



Irrelevant threats linger in high anxiety

Document Version:

Accepted author manuscript (peer-reviewed)

Citation for published version:

Aberg, KC, Toren, I & Paz, R 2022, 'Irrelevant threats linger in high anxiety', *The Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1186-22.2022

Total number of authors: 3

Digital Object Identifier (DOI): 10.1523/JNEUROSCI.1186-22.2022

Published In: The Journal of Neuroscience

License: Other

General rights

@ 2020 This manuscript version is made available under the above license via The Weizmann Institute of Science Open Access Collection is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognize and abide by the legal requirements associated with these rights.

How does open access to this work benefit you?

Let us know @ library@weizmann.ac.il

Take down policy

The Weizmann Institute of Science has made every reasonable effort to ensure that Weizmann Institute of Science content complies with copyright restrictions. If you believe that the public display of this file breaches copyright please contact library@weizmann.ac.il providing details, and we will remove access to the work immediately and investigate your claim.



Research Articles: Behavioral/Cognitive

Irrelevant threats linger in high anxiety

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

Cite as: J. Neurosci 2022; 10.1523/JNEUROSCI.1186-22.2022

Received: 17 June 2022 Revised: 18 November 2022 Accepted: 24 November 2022

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Copyright © 2022 the authors

- <u>JNeurosci Accepted Manuscript</u>
- 1 Title: Irrelevant threats linger and affect behavior in high anxiety
- 2 Abbreviated title: Irrelevant threats linger in high anxiety
- 3 Authors: Kristoffer C. Aberg¹, Ido Toren¹, Rony Paz¹
- 4 1 Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel
- 5 **Corresponding author:**
- 6 Kristoffer Carl Aberg
- 7 Department of Brain Sciences
- 8 Weizmann Institute of Science
- 9 Rehovot 76100 ISRAEL
- 10 Tel : +972 (0)58 448 35 47
- 11 Fax : <u>+972 (0)8 9344131</u>
- 12 Email: <u>kc.aberg@gmail.com</u>
- 13 Number of pages: 42
- 14 Number of Figures: 6
- 15 Number of Tables: 10
- 16 Number of words Abstract: 250
- 17 Number of words Introduction: 646
- 18 Number of words Discussion: 1614
- 19 Conflict of interest: The authors declare no competing financial interests.
- 20 Acknowledgements: K.C. Aberg was supported by the Swiss Society of Friends of the Weizmann
- 21 Institute Postdoctoral Fellowship grant, and is the incumbent of the Sam and Frances Belzberg
- 22 Research Fellow Chair in Memory and Learning. We thank Dr. Edna Furman-Haran and Fanny Attar
- 23 for MRI procedures. The work was supported by a Joy-Ventures grant, an ISF #2352/19 and an ERC-
- 24 2016-CoG #724910 grant to R. Paz.
- 25
- 26 Keywords: Anxiety; Maladaptive; Associative learning; Dorsolateral PFC; Irrelevant; Prediction error

27 Abstract

28 Threat-related information attracts attention and disrupts on-going behavior, and particularly so for 29 more anxious individuals. Yet, it is unknown how and to what extent threat-related information leave 30 lingering influences on behavior, e.g. by impeding on-going learning processes. Here, human male and female participants (N=47) performed probabilistic reinforcement learning tasks where 31 32 irrelevant distracting faces (neutral, happy, or fearful) were presented together with relevant 33 monetary feedback. Behavioral modeling was combined with fMRI data (N=27) to explore the 34 neurocomputational bases of learning relevant and irrelevant information. In two separate studies, 35 individuals with high trait anxiety showed increased avoidance of objects previously paired with the 36 combination of neutral monetary feedback and fearful faces (but not neutral or happy faces). 37 Behavioral modeling revealed that high anxiety increased the integration of fearful faces during 38 feedback learning, and fMRI results (regarded as provisional, due to a relatively small sample size) 39 further showed that variance in the prediction error signal - uniquely accounted for by fearful faces -40 correlated more strongly with activity in the right dorsolateral prefrontal cortex for more anxious 41 individuals. Behavioral and neuronal dissociations indicated that the threat-related distractors did 42 not simply disrupt learning processes. By showing that irrelevant threats exert long-lasting influences 43 on behavior, our results extend previous research that separately showed that anxiety increases 44 learning from aversive feedbacks and distractibility by threat-related information. Our behavioral 45 results, combined with the proposed neurocomputational mechanism, may help explain how 46 increased exposure to irrelevant affective information contributes to the acquisition of maladaptive 47 behaviors in more anxious individuals.

48

49

50 Significance statement

51 In modern-day society people are increasingly exposed to various types of irrelevant information, e.g. 52 intruding social media announcements. Yet, the neurocomputational mechanisms influenced by 53 irrelevant information during learning, and their interactions with increasingly distracted personality 54 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 55 56 irrelevant threat-related information (fearful faces) and inter-individual anxiety levels. Functional 57 neuroimaging (fMRI) show provisional evidence for an interaction between anxiety levels and the 58 coupling between activity in the dorsolateral prefrontal cortex and learning signals specifically 59 elicited by fearful faces. Our study reveals how irrelevant threat-related information may become 60 entrenched in the anxious psyche and contribute to long-lasting abnormal behaviors.

61

63 Introduction

In modern-day society people are increasingly exposed to emotionally loaded information that is 64 65 irrelevant for on-going and prospective behaviors (e.g. via online news and social media). Moreover, 66 efficient everyday learning requires the ability to ignore peripheral information that is not indicative 67 of, but presented in the vicinity of, actual performance feedback (e.g. intrusive social media 68 notifications). The ability to filter out irrelevant information is therefore important for an individual's 69 everyday function and well-being, even when not experienced first-hand. For example, media 70 exposure to disasters and violence relate to negative psychological outcomes (Holman et al., 2014; 71 Hopwood and Schutte, 2017), and information regarding potential threats, obtained via social 72 interactions, may induce maladaptive behaviors (Atlas, 2019; Lindstrom et al., 2019). Finally, 73 distracted learning has detrimental effects on learning performance in general (for a review, see 74 Schmidt, 2020). Surprisingly, the neurocomputational mechanisms influenced by affective irrelevant 75 information during learning, and how these interact with personality types that are more easily 76 distracted by affective information, are largely unknown.

Threat-related distractors attract attention and disrupt on-going 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. First, the null-hypothesis that affective distractors has no impact on the learning. Second, affective distractors disrupt the learning, i.e. learning performance in conditions with affective distractors should be reduced. Finally, 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.

92 To explore the neuronal correlates, our a priori analyses focused on the dorsolateral prefrontal 93 cortex (DLPFC) given that it has been implicated in attentional selection, such that the DLPFC is 94 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 95 al. (2010) first trained participants to attend either neutral or fearful faces, and reported increased 96 97 activity in the DLPFC when the attuned stimulus types were subsequently presented as distractors in 98 a different task. Second, converging evidence suggests that aberrant prediction error encoding in the 99 right DLPFC is involved in the acquisition of irrelevant associations (Corlett et al., 2007, 2016), with 100 the prediction error being the mismatch between an experienced and a predicted outcome (Sutton 101 and Barto, 1998). Accordingly, some studies report that prediction error encoding in the R DLPFC 102 correlated with an individual's tendency to learn associations in conditions that normally prevent the 103 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 104 105 events that should normally be ignored, eventually leading to the formation of maladaptive beliefs 106 and behaviors. Following reviewer suggestions, we also performed post-hoc analyses to elucidate 107 potential roles for the amygdala. This is relevant because the amygdala is activated by emotional 108 distractors (for a review, see Carretie, 2014a), plays a role in emotional learning (for a review, see 109 Phelps, 2006), and has been implicated in encoding prediction errors (Averbeck and Costa, 2017; 110 Aberg et al., 2020b). Additionally, amygdala activation during aversive learning and the presentation 111 of irrelevant distractors has been correlated with differences in anxiety levels (for reviews, see 112 Bishop et al., 2004; Lissek et al., 2005; Bishop, 2007; Aupperle and Paulus, 2010; Duval et al., 2015).

