0.004	0.951	<0.001
Error(GaLo × Feedback)	4.801	25	0.192			

[a]R DLPFC β parameter estimates for the "basic" prediction error term. TrAnx, Continuous covariate Trait anxiety; GaLo, factor gain/loss (gain or loss pair); Feedback, factor feedback (good, bad); η_p^2, partial η-squared.

Finally, to illustrate the robustness of the main modeling result, we demonstrate that the negative correlation between trait anxiety and the relative difference between fitted 0₪ feedback values for fearful (vs neutral and happy) faces are present across different behavioral models. First, the ϕ_{0FFG}, ϕ_{0FFG}, ϕ_{0NH} model differs from the most parsimonious model by estimating separate ϕ_{0FF}'s in Gain and Loss pairs (i.e., ϕ_{0FF} was separated into two parameters, ϕ_{0FFL} and ϕ_{0FFG}). Trait anxiety correlated negatively with the difference between ϕ_{0FFL} and ϕ_{NH} ($r = -0.426$, $p = 0.013$, one-tailed Pearson correlation; Fig. 3*J*), and with the difference between ϕ_{0FFG} and ϕ_{NH} ($r = -0.448$,

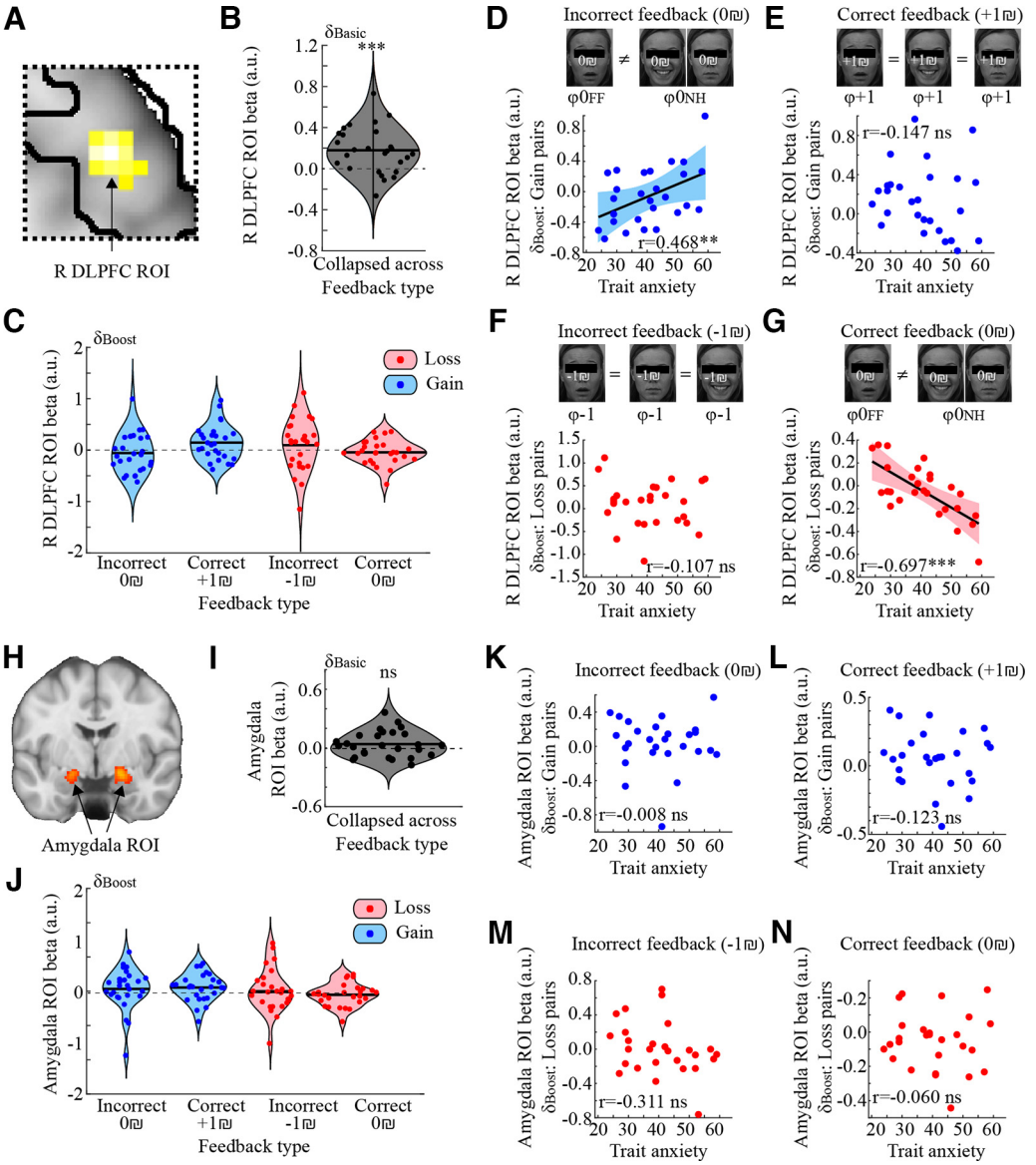

Figure 5. ***A***, The R DLPFC ROI used to analyze prediction error encoding in the learning task. ***B***, The average (solid line) and individual (dots) β parameters for the "basic" prediction error term, δ_{Basic}, averaged across voxels within the R DLPFC ROI for the four different feedback types. On average, activity in the R DLPFC ROI correlated significantly with the "basic" prediction error term independent of Feedback type and Trait anxiety. ***C***, The average (solid line) and individual (dots) β parameters for the prediction error "boost" term, δ_{Boost}, averaged across voxels within the R DLPFC ROI for the four different feedback types. On average, activity in the R DLPFC ROI did not correlate with δ_{Boost} across the four feedback types but showed significant interactions with trait anxiety and feedback types (see main text and ***D–G***). ***D–G***, Correlations between trait anxiety and δ_{Boost} within the four different feedback types. Trait anxiety correlated positively with δ_{Boost} for Incorrect feedback in Gain pairs (***D***) and negatively with δ_{Boost} for Correct feedback in Loss pairs (***G***). The different feedbacks presented in each feedback type is shown above the corresponding plot. The fitted subjective values for the 0₪ feedbacks differed between fearful and happy/neutral faces. ***H***, The amygdala ROI used to analyze prediction error encoding in the learning task. ***I***, The average (solid line) and individual (dots) β parameters for the "basic" prediction error term, δ_{Basic}, averaged across voxels within amygdala for the four different feedback types. On average, activity in the amygdala ROI did not correlate with the "basic" prediction error term, nor was there any interaction with trait anxiety or feedback types. ***J***, The average (solid line) and individual (dots) β parameters for the prediction error "boost" term, δ_{Boost}, averaged across voxels within the amygdala ROI for the four different feedback types. On average, activity in the amygdala ROI did not correlate with δ_{Boost}, nor were there any interactions with trait anxiety or feedback types. ***K–N***, For visualization purposes, correlations between trait anxiety and δ_{Boost} for the four different feedback types are displayed. $^{**}p < 0.01$, $^{***}p < 0.001$, ns = not significant ($p > 0.05$).

$p = 0.010$, one-tailed Pearson correlation; Fig. 3*K*), with a positive correlation between $\phi_{0\text{FFL}}$ and $\phi_{0\text{FFG}}$ ($r = 0.412$, $p = 0.016$, one-tailed Pearson correlation). Second, the $3\phi_0$ model differs from the most parsimonious model by separating the $\phi_{0\text{NH}}$ term into two terms: one term corresponding to the combination of 0₪ feedback paired with neutral faces ($\phi_{0\text{N}}$) and one term for happy faces ($\phi_{0\text{H}}$). Trait anxiety correlated negatively with the difference between $\phi_{0\text{FF}}$ and $\phi_{0\text{N}}$ ($r = -0.535$, $p = 0.002$, one-tailed Pearson correlation; Fig. 3*L*), as well as for the difference between $\phi_{0\text{FF}}$ and $\phi_{0\text{H}}$ ($r = -0.335$, $p = 0.044$, one-tailed Pearson correlation; Fig. 3*M*), and $\phi_{0\text{N}}$ and $\phi_{0\text{H}}$ were positively correlated ($r = 0.816$, $p < 0.001$).

In summary, the selected $\phi_{0\text{FF}}$, $\phi_{0\text{NH}}$ model provides the most parsimonious fit to behavior and provides a plausible and robust explanation for how anxiety interacts with threat-related distractors to modulate learning performance, namely, via a reduced subjective value of neutral 0₪ feedbacks.

Functional neuroimaging

A priori, we hypothesized that atypical prediction error encoding in the R DLPFC, caused by the presence of threat-related distractors, contributes to the learning bias displayed by more anxious individuals. Based on reviewer suggestions, we also conducted a *post hoc* analysis with focus on the amygdala. To this end, we first used a separate task to functionally define the ROIs to be used when analyzing the learning task. Notably, by selecting ROIs in a separate task, we avoid issues of double-dipping (Kriegeskorte et al., 2009), and by selecting ROIs based on group-level data, we minimize the possibility of inflated effect sizes when analyzing interindividual differences in brain activation (Vul et al., 2009).

Functional localization of the R DLPFC and the amygdala via the categorization task

In the functional localizer task, participants categorized numbers (−1, 0, +1) and faces (fearful, neutral, happy) as negative, neutral, or positive (Fig. 4*A*). The contrast between neutral and fearful faces revealed a region within an initial *a priori* defined R DLPFC mask which responded more strongly to neutral (vs fearful) faces (peak voxel coordinate: $x = 46$ year = 44 $z = 18$, $T_{(25)} = 5.151$, pFWE,SVC = 0.013; one-tailed paired t test, Fig. 4*B*,*C*). A negative correlation with trait anxiety shows that the difference in DLPFC BOLD signal between fearful and neutral faces is larger for more anxious individuals (Fig. 4*D*; $r = -0.514$, $p = 0.006$, two-tailed Pearson correlation).

The contrast between faces and numbers revealed bilateral activation within an initial *a priori* defined amygdala mask, which responded more strongly to faces (vs numbers) (peak voxel coordinates: $x = -20$ y = −6 $z = -14$, $T_{(25)} = 6.025$, pFWE,SVC = 0.001; $x = 20$ y = −4 $z = -14$, $T_{(25)} = 6.833$, pFWE,SVC < 0.001; one-tailed paired t tests, Fig. 4*E*,*F*). The collapsed activity within this bilateral ROI showed no correlation with trait anxiety (Fig. 4*G*; $r = -0.005$, $p = 0.981$, two-tailed Pearson correlation).

The obtained R DLPFC cluster and the bilateral amygdala cluster are subsequently used as ROIs in the analyses of prediction error encoding in the learning task.

Neural correlates of prediction errors

To assess the neural correlates of the unique contribution of ϕ_{0FF} (vs ϕ_{0NH}) to the prediction error signal, the prediction error term of the full model (δ_{Full}) is separated into two terms, δ_{Boost} and δ_{Basic} (see Materials and Methods). To assess their neuronal correlates, the β parameter estimates of the δ_{Boost} and the δ_{Basic} terms were extracted from all voxels within the functionally defined R DLPFC and amygdala ROIs. The resulting average β parameters for each ROI were entered into two separate repeated-measures ANOVAs, one for each prediction error type, with factors Gain/Loss (Gain, Loss) and Feedback type (Correct, Incorrect), and Trait anxiety as continuous covariate.

R DLPFC activity correlates with the "basic" prediction error

For the δ_{Basic} term, a repeated-measures ANOVA showed a significant intercept term ($F_{(1,25)} = 18.39$, $p < 0.001$, ANOVA), but no significant main effects or interactions (all p values > 0.16, ANOVA; Table 6). To illustrate this effect, the individual β parameters for the δ_{Basic} term collapsed across the four feedback conditions for the R DLPFC ROI are shown in Figure 5*B*.