JNeurosci Accepted Manuscript

114 Methods and Materials

115 *Participants*

After having provided written consent according to the ethical regulations of the Weizmann Institute of Science, fifty-one 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 neurological disorders. The study was performed in accordance with the Declaration of Helsinki.

121 To ensure sufficient power regarding the behavioral effects in the fMRI study, a power analysis was 122 conducted using data from the behavioral pilot study. This analysis showed that 16 participants are 123 required to detect a one-tailed Pearson correlation coefficient of 0.548 (as obtained in the pilot 124 study) with a power (1- β) of 0.8 and error probability (α) of 0.05. However, because 16 participants 125 are not sufficient to detect inter-individual differences in fMRI activation, we recruited additional 126 participants to be more in-line with previous fMRI studies that investigated fMRI activation as a 127 function of trait anxiety in learning and decision making tasks, e.g. n=31 (Browning et al., 2015b), 128 n=32 (Bijsterbosch et al., 2015), n=25 (Xu et al., 2013), n=30 (Fung et al., 2019), and n=28 (Aberg et 129 al., 2022).

130 Two participants frequently fell asleep in the MRI scanner (as indicated by frequently missed trials 131 and post-task interviews). One participant did not perform the task satisfactorily (they pressed the 132 same button in all trials of a block), and one participant displayed excessive movement in all three 133 blocks of learning (as indicated by translational movements in a direction larger than the relevant 134 voxel dimension; Wylie et al., 2014). Therefore, data from 27 participants were included in the subsequent analyses of fMRI data (20 females; average age ± STD: 25.667 ± 4.961), while data from 135 136 20 different participants were included in the behavioral pilot study (11 females; average age ± STD: 137 27.350 ± 4.171). Trait anxiety was estimated using the State-Trait Anxiety Inventory (Spielberger et 138 al., 1983).

139 Experimental design and statistical analyses

140 Reinforcement learning task with distracting emotional faces

141 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 (Fig. 1A). 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 Fig. 1B. If no response was made within 2.5s 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 3s and a maximum duration of 10s. 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.

152 The different feedback types provided in the experiment are shown in Fig. 1C (pilot study) and Fig. 2A 153 (fMRI study). To test for learning differences between appetitive and aversive conditions the Correct 154 and Incorrect feedbacks were respectively +1 回 (a gain of one shekel) or 0 回 (no shekel gained) in a 155 Gain condition, while in a Loss condition the Correct and Incorrect feedbacks were respectively O₪ 156 (no shekel lost) or -1回 (one shekel lost). The accumulated sum of shekels corresponded to a 157 monetary bonus provided at the end of the experiment. To assess the impact of affective distractors 158 on associative learning, the numerical feedbacks were superimposed on fearful, neutral, or happy faces (Fig. 1C; Fig. 2A). In Affirmative pairs, the facial expression was matched with the feedback type 159 160 (e.g. a positive face was presented together with Correct feedback), while in Contradictory pairs the 161 contingencies were reversed (e.g. a positive faces was presented with Incorrect feedback). Please 162 observe that there were slight differences in the different feedback types presented in the pilot and 163 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.

167

Figure 1 around here

168 The learning task was divided into three separate blocks, each consisting of four (pilot) or six (fMRI 169 study) different types of pairs: Affirmative Gain, Affirmative Loss, Contradictory Gain, and 170 Contradictory Loss (as well as Neutral Gain and Neutral Loss for the fMRI study). In total, 12/18 171 different pairs of objects were used and participants performed 120 trials per block (30 trials per pair 172 in the pilot, and 20 trials per pair in the fMRI study) for a total of 360 trials. Participants were allowed 173 a break between each block. Pairs were presented in an interleaved fashion, such that each pair was 174 presented once before any other pair was repeated. Moreover, the object pairs were randomly 175 assigned to a condition for each participant, and each object was presented equally many times to 176 the left and to the right. Finally, no facial identity was repeated until all facial identities has been 177 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.

184 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 188 were one-tailed when testing directed predictions (i.e. positive or negative correlations), while two-189 tailed tests were used when no direction was predicted. The Bonferroni-correction for multiple 190 comparisons were applied where required.

191 Categorization task

To provide a functional localization of the right dorsolateral prefrontal cortex (DLPFC) and the 192 193 amygdala, participants performed a categorization task prior to the learning task. This task was 194 inspired by a previous task in which participants categorized attended neutral and fearful faces, and 195 which showed increased DLPFC activation for neutral (vs. fearful) faces for participants that had been 196 previously been attuned to fearful faces (as compared participants that had been attuned to neutral 197 faces; see Browning et al. (2010), Figure 4B: Face Attended condition). Furthermore, the amygdala is robustly engaged by faces, suggesting it should be activated more strongly by faces than by numbers 198 199 (Todorov, 2012). We selected a task in which participants attended the faces, rather than presenting 200 them as distractors, because we wanted to prevent any task-related perceived difficulty to confound 201 the results. For example, a differential brain activity between distracting fearful and neutral faces 202 could be wrongly attributed to increased task-difficulty caused by, for example, a disruptive 203 attentional bias towards 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. 204

205 Task description

Stimuli were classified as negative, neutral, or positive (Fig. 4A). 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 fifteen blocks of six 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 trialprogressions are shown in Fig. 4A. The ITI durations were drawn from a truncated exponential distribution; Dale, 1999), with an average duration of 3s and a maximum duration of 10s. 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.

218 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 inter-individual factors (Vul et al., 2009).

224 Stimulus selection

225 **Objects**

Eighteen different pairs of objects were created from a colored version of the Snoddgrass and Vanderbilt object data set, 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.

230

Table 1 around here

231 Faces

Fearful, neutral, and happy faces from fifteen different identities (7 males and 8 females) were selected from the Karolinska Directed Emotional Faces (KDEF) data set (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 seven male and eight 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.

243 Behavioral modeling

244 Q-learning

Following standard reinforcement learning theory, each object *i* in a pair was assigned an expected value Q_i 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 :

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

$$\delta(t)_i = \varphi - Q(t)_i$$

250 The probability of selecting object i in a given trial t can be estimated by a soft-max choice probability

- 251 function (Sutton and Barto, 1998):
- 252 $p(t)_{i} = e^{Q(t)_{i}\cdot\beta} / (e^{Q(t)_{i}\cdot\beta} + e^{Q(t)_{j}\cdot\beta})$

253 The β parameter estimates the trade-off between exploration and exploration / randomness of 254 choice.

255 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, et

- 260 different types of feedback.
- 261 In the '12 ϕ ' model, 12 different ϕ 's were fitted: one ϕ for each type of face for +1 \mathbb{D} feedback (3 ϕ 's),
- 262 $-1\square$ feedback (3 φ 's), $0\square$ feedback in Gain pairs (3 φ 's), and $0\square$ feedback in Loss pairs (3 φ 's).

263 In the '9 ϕ ' model, 9 ϕ 's were fitted: one ϕ for each type of face for +1 \mathbb{D} feedback (3 ϕ 's), -1 \mathbb{D}

feedback (3 φ 's), and for $0\square$ feedback across Gain and Loss pairs (3 φ 's).