This result shows that BOLD signal in the R DLPFC correlates significantly with the magnitude of the "basic" prediction error signal.

Table 7. Repeated-measures ANOVA[a]

Predictor	Sum of squares	df	Mean of squares	F	p	η_p^2
(Intercept)	0.131	1	0.131	1.389	0.250	
TrAnx	0.055	1	0.055	0.584	0.452	0.023
Error	2.365	25	0.095			
GaLo	0.009	1	0.009	0.067	0.798	0.003
TrAnx × GaLo	0.763	1	0.763	6.041	0.021	0.195
Error(GaLo)	3.158	25	0.126			
Feedback	0.027	1	0.027	0.171	0.683	0.007
TrAnx × Feedback	0.737	1	0.737	4.676	0.040	0.158
Error(Feedback)	3.939	25	0.158			
GaLo × Feedback	0.802	1	0.802	5.742	0.024	0.187
TrAnx × GaLo × Feedback	0.087	1	0.087	0.622	0.438	0.024
Error(GaLo × Feedback)	3.492	25	0.140			

[a]R DLPFC β parameter estimates for the "boost" prediction error term. TrAnx, Continuous covariate Trait anxiety; GaLo, factor gain/loss (gain or loss pair); Feedback, factor feedback (good, bad); p_{GG}, Greenhouse-Geisser corrected p value; η_p^2, partial η-squared.

Table 8. Repeated-measures ANOVA[a]

Predictor	Sum of squares	df	Mean of squares	F	p	η_p^2
(Intercept)	0.168	1	0.168	2.311	0.141	
TrAnx	0.009	1	0.009	0.122	0.730	0.005
Error	1.819	25	0.073			
GaLo	0.051	1	0.051	0.731	0.401	0.028
TrAnx × GaLo	0.002	1	0.002	0.026	0.874	0.001
Error(GaLo)	1.757	25	0.070			
Feedback	0.068	1	0.068	1.122	0.300	0.043
TrAnx × Feedback	0.004	1	0.004	0.069	0.795	0.003
Error(Feedback)	1.525	25	0.061			
GaLo × Feedback	0.014	1	0.014	0.256	0.617	0.010
TrAnx × GaLo × Feedback	0.007	1	0.007	0.133	0.718	0.005
Error(GaLo × Feedback)	1.341	25	0.054			

[a]Amygdala β parameter estimates for the "basic" prediction error term. TrAnx, Continuous covariate Trait anxiety; GaLo, factor gain/loss (gain or loss pair); Feedback, factor feedback (good, bad); p_{GG}, Greenhouse-Geisser corrected p value; η_p^2, partial η-squared.

Table 9. Repeated-measures ANOVA[a]

Predictor	Sum of squares	df	Mean of squares	F	p	η_p^2
(Intercept)	0.013	1	0.013	0.190	0.667	
TrAnx	0.116	1	0.116	1.698	0.205	0.064
Error	1.706	25	0.068			
GaLo	0.116	1	0.116	1.911	0.179	0.071
TrAnx × GaLo	0.041	1	0.041	0.671	0.420	0.026
Error(GaLo)	1.512	25	0.061			
Feedback	0.012	1	0.012	0.274	0.605	0.011
TrAnx × Feedback	0.027	1	0.027	0.615	0.440	0.024
Error(Feedback)	1.082	25	0.043			
				0.547	0.466	0.021
TrAnx × GaLo × Feedback	0.077	1	0.077	0.944	0.341	0.036
Error(GaLo × Feedback)	2.034	25	0.081			

[a]Amygdala β parameter estimates for the "boost" prediction error term. TrAnx, Continuous covariate Trait anxiety; GaLo, factor gain/loss (gain or loss pair); Feedback, factor feedback (good, bad); p_{GG}, Greenhouse-Geisser corrected p value; η_p^2, partial η-squared.

R DLPFC activity correlates with the prediction error "boost" in anxious individuals

For the δ_{Boost} term, a repeated-measures ANOVA revealed significant interactions between Trait anxiety × Gain/Loss ($F_{(1,25)} = 6.04$, $p = 0.021$, ANOVA), and Trait anxiety × Feedback ($F_{(1,25)} = 4.68$, $p = 0.040$, ANOVA), but no significant Trait anxiety ×

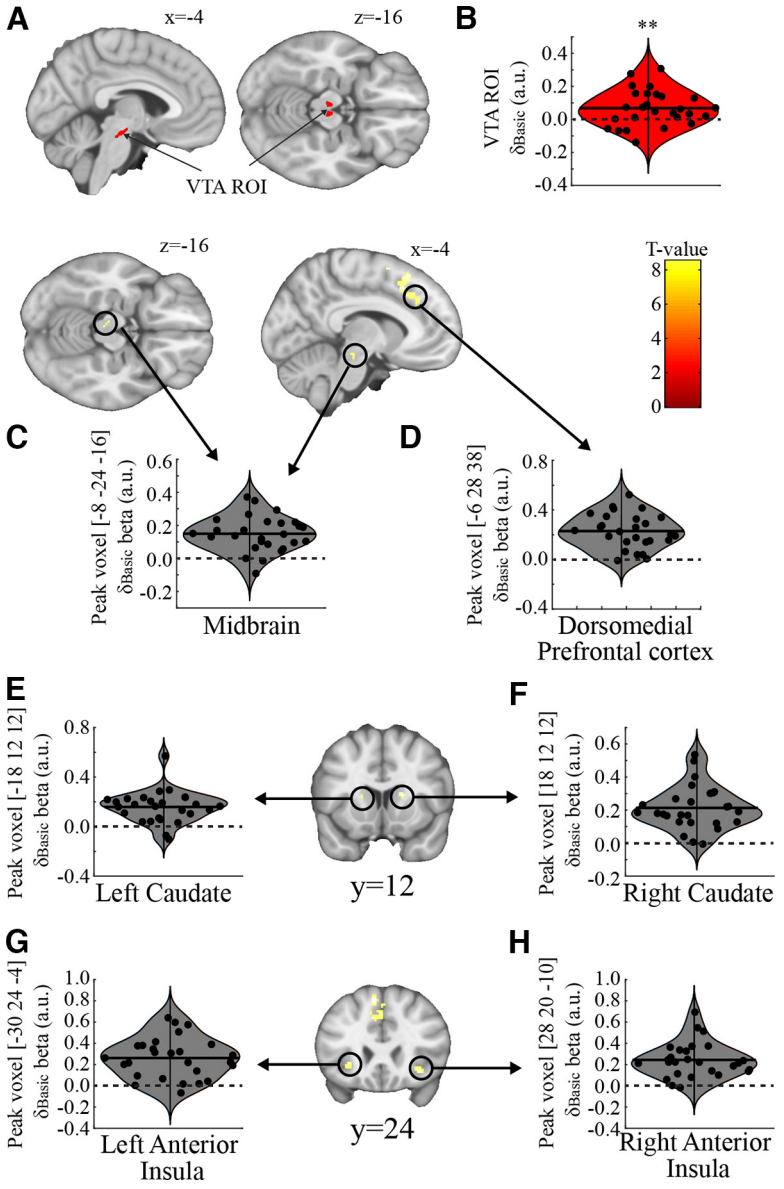

Figure 6. ***A***, Schematic of the VTA ROI. ***B***, Average (solid line) and individual (dots) β parameter estimates for the "basic" prediction error term, δ Basic, within the VTA ROI. On average, BOLD signal in the VTA ROI correlated significantly with δ Basic ($p = 0.001$, one-tailed t test). ***C–H***, BOLD signal in the midbrain, dorsomedial PFC, bilateral striatum, and bilateral anterior insula correlated significantly with δ Basic after applying familywise error rate correction for the whole brain. For visualization purposes, average (solid line) and individual (dots) β parameter estimates were extracted from the peak voxels within each respective cluster. $^{**}p < 0.01$.

Gain/Loss × Feedback interaction ($F_{(1,25)} = 0.62$, $p = 0.438$, ANOVA). Individual β parameters for the δ_{Boost} term in the four feedback conditions are shown in Figure 5*C*. Importantly, the two *a priori* hypotheses were confirmed via a positive correlation between trait anxiety and the β parameters of δ_{Boost} in the Gain Incorrect feedback condition ($r = 0.468$, $p = 0.007$, one-tailed Pearson correlation; Fig. 5*D*), and a negative correlation in the Loss Incorrect feedback condition ($r = -0.697$, $p < 0.001$, one-tailed Pearson correlation; Fig. 5*G*). By contrast, trait anxiety did not correlate with the β parameters of δ_{Boost} in the Gain Correct feedback condition ($r = -0.147$, $p = 0.463$, two-tailed Pearson correlation; Fig. 5*E*) nor in the Loss Incorrect feedback condition ($r = -0.107$, $p = 0.594$, two-tailed Pearson correlation; Fig. 5*F*). In addition to a significant Gain/Loss × Feedback interaction ($F_{(1,25)} = 5.74$, $p = 0.024$), no other effects or interactions are significant (all p values > 0.24; for a full ANOVA table, see Table 7).

In summary, these results confirm that threat-related distractors contribute to altered prediction error encoding in the R DLPFC for anxious individuals, and specifically so in conditions where anxiety correlated with learning performance.

Prediction error coding in the amygdala

As in the previous analysis, the β parameter estimates corresponding to the two prediction error terms, δ_{Boost} and δ_{Basic}, were extracted from all voxels within the amygdala ROI. The resulting average β parameters were entered into the same ANOVAs used for the R DLPFC ROI analysis.

Amygdala activity does not correlate with the "basic" prediction error

For the δ_{Basic} term, the repeated-measures ANOVA revealed no significant main effects or interactions (all p values > 0.14, ANOVA; Table 8). The intercept term, collapsed across conditions and anxiety levels, is shown in Figure 5*I* for visualization purposes.

Amygdala activity does not correlate with the prediction error "boost"

For the δ_{Boost} term, the repeated-measures ANOVA revealed no significant main effects or interactions (all p values > 0.17, ANOVA; Table 9). For visualization purposes, the β parameters for each condition are shown in Figure 5*J*, and correlations with trait anxiety are shown in Figure 5*K–N*.

In summary, no evidence supported a role for the amygdala in prediction error coding.

Whole-brain correlates of the "basic" prediction error

To validate our model-based fMRI procedure, we tested whether activity in the VTA, a region well known for its role in encoding different aspects of reward, including prediction errors (D'Ardenne et al., 2008; Bromberg-Martin et al., 2010; Aberg et al., 2015, 2020a; Schultz, 2016), correlated with the δ_{Basic} term. This analysis was performed by averaging the β parameters related to the δ_{Basic} term for all voxels within a recently developed probabilistic *in vivo* atlas of the VTA (Fig. 6*A*) (Pauli et al., 2018). Indeed, the average β parameters of this VTA ROI were significantly larger than 0.0 (mean ± SEM: 0.068 ± 0.021), $t_{(26)} = 3.291$, $p = 0.001$, one-tailed t test, Fig. 6*B*). Next, correlations with the δ_{Basic} term were tested across the whole-brain using an FWE-corrected threshold of 0.05. A full list of regions correlating with δ_{Basic}, surviving a threshold of an FWE-corrected threshold of 0.05, is reported in Table 10. In short, significant activation was observed in a midbrain region close to the previously used VTA mask (Fig. 6*C*,*D*), in the dorsal anterior cingulate cortex/dorsomedial PFC (Fig. 6*D*), in the bilateral striatum (Fig. 6*E*,*F*), and in the bilateral anterior insula (Fig. 6*G*,*H*). These regions have previously been implicated in the neuronal coding of prediction errors (Garrison et al., 2013).