265 In the '6 ϕ_0 ' model, 8 ϕ 's were fitted: one ϕ collapsed across faces for +1 \square feedback (1 ϕ) and -1 \square

- 266 feedback (1 ϕ), and one ϕ for each type of face separately for $0\square$ feedback in Gain (3 ϕ 's) and Loss 267 pairs (3 ϕ 's).
- 268 In the ' $3\phi_0$ ' model, 5 ϕ 's were fitted: one ϕ collapsed across faces for +1 \mathbb{D} feedback (1 ϕ) and -1 \mathbb{D}
- 269 feedback (1 φ), and one φ for each type of face for 0 \mathbb{D} collapsed across Gain and Loss pairs (3 φ 's).
- 270 In the ' ϕ_{OFF} , ϕ_{ONH} ' model, 4 ϕ 's were fitted: one ϕ collapsed across faces for +1 \mathbb{D} feedback (1 ϕ) and -

271 1[®] feedback (1 φ), one $φ_{FF}$ for fearful faces paired with 0[®] feedback, and one $φ_{NH}$ for neutral/happy 272 faces paired with 0[®] feedback.

273 In the ' ϕ_{0FFG} , ϕ_{0FFL} , ϕ_{0NH} ' model, 5 ϕ 's were fitted: one ϕ collapsed across faces for +1 \square feedback (1 274 ϕ) and -1 \square feedback (1 ϕ), one ϕ_{FFG} for fearful faces paired with 0 \square feedback in Gain pairs, one ϕ_{FFL} 275 for fearful faces paired with 0 \square feedback in Loss pairs, and one ϕ_{NH} for neutral/happy faces paired 276 with 0 \square feedback.

277 In the ' ϕ_{+1}, ϕ_{-1} ' model, 5 ϕ 's were fitted: one ϕ collapsed across faces for $0 \square$ feedback (1 ϕ), one ϕ 278 for happy faces paired with +1 \square feedback (1 ϕ), one ϕ for happy faces paired with -1 \square feedback (1 279 ϕ), one ϕ for fearful/neutral faces paired with +1 \square feedback (1 ϕ), and one ϕ for fearful/neutral 280 faces paired with -1 \square 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).

- 284 In the '3 ϕ , 3ϵ ' model, 3 ϕ 's were fitted: one ϕ for each numerical feedback type (-1, 0, +1; 3 ϕ 's), and
- 285 one ϵ for each emotional face type (fearful, neutral, happy; 3 ϵ 's).
- 286 In the '3 ϕ , ϵ_{FF} , ϵ_{NH} ' model, 3 ϕ 's were fitted: one ϕ for each numerical feedback type (-1, 0, +1; 3 ϕ 's),
- 287 one ε_{FF} for fearful faces (1 ε), and one ε_{NH} for neutral/happy faces combined (1 ε).
- 288 Finally, in the ' 3ϕ , ϵ_{0FF} , ϵ_{0NH} ' model, 3 ϕ 's were fitted: one ϕ for each numerical feedback type (-1, 0,
- 289 +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 $0\square$ feedback (1 ε).

291 Model fitting and model selection procedures

292 For each model, the free parameters were fitted individually to each participant's learning behavior

293 by minimizing the negative log-likelihood estimate:

$$LLE = -\ln(\prod_{1}^{n} p(t)_{i})$$

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:

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

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).

303 Model simulations

304 Two different model-simulations of behavior were performed to validate the most parsimonious305 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-ofinterest, 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-ofinterest, 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 if 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.

317 **MRI Data**

318 Image Acquisition

319 MRI images were acquired using a 3T whole body MRI scanner (Prisma, Siemens, Germany) with 320 a 20-channel head coil. Standard structural images were acquired with a T1 weighted 3D sequence 321 (MPRAGE, Repetition time (TR)/Inversion delay time (TI)/Echo time (TE)=2300/900/2.32 ms, flip 322 angle=8 degrees, voxel dimensions=0.9 mm isotropic, 192 slices). Functional images were acquired 323 with a susceptibility weighted EPI sequence (TR=2000, TE=30 ms, flip angle=75 degrees, voxel dimensions=3x3x3.5 mm, 32 slices). The phase encoding direction was anterior-posterior, the slice 324 325 order was all even (2 to 32) followed by all odd (1 to 31), with a 0% distance factor. No acceleration 326 technique was applied. The MRI scanner was stopped between each block of the learning task (each block lasted~15 minutes), while the functional localizer task lasted ~7 minutes. 327

328 Preprocessing

329	Functional MRI data were preprocessed and then analyzed using the general linear model (GLM) for
330	event-related designs in SPM12 (Welcome Department of Imaging Neuroscience, London, UK;
331	http://www.fil.ion.ucl.ac.uk/spm). During preprocessing, all functional volumes were realigned to the
332	mean image (with auto-masking applied), co-registered to the structural T1 image, corrected for slice
333	timing, resampled to 2x2x2 mm voxel size (upsampling of the voxel size to these dimensions has
334	been suggested to increase the sensitivity of fMRI analyses; Hopfinger et al., 2000), normalized to the
335	MNI EPI-template, and smoothed using a 6 mm FWHM Gaussian kernel. Please observe that the
336	resampling of voxels is mainly relevant for ROI identification in the functional localizer task.

337 *First-level analyses*

338 General procedure

At the first level, individual event-types, e.g. feedbacks, stimuli, or button presses (depending on task, see below), were modelled by a standard synthetic hemodynamic response function (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 twelve corresponding squared items; Friston et al., 1996). Statistical analyses were performed on a voxel-wise basis across the whole brain.

345 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, twelve different event-types were created, together with a regressor of no interest which included the onset of trials in which no response was made.

350 Region of interest (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 pre-defined ROIs for Broadmann areas 9, 46, and the middle frontal gyrus (Tzourio-Mazoyer et al., 2002; Maldjian et al., 2003; Maldjian et al., 2004). For the post-hoc analysis regarding amygdala involvement, an initial amygdala mask was obtained by including all available amygdala sub-regions provided by the SPM Anatomy toolbox (Eickhoff et al., 2005).

359 Statistical Analyses

To localize the R DLPFC, we contrasted the BOLD signal evoked by neutral and fearful faces, while the 360 361 amygdala was localized by contrasting BOLD signal evoked by faces and numbers. Significant 362 differential activations within the initial R DLPFC and amygdala masks were tested via t-tests 363 implemented in SPM using an initial search threshold of p=0.001, and small volume correction (SVC) 364 using a threshold of p<0.05 Family-Wise Error rate (FWE) to correct for multiple comparisons. For 365 display purposes and follow-up analyses (e.g. correlation with individual anxiety levels), beta 366 parameter estimates were extracted and averaged from all voxels within significant clusters of 367 activation.

368 First-level analysis of the learning task

An event-related fMRI design was created with three different event-types (stimulus onset, response 369 370 onset, and feedback onset) for each of four trial types (Gain Correct feedback, Gain Incorrect 371 feedback, Loss Correct feedback, and Loss Incorrect feedback). Besides these twelve event-types for 372 each of three blocks, trials in which no response was made during the picture display were included 373 as a regressor of no interest. To isolate the contribution of the distractors to the prediction error 374 signal, the prediction error term for the selected model (δ Full) was separated into two parts (see 375 Wittman et al., 2008; Eldar and Niv, 2015, for similar procedures). In brief, a 'Basic' prediction error 376 term (δ Basic) accounted for variance in the prediction error signal when there is no differential 377 modulation by distractor type, i.e. the values of parameters-of-interest are set to be equal. Next, a 378 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 was orthogonalized with respect to the values of δ Basic (Mumford et al., 2015).

384 Regions of interest (ROIs)

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

386 Statistical Analyses

387 Correlations between prediction errors and BOLD signal in the ROIs was tested using a ROI 388 approach were the average beta parameter estimates for each type of prediction error (δ_{Basic} , δ_{Boost}) 389 were extracted from all voxels within the ROIs. These beta parameters were then entered into two 390 separate repeated measures ANOVAs (i.e. one for each prediction error type) with factors Gain/Loss 391 (Gain, Loss pairs) and Feedback (Correct, Incorrect), and Trait anxiety as continuous covariate. 392 Follow-up analyses were conducted using paired *t*-tests and Pearson correlations.