Table 10. Brain regions showing significantly positive correlations between BOLD signal and the basic prediction error term[a]

		MNI peak coordinate				
Brain region	Hemisphere	*x*	*y*	*z*	*T*	P_{FWE}
Thalamus	Right	12	−6	12	8.517	<0.001
Superior frontal gyrus	Left	−6	26	54	8.498	<0.001
Supplemental motor area	Left	−8	20	44	8.325	0.001
Medial frontal gyrus	Left	−4	28	38	8.282	0.002
Caudate	Right	18	12	12	8.396	0.001
Supplemental motor area	Left	−6	8	62	8.347	0.001
Superior frontal gyrus	Left	−28	56	20	8.343	0.001
Putamen	Left	−22	4	8	8.293	0.001
Caudate	Left	−18	12	12	6.897	0.034
Putamen	Left	−32	−12	−4	8.205	0.002
Putamen	Left	−26	−10	2	7.283	0.015
Superior frontal gyrus	Left	−28	56	0	8.093	0.002
Supramarginal gyrus	Left	−54	−46	36	7.899	0.004
Insula	Right	28	20	−10	7.796	0.004
Insula	Left	−30	26	−4	7.609	0.007
Insula	Left	−34	18	−10	7.397	0.011
Superior frontal gyrus	Right	22	58	28	7.544	0.008
Medial frontal gyrus	Left / Right	0	36	46	7.390	0.011
Supramarginal gyrus	Right	54	−50	38	7.330	0.013
White matter	Left	−12	−8	4	7.252	0.016
Midbrain	Left	−8	−24	−16	7.131	0.021

[a]P_{FWE} indicates familywise error rate (FWEr) corrected *p* values for peak voxel activities across the whole brain. *T* statistics were obtained from *t* tests. Initial search threshold: $p = 0.001$; minimum cluster size: 5 voxels.

Discussion

An increased sensitivity to threat-related information is advantageous in the context of immediate and actual threat avoidance (e.g., when hearing a threatening growl in the forest) (Ohman, 1986). However, it is maladaptive if neutral/safe cues in the environment acquire aversive associations based on irrelevant threat-related information, and these associations subsequently guide behavior.

Here, we report that anxious individuals avoided a neutral stimulus following its pairing with the feedback combination of relevant 0₪ neutral feedback and irrelevant fearful faces, although participants were explicitly instructed that the faces are unrelated to task performance. By showing that exposure to irrelevant affective information lingers and affects behavior beyond the immediate situation, our study extends previous research which focused on the immediate impact of affective distractors, such as alterations in response times, hit rates, or brain activations (e.g., within the same trial) (Bar-Haim et al., 2007).

Importantly, the threat-related distractors did not simply disrupt the learning process, as would be indicated by an overall reduced learning performance in conditions with the feedback combination of fearful faces and neutral 0₪ feedback. By contrast, anxious individuals displayed, respectively, reduced or improved performance in conditions where this feedback combination represented the Correct or the Incorrect outcome. In support, behavioral modeling further showed that high anxiety was associated with a reduced subjective value of the neutral 0₪ feedback when paired with fearful faces both when it signaled Correct and Incorrect outcomes (compared with happy and neutral faces). A third dissociation was observed in the fMRI data, with a stronger/weaker coupling between the prediction error signal, uniquely accounted for by fearful faces, and R DLPFC BOLD signal for feedbacks where anxious individuals showed increased/decreased learning performance. Together, these results indicate that anxiety is associated with an increased integration of irrelevant threat-related information during feedback processing (and not just disrupted learning processes). From an evolutionary perspective, it makes sense that information related to potential threats is integrated during learning, rather than disrupting it. However, this ability comes at the cost of increased avoidance of beneficial situations in which a potential treat was occasionally detected.

It has been suggested that anxiety disorders develop from abnormal learning processes, for example, amplified fear learning (Lissek et al., 2005) and overgeneralization (i.e., the transfer of aversive properties from a fear-conditioned neutral stimulus to other perceptually and conceptually similar neutral stimuli) (Lissek et al., 2014). Additionally, trait anxiety (e.g., the general tendency to experience distress in everyday life situations) may indicate a vulnerability to develop a mental illness (Chambers et al., 2004; Weger and Sandi, 2018). The identification of abnormal learning processes in trait anxiety could therefore help understand external factors and internal mechanisms that contribute to the development of dysfunctional behaviors and mental illness. Based on the present results, we propose that this includes the maladaptive formation of associations between neutral stimuli/events and irrelevant threat-related information, as these may result in inappropriate avoidance behaviors.

Anxious individuals showed increased integration of fearful faces with the neutral 0₪ feedback, but not with +1₪ and −1₪ feedbacks. One potential explanation for this result could be that anxious individuals call on additional, salient sources of information to resolve uncertain feedbacks. To clarify, in the present study, the −1₪ and +1₪ feedbacks always indicated the worst and best possible outcomes, while the 0₪ feedback signaled either a correct (in Loss conditions) or an incorrect (in Gain conditions) outcome, causing it to be more uncertain. This interpretation is in line with findings that anxiety increases aversion to uncertainty (Hartley and Phelps, 2012; Grupe and Nitschke, 2013), the motivation to reduce uncertainty (Aberg et al., 2022), and distractibility by threat-related information (Bar-Haim et al., 2007). Future research may profit from looking at how irrelevant salient information guides the processing of uncertain feedback in high anxiety.

The present study used a between-subject design to study the interaction between anxiety and threat-related distractors during learning. A complementary way to assess behavioral interactions with anxiety is via alterations of stress and state anxiety, something which could be accomplished by, for example threat-of-shock manipulations (Schmitz and Grillon, 2012; Robinson et al., 2013). This approach is beneficial because it could be used to combine a powerful within-subject design (i.e., conditions with and without stress, or induced anxiety) with a between-subject design (e.g., trait anxiety measures, or patient vs control groups). This approach may be particularly fruitful to further research on findings that individuals with anxious predispositions respond differently in stressful situations (Meijer, 2001; Indovina et al., 2011; Aberg and Paz, 2022).

In accord with previous studies (for review, see Garrison et al., 2013), a number of brain regions in the present study, including the VTA, the striatum, anterior cingulate cortex, anterior insula, and the R DLPFC, encoded a "basic" prediction error signal. However, only the R DLPFC of anxious individuals correlated with additional variance in the prediction error signal that was uniquely attributed to the fearful faces. This correlation was positive in conditions where anxiety improved performance, but negative when anxiety show impaired performance. These results

are in accordance with previous research showing that the strength of neuronal prediction error encoding correlates with the amount of learning (Schönberg et al., 2007; Aberg et al., 2015, 2016a), and complement behavioral and physiological reports of links between personality traits and reinforcement learning biases (Browning et al., 2015; Aberg et al., 2016b, 2017). These results also bridge separate reports of an involvement of the DLPFC in attentional bias to threat (Bishop, 2009), prediction error encoding (Fletcher et al., 2001; Corlett et al., 2004), and the acquisition of irrelevant associations via altered prediction error encoding (Corlett and Fletcher, 2012, 2015). Indeed, similar and converging pieces of evidence support a theory in which aberrant prediction error encoding in the R DLPFC is believed to enable maladaptive learning about stimuli, events, and outcomes that are not related (Corlett et al., 2007, 2016). The present study adds to this theory by suggesting that one source of "aberrancy" stems from failures in suppressing attention to irrelevant sources of threat-related information; that is, these stimuli may grab attention and engage learning processes like any other (relevant) stimulus, and particularly so with high anxiety.

By contrast, we did not observe any involvement of the amygdala in coding prediction errors, nor any interaction with trait anxiety. Although the amygdala plays a prominent role in fear learning and anxiety (Phelps, 2006; Duval et al., 2015; Tovote et al., 2015), only scarce evidence report correlations between amygdala activity and prediction errors (McHugh et al., 2014; Meffert et al., 2015; Aberg et al., 2020b). One possibility is that the amygdala codes for other features related to learning and the prediction error signal, such as surprise, sometimes defined as the unsigned prediction error signal (Li et al., 2011; Klavir et al., 2013). Further, although the amygdala is activated by affective distractors, and particularly so for more anxious individuals (Bishop et al., 2004; Bishop, 2009), it has to our best knowledge not been implicated in the learning of irrelevant information.

Limitations

Anxiety was estimated using the standard Spielberger's Trait-Anxiety Inventory (Spielberger et al., 1983), which provides a gradual scale for the normal (subclinical) range of anxiety. Using a continuous scale has the benefit of correlating behavior across a distribution of anxiety scores, rather than just comparing performance across two somewhat arbitrarily divided populations (patients vs controls). Additionally, by studying anxiety within the normal range, we can determine how maladaptive decisions are mediated by irrelevant distractors even in healthy individuals. Because such maladaptive decision may have a huge impact on daily-life in all individuals, and definitely on societies and industry, we actually believe that more studies should use gradual scales over non-clinical populations (Browning et al., 2015; Fung et al., 2019; Gagne et al., 2020). That said, Spielberger's Trait-Anxiety Inventory has been debated for its lack of convergent and discriminant validity, suggesting that it estimates "negative affectivity" rather than proneness to anxiety per-SE (Balsamo et al., 2013). Yet, because negative affectivity is closely linked to psychopathology (Kotov et al., 2010; Stanton and Watson, 2014), and has been noted as a vulnerability factor for developing anxiety and depression (Clark et al., 1994), our results still bear significant relevance.

Our main behavioral results were replicated between two separate groups of participants (i.e., negative correlations between trait anxiety and learning performance in the Contradictory Loss condition was observed in both the pilot and in the fMRI study), and were replicated within another condition in the fMRI study (i.e., the feedback combination of fearful faces + neutral 0₪ feedback increased behavioral switching in both the Contradictory Loss and the Affirmative Gain condition). Furthermore, these latter results were associated with two dissociations in the fMRI results, namely, opposite correlations with trait anxiety and the coupling between R DLPFC activity and the prediction error signal associated with fearful faces. Being able to replicate results across- and within-groups speaks in favor of the robustness of our results.

Importantly, we would like to stress that our fMRI findings were obtained with a relatively small sample size ($N = 27$), and therefore needs to be regarded as provisional. In particular, although many factors may contribute to the reliability of brain-behavior correlations in fMRI data, including behavioral task, amount of data per participant, targeted brain regions, and method of analysis, recent efforts suggest "... that with sample sizes in the range of those often used in fMRI studies (i.e., 20-30 participants), one cannot be confident that all of the regions appearing to correlate with individual differences in behavior are reliable, or that other regions have not been missed altogether" (Grady et al., 2021). Future studies should therefore expand on the issue and validate the robustness of the present fMRI results.

In conclusion, the present study displays a learning bias for individuals with high trait anxiety caused by an entanglement between threat-related distractors and ongoing learning processes. This bias may be particularly unhealthy in modern society, where exposure to irrelevant threat-related information is increasingly prevalent via online news reporting and social networking sites. The present study describes a new pathway for how threat-related information may become entrenched in the anxious psyche.

References

Aberg KC, Paz R (2022) Stress-induced avoidance in mood disorders. Nat Hum Behav 6:915–918.

Aberg KC, Doell KC, Schwartz S (2015) Hemispheric asymmetries in striatal reward responses relate to approach-avoidance learning and encoding of positive-negative prediction errors in dopaminergic midbrain regions. J Neurosci 35:14491–14500.