393 Results

394 Behavior

395 Behavioral pilot study

An initial pilot study was conducted with twenty 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 Fig. 1D,E. Trait anxiety scores correlated negatively with the average performance only in the Contradictory Loss condition (trials 16-20: Pearson's r=-0.548, 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.

407

Table 2 around here

408 fMRI study

409 The behavioral paradigm of the fMRI study was similar to the one used in the pilot study, with 410 the main addition of a control condition used to normalize learning performance by subtracting the 411 learning performance in the absence of affective distractors (i.e. with neutral faces; Fig. 2A). Learning 412 curves are shown in Fig. 2B, and normalized average learning performances are shown in Fig. 2C-F. 413 First, we replicated the main result of the pilot study, namely a negative correlation between 414 trait anxiety and learning performance in the Contradictory (vs. Neutral) Loss condition [Fig. 1F; r=-415 0.394, p=0.021, one-tailed Pearson correlation]. Because anxiety increases the tendency to display 416 behavioral switching following aversive feedbacks (Aberg and Paz, 2022), we tested whether anxious 417 participants displayed a reduced proportion of win-stay decisions for the Correct feedback in the 418 Contradictory Loss condition (i.e. because the fearful faces were paired with the neutral O₪ 419 feedback). As predicted the proportion of win-stay decisions correlated negatively with trait anxiety 420 in Contradictory (vs. Neutral) Loss pairs [Fig. 2J; r=-0.420, p=0.015, one-tailed Pearson-correlation]. A 421 similar trend was observed in the Contradictory Loss condition of the pilot study [r=-0.359, p=0.060, 422 one-tailed Pearson-correlation]. 423 In contrast to the pilot study, we observed a positive correlation between anxiety and learning performance in Affirmative (vs. Neutral) Gain pairs [Figure 2C; r=0.640, p=0.001, two-tailed Pearson 424

425 correlation, p-value was corrected for three unplanned comparisons]. Notably, in the fMRI study (but

426 not the pilot study), the neutral 0[™] feedback in Affirmative Gain pairs was presented together with a

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. 2G]. Finally, trait anxiety did not correlate significantly with learning performance in the remaining two conditions [Fig. 2D,E, all uncorrected p-values > 0.05, two-tailed Pearson-correlations; Table 3],

fearful face, therefore providing another opportunity to test whether fearful faces increase the

433 nor with the behavioral switching for neutral 0[™] feedbacks paired with happy faces [Fig. 2H,I; all

434 uncorrected p-values > 0.05, two-tailed Pearson-correlations; Table 4].

435

427

Table 3 around here

- 436 Table 4 around here
- In summary, the behavioral results from the pilot and the fMRI study suggest that distracting
 fearful faces increases the averseness of the neutral OID 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.

442 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 Methods).

447 A fixed-effect analysis showed an overall lower BIC for the ' ϕ_{0FF} , ϕ_{0NH} ' model (Fig. 3A), indicating 448 a better fit to behavior on average. Additionally, a random-effects analysis indicated a protected 449 exceedance probability of 1.0 for the same model (Inset, Fig. 3A), a result which suggests that the 450 selected model is the most likely model to generate the observed behavior (Stephan et al., 2009). 451 Together, these two complementary ways of comparing model indicate the ' ϕ_{0FF} , ϕ_{0NH} ' model as 452 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 estimate the subjective value of +1 \square and -1 \square feedbacks and are independent of the distracting faces, while ϕ_{0FF} and ϕ_{0NH} respectively estimates the subjective value of the neutral 0 \square feedback paired with fearful (ϕ_{0FF}) and neutral/happy (ϕ_{0NH}) faces. Average fitted model parameters for all models are displayed in Table 5.

459

Table 5 around here

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 Fig. 3B. More importantly, the model successfully reproduced the behavioral effects-of-interest (Fig. 3C-F, c.f. Fig. 2C-F).

One possible explanation for the behavioral results is that the interaction between fearful faces and anxiety reduces the subjective value of the neutral $0\square$ feedback. Corroborating this notion, trait anxiety correlated negatively with the difference in the fitted subjective values of the $0\square$ feedback paired with fearful and neutral/happy faces [φ_{0FF} - φ_{0NH} , r=-0.467, p=0.007, one-tailed Pearsoncorrelation; Figure 2G]. 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].

472 Additional model simulations were performed to ensure that the impact of the interaction 473 between distractor type and anxiety on learning can actually be attributed to the differential 474 subjective values of ϕ_{OFF} and ϕ_{ONH} . In these simulations, all model parameters are initially set to the average values of the fitted parameters across participants' values (i.e. α =0.25, ϕ_{+1} =-0.25, ϕ_{0FF} =0.35, 475 476 φ_{0NH} =0.35, φ_{+1} =0.80, and β =0.15). The values are held constant, except for the values of φ_{0FF} and 477 φ_{ONH} , which are gradually decreased and increased, respectively, in order to simulate the modulation 478 by trait anxiety (Fig. 3G). Simulated performance improvements are calculated for all conditions, and 479 visualized as a comparison between conditions (see Methods for further details). As would be 480 expected, decreases in the difference between ϕ_{OFF} and ϕ_{ONH} , improved performance in Affirmative 481 Gain pairs (relative Contradictory and Neutral Gain pairs; Fig. 2H) while reducing performance in 482 Contradictory Loss pairs (relative Affirmative and Neutral Loss pairs; Fig. 2I).

483 Finally, to illustrate the robustness of the main modeling result, we demonstrate that the 484 negative correlation between trait anxiety and the relative difference between fitted 0n feedback 485 values for fearful (vs. neutral and happy) faces are present across different behavioral models. First, 486 the ' ϕ_{OFFG} , ϕ_{OFFG} , ϕ_{ONH} ' model differs from the most parsimonious model by estimating separate ϕ_{OFF} 's 487 in Gain and Loss pairs (i.e. ϕ_{0FF} was separated into two parameters, ϕ_{0FFL} and ϕ_{0FFG}). Trait anxiety 488 correlated negatively with the difference between ϕ_{OFFL} and ϕ_{NH} [r=-0.426, p=0.013, one-tailed 489 Pearson correlation; Fig. 3J], and with the difference between ϕ_{OFFG} and ϕ_{NH} [r=-0.448, p=0.010, one-490 tailed Pearson correlation; Fig. 3K], with a positive correlation between ϕ_{OFFL} and ϕ_{OFFG} [r=0.412, 491 p=0.016, one-tailed Pearson correlation]. Second, the ${}^\prime 3\phi_0{}^\prime$ model differs from the most 492 parsimonious model by separating the ϕ_{ONH} term into two terms, one term corresponding to the 493 combination of $0\mathbb{D}$ feedback paired with neutral faces (ϕ_{0N}) and one term for happy faces (ϕ_{0H}). Trait 494 anxiety correlated negatively with the difference between ϕ_{OFF} and ϕ_{ON} [r=-0.535, p=0.002, one-tailed 495 Pearson-correlation; Fig. 3L], as well as for the difference between ϕ_{OFF} and ϕ_{OH} [r=-0.335, p=0.044, 496 one-tailed Pearson correlation; Fig. 3M], and ϕ_{ON} and ϕ_{OH} were positively correlated [r=0.816, 497 p<0.001].

498 In summary, the selected ' φ_{0FF} , φ_{0NH} ' model provides the most parsimonious fit to behavior and 499 provides a plausible and robust explanation for how anxiety interacts with threat-related distractors 500 to modulate learning performance, namely via a reduced subjective value of neutral OD feedbacks.