Aberg KC, Doell KC, Schwartz S (2016a) The left hemisphere learns what is right: hemispatial reward learning depends on reinforcement learning processes in the contralateral hemisphere. Neuropsychologia 89:1–13.

Aberg KC, Doell KC, Schwartz S (2016b) Linking individual learning styles to approach-avoidance motivational traits and computational aspects of reinforcement learning. PLoS One 11:e0166675.

Aberg KC, Muller J, Schwartz S (2017) Trial-by-trial modulation of associative memory formation by reward prediction error and reward anticipations revealed by a biologically plausible computational model. Front Hum Neurosci 11:56.

Aberg KC, Kramer EE, Schwartz S (2020a) Interplay between midbrain and dorsal anterior cingulate regions arbitrates lingering reward effects on memory encoding. Nat Commun 11:1829.

Aberg KC, Kramer EE, Schwartz S (2020b) Neurocomputational correlates of learned irrelevance in humans. Neuroimage 213:116719.

Aberg KC, Toren I, Paz R (2022) A neural and behavioral trade-off between value and uncertainty underlies exploratory decisions in normative anxiety. Mol Psychiatry 27:1573–1587.

Atlas LY (2019) How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. Curr Opin Behav Sci 26:121–129.

Aupperle RL, Paulus MP (2010) Neural systems underlying approach and avoidance in anxiety disorders. Dialogues Clin Neurosci 12:517–531.

Averbeck BB, Costa VD (2017) Motivational neural circuits underlying reinforcement learning. Nat Neurosci 20:505–512.

Balsamo M, Romanelli R, Innamorati M, Ciccarese G, Carlucci L, Saggino A (2013) The State-Trait Anxiety Inventory: shadows and lights on its construct validity. J Psychopathol Behav Assess 35:475–486.

Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IJzendoorn MH (2007) Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychol Bull 133:1–24.

Bijsterbosch J, Smith S, Bishop SJ (2015) Functional connectivity under anticipation of shock: correlates of trait anxious affect versus induced anxiety. J Cogn Neurosci 27:1840–1853.

Bishop SJ (2007) Neurocognitive mechanisms of anxiety: an integrative account. Trends Cogn Sci 11:307–316.

Bishop SJ (2009) Trait anxiety and impoverished prefrontal control of attention. Nat Neurosci 12:92–98.

Bishop SJ, Duncan J, Brett M, Lawrence AD (2004) Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. Nat Neurosci 7:184–188.

Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68:815–834.

Browning M, Holmes EA, Murphy SE, Goodwin GM, Harmer CJ (2010) Lateral prefrontal cortex mediates the cognitive modification of attentional bias. Biol Psychiatry 67:919–925.

Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. Nat Neurosci 18:590–596.

Calvo MG, Lundqvist D (2008) Facial expressions of emotion (KDEF): identification under different display-duration conditions. Behav Res Methods 40:109–115.

Carretié L (2014) Exogenous (automatic) attention to emotional stimuli: a review. Cogn Affect Behav Neurosci 14:1228–1258.

Chambers JA, Power KG, Durham RC (2004) The relationship between trait vulnerability and anxiety and depressive diagnoses at long-term follow-up of Generalized Anxiety Disorder. J Anxiety Disord 18:587–607.

Cisler JM, Koster EH (2010) Mechanisms of attentional biases towards threat in anxiety disorders: an integrative review. Clin Psychol Rev 30:203–216.

Clark LA, Watson D, Mineka S (1994) Temperament, personality, and the mood and anxiety disorders. J Abnorm Psychol 103:103–116.

Corlett PR, Fletcher PC (2012) The neurobiology of schizotypy: fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. Neuropsychologia 50:3612–3620.

Corlett PR, Fletcher PC (2015) Delusions and prediction error: clarifying the roles of behavioural and brain responses. Cogn Neuropsychiatry 20:95–105.

Corlett PR, Aitken MR, Dickinson A, Shanks DR, Honey GD, Honey RA, Robbins TW, Bullmore ET, Fletcher PC (2004) Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. Neuron 44:877–888.

Corlett PR, Honey GD, Fletcher PC (2007) From prediction error to psychosis: ketamine as a pharmacological model of delusions. J Psychopharmacol 21:238–252.

Corlett PR, Honey GD, Fletcher PC (2016) Prediction error, ketamine and psychosis: an updated model. J Psychopharmacol 30:1145–1155.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267.

Dale AM (1999) Optimal experimental design for event-related fMRI. Hum Brain Mapp 8:109–114.

Duval ER, Javanbakht A, Liberzon I (2015) Neural circuits in anxiety and stress disorders: a focused review. Ther Clin Risk Manag 11:115–126.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325–1335.

Eldar E, Niv Y (2015) Interaction between emotional state and learning underlies mood instability. Nat Commun 6:6149.

Fales CL, Barch DM, Rundle MM, Mintun MA, Snyder AZ, Cohen JD, Mathews J, Sheline YI (2008) Altered emotional interference processing in affective and cognitive-control brain circuitry in major depression. Biol Psychiatry 63:377–384.

Fletcher PC, Anderson JM, Shanks DR, Honey R, Carpenter TA, Donovan T, Papadakis N, Bullmore ET (2001) Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nat Neurosci 4:1043–1048.

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI time-series. Magn Reson Med 35:346–355.

Fung BJ, Qi S, Hassabis D, Daw N, Mobbs D (2019) Slow escape decisions are swayed by trait anxiety. Nat Hum Behav 3:702–708.

Gagne C, Zika O, Dayan P, Bishop SJ (2020) Impaired adaptation of learning to contingency volatility in internalizing psychopathology. Elife 9:e61387.

Garrison J, Erdeniz B, Done J (2013) Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci Biobehav Rev 37:1297–1310.

Grady CL, Rieck JR, Nichol D, Rodrigue KM, Kennedy KM (2021) Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. Hum Brain Mapp 42:204–219.