501 Functional neuroimaging

502 A priori, we hypothesized that atypical prediction error encoding in the R DLPFC, caused by the 503 presence of threat-related distractors, contributes to the learning bias displayed by more anxious 504 individuals. Based on reviewer suggestions, we also conducted a post-hoc analysis with focus on the 505 amygdala. To this end, we first used a separate task to functionally define the ROIs to be used when 506 analyzing the learning task. Notably, by selecting ROIs in a separate task, we avoid issues of double-507 dipping (Kriegeskorte et al., 2009), and by selecting ROIs based on group-level data, we minimize the 508 possibility of inflated effect sizes when analyzing inter-individual differences in brain activation (Vul 509 et al., 2009).

510 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 either negative, neutral, or positive (Fig. 4A). 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 (versus fearful) faces [peak voxel coordinate: x=46 y=44 z=18, T(25)=5.151, pFWE,SVC=0.013; one-tailed paired *t*-test, Fig. 4B,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. 4D; 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 (versus 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. 4E,F]. The collapsed activity within this bilateral ROI showed no correlation with trait anxiety [Fig. 4G; 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.

525 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 Methods). To assess their neuronal correlates, the beta 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 beta 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.

533 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 beta parameters for the δ_{Basic} term collapsed across the four feedback conditions for the R DLPFC ROI are shown in Fig. 5B.

538

Table 6 around here

539 This result shows that BOLD signal in the R DLPFC correlates significantly with the magnitude of 540 the 'Basic' prediction error signal.

541 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 x Gain/Loss [F(1, 25)=6.04, p=0.021, ANOVA], and Trait anxiety x Feedback [F(1, 25)=4.68, p=0.040, ANOVA], but no significant Trait anxiety x Gain/Loss x Feedback interaction [F(1,25)=0.62, p=0.438, ANOVA]. Individual beta parameters for the δ_{Boost} term in the four feedback conditions are shown in Fig. 5C. Importantly, the two a priori hypotheses were confirmed via a positive correlation 547 between trait anxiety and the beta parameters of δ_{Boost} in the Gain Incorrect feedback condition [r=0.468, p=0.007, one-tailed Pearson correlation; Fig. 5D], and a negative correlation in the Loss 548 Incorrect feedback condition [r=-0.697, p<0.001, one-tailed Pearson correlation; Fig. 5G]. By contrast, 549 trait anxiety did not correlate with the beta parameters of δ_{Boost} in the Gain Correct feedback 550 condition [r=-0.147, p=0.463, two-tailed Pearson correlation; Fig. 5E] nor in the Loss Incorrect 551 552 feedback condition [r=-0.107, p=0.594, two-tailed Pearson correlation; Fig. 5F]. Besides a significant Gain/Loss x Feedback interaction [F(1, 25)=5.74, p=0.024], no other effects or interactions are 553 554 significant (all p-values > 0.24; see Table 7 for a full ANOVA table).

555

Table 7 around here

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.

559 Prediction error coding in the Amygdala

560 As in the previous analysis, the beta parameter estimates corresponding to the two prediction

- 561 error terms, δ_{Boost} and δ_{Basic} , were extracted from all voxels within the amygdala ROI. The resulting
- 562 average beta parameters where entered into the same ANOVAs used for the R DLPFC ROI analysis.

563 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

- 566 anxiety levels is shown in Fig. 5I for visualization purposes.
- 567 Table 8 around here
- 568 Amygdala activity does not correlate with the prediction error 'Boost'

570	interactions [all p-values>.17, ANOVA; Table 9]. For visualization purposes, the beta parameters for
571	each condition are shown in Fig. 5J, and correlations with trait anxiety are shown in Fig. 5K-N.
572	Table 9 around here
573	In summary, no evidence supported a role for the amygdala in prediction error coding.
574	Whole-brain correlates of the 'Basic' prediction error
575	To validate our model-based fMRI procedure, we tested whether activity in the ventral
576	tegmental area (VTA), a region well known for its role in encoding different aspects of reward,
577	including prediction errors (D'Ardenne et al., 2008; Bromberg-Martin et al., 2010; Aberg et al., 2015;
578	Schultz, 2016; Aberg et al., 2020a), correlated with the $\delta_{\scriptscriptstyle Basic}$ term. This analysis was performed by
579	averaging the beta parameters related to the $\delta_{\mbox{\tiny Basic}}$ term for all voxels within a recently developed
580	probabilistic in vivo atlas of the VTA (Fig. 6A; Pauli, Nili, & Tyszka, 2018(Pauli et al., 2018)). Indeed,
581	the average beta parameters of this VTA ROI was significantly larger than 0.0 [mean (±SEM)=0.068
582	(±0.021), t(26)=3.291, p=0.001, one-tailed <i>t</i> -test, Fig. 6B]. Next, correlations with the δ_{Basic} term were
583	tested across the whole-brain using a FWE-corrected threshold of 0.05. A full list of regions
584	correlating with δ_{Basic} , surviving a threshold of a FWE-corrected threshold of 0.05, is reported in Table
585	10. In short, significant activation was observed in a midbrain region close to the previously used VTA
586	mask (Fig. 6C,D), in the dorsal anterior cingulate cortex (dACC)/dorsomedial prefrontal cortex
587	(dmPFC; Fig. 6D), in the bilateral striatum (Fig. 6E,F), and in the bilateral anterior insula (Fig 6G,H).
588	These regions have previously been implicated in the neuronal coding of prediction errors (Garrison
589	et al., 2013).
590	Table 10 around here

For the δ_{Boost} term, the repeated measures ANOVA revealed no significant main effects or

569

591

JNeurosci Accepted Manuscript

595 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 OID neutral feedback and irrelevant fearful faces, even though participants were explicitly instructed that the faces are unrelated to task performance. By showing that exposure to irrelevant affective information lingers and affect 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).

608 Importantly, the threat-related distractors did not simply disrupt the learning process, as would 609 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 610 respectively reduced or improved performance in conditions where this feedback combination 611 612 represented the Correct or the Incorrect outcome. In support, behavioral modeling further showed 613 that high anxiety was associated with a reduced subjective value of the neutral O
 feedback when 614 paired with fearful faces both when it signaled Correct and Incorrect outcomes (as compared to 615 happy and neutral faces). A third dissociation was observed in the fMRI data, with a stronger / 616 weaker coupling between the prediction error signal - uniquely accounted for by fearful faces - and R 617 DLPFC BOLD signal for feedbacks where anxious individuals showed increased / decreased learning 618 performance. Together, these results indicate that anxiety is associated with an increased integration 619 of irrelevant threat-related information during feedback processing (and not just disrupted learning 620 processes). From an evolutionary perspective, it makes sense that information related to potential 621 threats are integrated during learning, rather than disrupting it. However, this ability comes at the 622 cost of increased avoidance of beneficial situations in which a potential treat was occasionally 623 detected.

624 It has been suggested that anxiety disorders develop from abnormal learning processes, e.g. 625 amplified fear learning (Lissek et al., 2005) and over-generalization (i.e. the transfer of aversive 626 properties from a fear-conditioned neutral stimulus to other perceptually and conceptually similar 627 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 628 629 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 630 631 that contribute to the development of dysfunctional behaviors and mental illness. Based on the 632 present results, we propose that this includes the maladaptive formation of associations between 633 neutral stimuli/events and irrelevant threat-related information, as these may result in inappropriate 634 avoidance behaviors.

635 Anxious individuals showed increased integration of fearful faces with the neutral OD feedback, but 636 not with +1D and -1D feedbacks. One potential explanation for this result could be that anxious 637 individuals call upon additional, salient sources of information to resolve uncertain feedbacks. To 638 clarify, in the present study the -1₪ and +1₪ feedbacks always indicated the worst and best possible 639 outcomes, while the Op feedback signaled either a correct (in Loss conditions) or an incorrect (in 640 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), 641 642 the motivation to reduce uncertainty (Aberg et al., 2022), and distractibility by threat-related 643 information (Bar-Haim et al., 2007). Future research may profit from looking at how irrelevant salient 644 information guides the processing of uncertain feedback in high anxiety.

645 The present study used a between-subject design to study the interaction between anxiety and 646 threat-related distractors during learning. A complementary way to assess behavioral interactions 647 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 648 649 approach is beneficial because it could be used to combine a powerful within-subject design (i.e. 650 conditions with and without stress, or induced anxiety) with a between-subject design (e.g. trait 651 anxiety measures, or patient versus control groups). This approach may be particularly fruitful to 652 further research on findings that individuals with anxious predispositions respond differently in 653 stressful situations (Meijer, 2001; Indovina et al., 2011; Aberg and Paz, 2022).

654 In accord with previous studies (for a review, see Garrison et al., 2013), a number of brain 655 regions in the present study, including the ventral tegmental area, the striatum, anterior cingulate cortex, anterior insula, and the R DLPFC, encoded a 'basic' prediction error signal. However, only the 656 657 R DLPFC of anxious individuals correlated with additional variance in the prediction error signal that 658 was uniquely attributed to the fearful faces. This correlation was positive in conditions where anxiety 659 improved performance, but negative when anxiety show impaired performance. These results are in 660 accordance with previous research showing that the strength of neuronal prediction error encoding 661 correlates with the amount of learning (Schönberg et al., 2007; Aberg et al., 2015, 2016a), and 662 complement behavioral and physiological reports of links between personality traits and 663 reinforcement learning biases (Browning et al., 2015b; Aberg et al., 2016b; Aberg et al., 2017). These 664 results also bridge separate reports of an involvement of the DLPFC in attentional bias to threat 665 (Bishop, 2009), prediction error encoding (Fletcher et al., 2001; Corlett et al., 2004), and the 666 acquisition of irrelevant associations via altered prediction error encoding (Corlett and Fletcher, 667 2012, 2015). Indeed, similar and converging pieces of evidence support a theory in which aberrant 668 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 669 670 theory by suggesting that one source of 'aberrancy' stems from failures in suppressing attention to

<u>JNeurosci Accepted Manuscript</u>

irrelevant sources of threat-related information, i.e. these stimuli may grab attention and engagelearning processes like any other (relevant) stimulus, and particularly so with high anxiety.

673 By contrast, we did not observe any involvement of the amygdala in coding prediction errors, nor 674 any interaction with trait anxiety. Although the amygdala plays a prominent role in fear learning and 675 anxiety (Phelps, 2006; Duval et al., 2015; Tovote et al., 2015), only scarce evidence report 676 correlations between amygdala activity and prediction errors (McHugh et al., 2014; Meffert et al., 677 2015; Aberg et al., 2020b). One possibility is that the amygdala codes for other features related to 678 learning and the prediction error signal, such as surprise, sometimes defined as the unsigned 679 prediction error signal (Li et al., 2011; Klavir et al., 2013). Further, although the amygdala is activated 680 by affective distractors, and particularly so for more anxious individuals (Bishop et al., 2004; Bishop, 681 2009), it has to our best knowledge not been implicated in the learning of irrelevant information.

682 Limitations

683 Anxiety was estimated using the standard Spielberger's Trait-Anxiety Inventory (Spielberger et 684 al., 1983), which provides a gradual scale for the normal (sub-clinical) range of anxiety. Using a 685 continuous scale has the benefit of correlating behavior across a distribution of anxiety scores, rather 686 than just comparing performance across two somewhat arbitrarily divided populations (patients vs. 687 controls). Additionally, by studying anxiety within the normal range, we can determine how 688 maladaptive decisions are mediated by irrelevant distractors even in healthy individuals. Because 689 such maladaptive decision may have a huge impact on daily-life in all individuals, and definitely on 690 societies and industry, we actually believe that more studies should use gradual scales over non-691 clinical populations (Browning et al., 2015a; Fung et al., 2019; Gagne et al., 2020). That said, 692 Spielberger's Trait-Anxiety Inventory has been debated for its lack of convergent and discriminant 693 validity, suggesting that it estimates 'negative affectivity' rather than proneness to anxiety per-se 694 (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.

697 Our main behavioral results were replicated between two separate groups of participants (i.e. 698 negative correlations between trait anxiety and learning performance in the Contradictory Loss 699 condition was observed in both the pilot and in the fMRI study), and were replicated within another 700 condition in the fMRI study (i.e. the feedback combination of fearful faces + neutral 0₪ feedback 701 increased behavioral switching in both the Contradictory Loss and the Affirmative Gain condition). 702 Furthermore, these latter results were associated with two dissociations in the fMRI results, namely 703 opposite correlations with trait anxiety and the coupling between R DLPFC activity and the prediction 704 error signal associated with fearful faces. Being able to replicate results across- and within groups 705 speaks in favor of the robustness of our results.

706 Importantly, we would like to stress that our fMRI findings were obtained with a relatively small 707 sample size (N=27), and therefore needs to be regarded as provisional. In particular, although many 708 factors may contribute to the reliability of brain-behavior correlations in fMRI data, including 709 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., 710 711 20-30 participants), one cannot be confident that all of the regions appearing to correlate with 712 individual differences in behavior are reliable, or that other regions have not been missed 713 altogether." (Grady et al., 2021). Future studies should therefore expand on the issue and validate 714 the robustness of the present fMRI results.

715 *Conclusion*

In conclusion, the present study displays a learning bias for individuals with high trait anxiety caused by an entanglement between threat-related distractors and on-going 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

- 720 present study describes a new pathway for how threat-related information may become entrenched
- in the anxious psyche.

722

References

725 726	Aberg KC, Paz R (2022) Stress-induced avoidance in mood disorders. Nature human behaviour 6:915- 918.
727	Aberg KC, Doell KC, Schwartz S (2015) Hemispheric Asymmetries in Striatal Reward Responses Relate
728	to Approach-Avoidance Learning and Encoding of Positive-Negative Prediction Errors in
729	Dopaminergic Midbrain Regions. The Journal of neuroscience : the official journal of the
730	Society for Neuroscience 35:14491-14500.
731	Aberg KC. Doell KC. Schwartz S (2016a) The left hemisphere learns what is right: Hemispatial reward
732	learning depends on reinforcement learning processes in the contralateral hemisphere.
733	Neuropsychologia 89:1-13.
734	Aberg KC. Doell KC. Schwartz S (2016b) Linking Individual Learning Styles to Approach-Avoidance
735	Motivational Traits and Computational Aspects of Reinforcement Learning, PloS one 11.
736	Aberg KC. Muller J. Schwartz S (2017) Trial-by-Trial Modulation of Associative Memory Formation by
737	Reward Prediction Error and Reward Anticipationas Revealed by a Biologically Plausible
738	Computational Model. Front Hum Neurosci 11.
739	Aberg KC. Kramer EE. Schwartz S (2020a) Interplay between midbrain and dorsal anterior cingulate
740	regions arbitrates lingering reward effects on memory encoding. Nature communications
741	11:1829.
742	Aberg KC, Kramer EE, Schwartz S (2020b) Neurocomputational correlates of learned irrelevance in
743	humans. NeuroImage 213.
744	Aberg KC, Toren I, Paz R (2022) A neural and behavioral trade-off between value and uncertainty
745	underlies exploratory decisions in normative anxiety. Molecular psychiatry 27:1573-1587.
746	Atlas LY (2019) How instructions shape aversive legarning: higher order knowledge, reversal learning,
747	and the role of the amygdala. Curr Opin Behav Sci 26:121-129.
748	Aupperle RL, Paulus MP (2010) Neural systems underlying approach and avoidance in anxiety
749	disorders. Dialogues in clinical neuroscience 12:517-531.
750	Averbeck BB, Costa VD (2017) Motivational neural circuits underlying reinforcement learning. Nature
751	neuroscience 20:505-512.
752	Balsamo M, Romanelli R, Innamorati M, Ciccarese G, Carlucci L, Saggino A (2013) The State-Trait
753	Anxiety Inventory: Shadows and Lights on its Construct Validity. J Psychopathol Behav
754	35:475-486.
755	Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IMH (2007) Threat-related
756	attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychological
757	bulletin 133:1-24.
758	Bijsterbosch J, Smith S, Bishop SJ (2015) Functional Connectivity under Anticipation of Shock:
759	Correlates of Trait Anxious Affect versus Induced Anxiety. Journal of cognitive neuroscience
760	27:1840-1853.
761	Bishop S, Duncan J, Brett M, Lawrence AD (2004) Prefrontal cortical function and anxiety: controlling
762	attention to threat-related stimuli. Nature neuroscience 7:184-188.
763	Bishop SJ (2007) Neurocognitive mechanisms of anxiety: an integrative account. Trends in cognitive
764	sciences 11:307-316.
765	Bishop SJ (2009) Trait anxiety and impoverished prefrontal control of attention. Nature neuroscience
766	12:92-98.
767	Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Dopamine in motivational control:
768	rewarding, aversive, and alerting. Neuron 68:815-834.
769	Browning M, Holmes EA, Murphy SE, Goodwin GM, Harmer CJ (2010) Lateral prefrontal cortex
770	mediates the cognitive modification of attentional bias. Biological psychiatry 67:919-925.
771	Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015a) Anxious individuals have difficulty
772	learning the causal statistics of aversive environments. Nature neuroscience 18:590-+.
/73	Browning M, Benrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015b) Anxious individuals have difficulty
//4	learning the causal statistics of aversive environments. Nature neuroscience 18:590-596.