Grillon C (2002) Startle reactivity and anxiety disorders: aversive conditioning, context, and neurobiology. Biol Psychiatry 52:958–975.

Grupe DW, Nitschke JB (2013) Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. Nat Rev Neurosci 14:488–501.

Hartley CA, Phelps EA (2012) Anxiety and decision-making. Biol Psychiatry 72:113–118.

Holman EA, Garfin DR, Silver RC (2014) Media's role in broadcasting acute stress following the Boston Marathon bombings. Proc Natl Acad Sci USA 111:93–98.

Hopfinger JB, Buchel C, Holmes AP, Friston KJ (2000) A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. Neuroimage 11:326–333.

Hopwood TL, Schutte NS (2017) Psychological outcomes in reaction to media exposure to disasters and large-scale violence: a meta-analysis. Psychol Violence 7:316–327.

Indovina I, Robbins TW, Nunez-Elizalde AO, Dunn BD, Bishop SJ (2011) Fear-conditioning mechanisms associated with trait vulnerability to anxiety in humans. Neuron 69:563–571.

Klavir O, Genud-Gabai R, Paz R (2013) Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning. Neuron 80:1290–1300.

Kotov R, Gamez W, Schmidt F, Watson D (2010) Linking 'big' personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychol Bull 136:768–821.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.

Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND (2011) Differential roles of human striatum and amygdala in associative learning. Nat Neurosci 14:1250–1252.

Lindstrom B, Golkar A, Jangard S, Tobler PN, Olsson A (2019) Social threat learning transfers to decision making in humans. Proc Natl Acad Sci USA 116:4732–4737.

Lissek S, Powers AS, McClure EB, Phelps EA, Woldehawariat G, Grillon C, Pine DS (2005) Classical fear conditioning in the anxiety disorders: a meta-analysis. Behav Res Ther 43:1391–1424.

Lissek S, Kaczkurkin AN, Rabin S, Geraci M, Pine DS, Grillon C (2014) Generalized anxiety disorder is associated with overgeneralization of classically conditioned fear. Biol Psychiatry 75:909–915.

Lundquist D, Flykt A, Öhman A (1998) The Karolinska Directed Emotional Faces—KDEF [CD-ROM]. Stockholm, Sweden: Department of Clinical Neuroscience, Karolinska Institutet.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Maldjian JA, Laurienti PJ, Burdette JH (2004) Precentral gyrus discrepancy in electronic versions of the Talairach atlas. Neuroimage 21:450–455.

McHugh SB, Barkus C, Huber A, Capitao L, Lima J, Lowry JP, Bannerman DM (2014) Aversive prediction error signals in the amygdala. J Neurosci 34:9024–9033.

Meffert H, Brislin SJ, White SF, Blair JR (2015) Prediction errors to emotional expressions: the roles of the amygdala in social referencing. Soc Cogn Affect Neurosci 10:537–544.

Meijer J (2001) Stress in the relation between trait and state anxiety. Psychol Rep 88:947–964.

Mumford JA, Poline JB, Poldrack RA (2015) Orthogonalization of regressors in FMRI models. PLoS One 10:e0126255.

Ohman A (1986) Face the beast and fear the face: animal and social fears as prototypes for evolutionary analyses of emotion. Psychophysiology 23:123–145.

Palminteri S, Wyart V, Koechlin E (2017) The importance of falsification in computational cognitive modeling. Trends Cogn Sci 21:425–433.
Pauli WM, Nili AN, Tyszka JM (2018) A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. Sci Data 5:180063.
Phelps EA (2006) Emotion and cognition: insights from studies of the human amygdala. Annu Rev Psychol 57:27–53.
Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies: revisited. Neuroimage 84:971–985.
Robinson OJ, Charney DR, Overstreet C, Vytal K, Grillon C (2012) The adaptive threat bias in anxiety: amygdala-dorsomedial prefrontal cortex coupling and aversive amplification. Neuroimage 60:523–529.
Robinson OJ, Vytal K, Cornwell BR, Grillon C (2013) The impact of anxiety upon cognition: perspectives from human threat of shock studies. Front Hum Neurosci 7:203.
Rossion B, Pourtois G (2004) Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. Perception 33:217–236.
Schmidt SJ (2020) Distracted learning: big problem and golden opportunity. J Food Sci Educ 19:278–291.
Schmitz A, Grillon C (2012) Assessing fear and anxiety in humans using the threat of predictable and unpredictable aversive events (the NPU-threat test). Nat Protoc 7:527–532.
Schönberg T, Daw ND, Joel D, O'Doherty J (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867.
Schultz W (2016) Dopamine reward prediction-error signalling: a two-component response. Nat Rev Neurosci 17:183–195.
Schwarz GE (1978) Estimating the dimension of a model. Ann Statist 6:461–464.
Spielberger CD, Gorsuch RL, Lushene R, Vagg PR, Jacobs GA (1983) Manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists.
Stanton K, Watson D (2014) Positive and negative affective dysfunction in psychopathology. Soc Pers Psychol Compass 8:555–567.
Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017.
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: Massachusetts Institute of Technology.
Todorov A (2012) The role of the amygdala in face perception and evaluation. Motiv Emot 36:16–26.
Tovote P, Fadok JP, Luthi A (2015) Neuronal circuits for fear and anxiety. Nat Rev Neurosci 16:317–331.
Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289.
Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4:274–290.
Weger M, Sandi C (2018) High anxiety trait: a vulnerable phenotype for stress-induced depression. Neurosci Biobehav Rev 87:27–37.
Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58:967–973.
Wylie GR, Genova H, DeLuca J, Chiaravalloti N, Sumowski JF (2014) Functional magnetic resonance imaging movers and shakers: does subject-movement cause sampling bias? Hum Brain Mapp 35:1–13.
Xu P, Gu R, Broster LS, Wu R, Van Dam NT, Jiang Y, Fan J, Luo YJ (2013) Neural basis of emotional decision making in trait anxiety. J Neurosci 33:18641–18653.