775	Calvo MG, Lundqvist D (2008) Facial expressions of emotion (KDEF): identification under different
776	display-duration conditions. Behavior research methods 40:109-115.
777	Carretie L (2014) Exogenous (automatic) attention to emotional stimuli: a review. Cogn Affect Behav
778	Ne 14:1228-1258.
779	Chambers JA, Power KG, Durham RC (2004) The relationship between trait vulnerability and anxiety
780	and depressive diagnoses at long-term follow-up of Generalized Anxiety Disorder. J Anxiety
781	Disord 18:587-607.
782	Cisler JM, Koster EH (2010) Mechanisms of attentional biases towards threat in anxiety disorders: An
783	integrative review. Clinical psychology review 30:203-216.
784	Clark LA, Watson D, Mineka S (1994) Temperament, Personality, and the Mood and Anxiety
785	Disorders. J Abnorm Psychol 103:103-116.
786	Corlett PR, Fletcher PC (2012) The neurobiology of schizotypy: fronto-striatal prediction error signal
787	correlates with delusion-like beliefs in healthy people. Neuropsychologia 50:3612-3620.
788	Corlett PR, Fletcher PC (2015) Delusions and prediction error: clarifying the roles of behavioural and
789	brain responses. Cognitive neuropsychiatry 20:95-105.
790	Corlett PR, Honey GD, Fletcher PC (2007) From prediction error to psychosis: ketamine as a
791	pharmacological model of delusions. Journal of psychopharmacology 21:238-252.
792	Corlett PR, Honey GD, Fletcher PC (2016) Prediction error, ketamine and psychosis: An updated
793	model. Journal of psychopharmacology 30:1145-1155.
794	Corlett PR, Aitken MR, Dickinson A, Shanks DR, Honey GD, Honey RA, Robbins TW, Bullmore ET,
795	Fletcher PC (2004) Prediction error during retrospective revaluation of causal associations in
796	humans: fMRI evidence in favor of an associative model of learning. Neuron 44:877-888.
797	D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic
798	signals in the human ventral tegmental area. Science 319:1264-1267.
799	Dale AM (1999) Optimal experimental design for event-related fMRI. Human brain mapping 8:109-
800	114.
801	Duval ER, Javanbakht A, Liberzon I (2015) Neural circuits in anxiety and stress disorders: a focused
802	review. Ther Clin Risk Manag 11:115-126.
803	Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM
804	toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data.
805	NeuroImage 25:1325-1335.
806	Eldar E, Niv Y (2015) Interaction between emotional state and learning underlies mood instability.
807	Nature communications 6:6149.
808	Fales CL, Barch DM, Rundle MM, Mintun MA, Snyder AZ, Cohen JD, Mathews J, Sheline YI (2008)
809	Altered emotional interference processing in affective and cognitive-control brain circuitry in
810	major depression. Biological psychiatry 63:377-384.
811	Fletcher PC, Anderson JM, Shanks DR, Honey R, Carpenter TA, Donovan T, Papadakis N, Bullmore ET
812	(2001) Responses of human frontal cortex to surprising events are predicted by formal
813	associative learning theory. Nature neuroscience 4:1043-1048.
814	Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI
815	time-series. Magnetic resonance in medicine 35:346-355.
816	Fung BJ, Qi S, Hassabis D, Daw N, Mobbs D (2019) Slow escape decisions are swayed by trait anxiety.
817	Nat Hum Behav 3:702-708.
818	Gagne C, Zika O, Dayan P, Bishop SJ (2020) Impaired adaptation of learning to contingency volatility
819	in internalizing psychopathology. Elife 9.
820	Garrison J, Erdeniz B, Done J (2013) Prediction error in reinforcement learning: A meta-analysis of
821	neuroimaging studies. Neurosci Biobehav R 37:1297-1310.
822	Grady CL, Rieck JR, Nichol D, Rodrigue KM, Kennedy KM (2021) Influence of sample size and analytic
823	approach on stability and interpretation of brain-behavior correlations in task-related fMRI
824	data. Human brain mapping 42:204-219.
825	Grillon C (2002) Startle reactivity and anxiety disorders: Aversive conditioning, context, and
826	neurobiology. Biological psychiatry 52:958-975.

827	Grupe DW, Nitschke JB (2013) Uncertainty and anticipation in anxiety: an integrated neurobiological
828	and psychological perspective. Nature reviews Neuroscience 14:488-501.
829	Hartley CA, Phelps EA (2012) Anxiety and decision-making. Biological psychiatry 72:113-118.
830	Holman EA, Garfin DR, Silver RC (2014) Media's role in broadcasting acute stress following the Boston
831	Marathon bombings. Proceedings of the National Academy of Sciences of the United States
832	of America 111:93-98.
833	Hopfinger JB, Buchel C, Holmes AP, Friston KJ (2000) A study of analysis parameters that influence
834	the sensitivity of event-related fMRI analyses. NeuroImage 11:326-333.
835	Hopwood TL, Schutte NS (2017) Psychological Outcomes in Reaction to Media Exposure to Disasters
836	and Large-Scale Violence: A Meta-Analysis. Psychol Violence 7:316-327.
837	Indovina I, Robbins TW, Nunez-Elizalde AO, Dunn BD, Bishop SJ (2011) Fear-Conditioning Mechanisms
838	Associated with Trait Vulnerability to Anxiety in Humans. Neuron 69:563-571.
839	Klavir O, Genud-Gabai R, Paz R (2013) Functional Connectivity between Amygdala and Cingulate
840	Cortex for Adaptive Aversive Learning. Neuron 80:1290-1300.
841	Kotov R, Gamez W, Schmidt F, Watson D (2010) Linking "Big" Personality Traits to Anxiety,
842	Depressive, and Substance Use Disorders: A Meta-Analysis. Psychological bulletin 136:768-
843	821.
844	Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems
845	neuroscience: the dangers of double dipping. Nature neuroscience 12:535-540.
846	Li J. Schiller D. Schoenbaum G. Phelps EA. Daw ND (2011) Differential roles of human striatum and
847	amygdala in associative learning. Nature neuroscience 14:1250-1252.
848	Lindstrom B. Golkar A. Jangard S. Tobler PN. Olsson A (2019) Social threat learning transfers to
849	decision making in humans. Proceedings of the National Academy of Sciences of the United
850	States of America 116:4732-4737.
851	Lissek S. Kaczkurkin AN. Rabin S. Geraci M. Pine DS. Grillon C (2014) Generalized Anxiety Disorder Is
852	Associated With Overgeneralization of Classically Conditioned Fear. Biological psychiatry
853	75:909-915.
854	Lissek S. Powers AS. McClure EB. Phelps EA. Woldehawariat G. Grillon C. Pine DS (2005) Classical fear
855	conditioning in the anxiety disorders: a meta-analysis. Behaviour research and therapy
856	43:1391-1424.
857	Lundguist D, Flykt A, Öhman A (1998) The Karolinska Directed Emotional Faces—KDEF [CD-ROM]. In:
858	(Department of Clinical Neuroscience Ps, Karolinska Institutet, Stockholm, Sweden, ed).
859	Maldjian JA, Laurienti PJ, Burdette JH (2004) Precentral gyrus discrepancy in electronic versions of
860	the Talairach atlas. NeuroImage 21:450-455.
861	Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and
862	cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage 19:1233-1239.
863	McHugh SB, Barkus C, Huber A, Capitao L, Lima J, Lowry JP, Bannerman DM (2014) Aversive
864	prediction error signals in the amygdala. The Journal of neuroscience : the official journal of
865	the Society for Neuroscience 34:9024-9033.
866	Meffert H, Brislin SJ, White SF, Blair JR (2015) Prediction errors to emotional expressions: the roles of
867	the amygdala in social referencing. Social cognitive and affective neuroscience 10:537-544.
868	Meijer J (2001) Stress in the relation between trait and state anxiety. Psychol Rep 88:947-964.
869	Mumford JA, Poline JB, Poldrack RA (2015) Orthogonalization of regressors in FMRI models. PloS one
870	10:e0126255.
871	Ohman A (1986) Face the Beast and Fear the Face - Animal and Social Fears as Prototypes for
872	Evolutionary Analyses of Emotion. Psychophysiology 23:123-145.
873	Palminteri S, Wyart V, Koechlin E (2017) The Importance of Falsification in Computational Cognitive
874	Modeling. Trends in cognitive sciences 21:425-433.
875	Pauli WM, Nili AN, Tyszka JM (2018) A high-resolution probabilistic in vivo atlas of human subcortical
876	brain nuclei. Scientific data 5:180063.
877	Phelps EA (2006) Emotion and cognition: Insights from studies of the human amygdala. Annu Rev
878	Psychol 57:27-53.

879	Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies -
880	revisited. NeuroImage 84:971-985.
881	Robinson OJ, Vytal K, Cornwell BR, Grillon C (2013) The impact of anxiety upon cognition:
882	perspectives from human threat of shock studies. Front Hum Neurosci 7.
883	Robinson OJ, Charney DR, Overstreet C, Vytal K, Grillon C (2012) The adaptive threat bias in anxiety:
884	Amygdala-dorsomedial prefrontal cortex coupling and aversive amplification. NeuroImage
885	60:523-529.
886	Rossion B, Pourtois G (2004) Revisiting Snodgrass and Vanderwart's object pictorial set: The role of
887	surface detail in basic-level object recognition. Perception 33:217-236.
888	Schmidt SJ (2020) Distracted learning: Big problem and golden opportunity. J Food Sci Educ 19:278-
889	291.
890	Schmitz A, Grillon C (2012) Assessing fear and anxiety in humans using the threat of predictable and
891	unpredictable aversive events (the NPU-threat test). Nat Protoc 7:527-532.
892	Schönberg T, Daw ND, Joel D, O'Doherty J (2007) Reinforcement Learning Signals in the Human
893	Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. The
894	Journal of Neuroscience 27:12860-12867.
895	Schultz W (2016) Dopamine reward prediction-error signalling: a two-component response. Nature
896	reviews Neuroscience 17:183-195.
897	Schwarz GE (1978) Estimating the dimension of a model. Annals of Statistics 6:461-464.
898	Spielberger CD, Gorsuch RL, Lushene R, Vagg PR, Jacobs GA (1983) Manual for the State-Trait Anxiety
899	Inventory. Palo Alto, CA: Consulting Psychologists Press.
900	Stanton K, Watson D (2014) Positive and Negative Affective Dysfunction in Psychopathology. Social
901	and Personality Psychology Compass 8:555-567.
902	Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group
903	studies. NeuroImage 46:1004-1017.
904	Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.
905	Todorov A (2012) The role of the amygdala in face perception and evaluation. Motivation and
906	emotion 36:16-26.
907	Tovote P, Fadok JP, Luthi A (2015) Neuronal circuits for fear and anxiety. Nature Reviews
908	Neuroscience 16:317-331.
909	Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M
910	(2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical
911	parcellation of the MNI MRI single-subject brain. NeuroImage 15:273-289.
912	Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly High Correlations in fMRI Studies of
913	Emotion, Personality, and Social Cognition. Perspectives on psychological science : a journal
914	of the Association for Psychological Science 4:274-290.
915	Weger M, Sandi C (2018) High anxiety trait: A vulnerable phenotype for stress-induced depression.
916	Neurosci Biobehav Rev 87:27-37.
917	Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in
918	humans. Neuron 58:967-973.
919	Wylie GR, Genova H, DeLuca J, Chiaravalloti N, Sumowski JF (2014) Functional Magnetic Resonance
920	Imaging Movers and Shakers: Does Subject-Movement Cause Sampling Bias? Human brain
921	mapping 35:1-13.
922	Xu P, Gu R, Broster LS, Wu R, Van Dam NT, Jiang Y, Fan J, Luo YJ (2013) Neural basis of emotional
923	decision making in trait anxiety. J Neurosci 33:18641-18653.
924	

926 **Tables**

927 **Table 1.** Object pairs used in the reinforcement learning task.

		Object
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

928

929

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

Condition	Trials	16-20	Trials 26-30			
Condition	r	p-value	r	p-value		
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		

930 r: Two-tailed uncorrected Pearson's correlation coefficient.

Table 3. Correlations between trait anxiety and learning performance in the fMRI study. Datawas normalized based on the Neutral condition.

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

933 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. Data was normalized based on the Neutral condition.

Condition	Win	-stay	Lose-shift		
Condition	r	r p-value		p-value	
Affirmative Gain	х	Х	0.343	0.080	
Affirmative Loss	х	Х	0.108	0.591	
Contradictory Gain	-0.134	0.505	х	Х	
Contradictory Loss	-0.420	0.029	х	х	

936

937 r: Two-tailed uncorrected Pearson's correlation coefficient. X indicates that no such action was

938 available for the neutral feedback in this condition.

939 **Table 5.** Model fits and parameters. Mean (SEM)

Models									
				(0	ϕ_{OFFG}	(0		3φ,	3φ,
12φ	9φ	6φ₀	3φ₀	Ψ0FF,	$\phi_{\text{OFFL},}$	Ψ+1H,	3φ, 3ε	ε _{FF} ,	ε _{0FF} ,
				ΨONH	ϕ_{ONH}	Ψ-1Η		ϵ_{NH}	ϵ_{ONH}
110.8	113.6	113.9	117.8	118.7	117.8	117.3	117.7(118.5(118.6
(7.8)	(7.8)	(7.8)	(7.8)	(7.8)	(7.7)	(7.7)	7.8)	7.8)	(7.8)
303.9	291.9	286.8	276.9	272.8	276.8	275.9(282.4	278.3	278.5
(15.6)	(15.6)	(15.5)	(15.5)	(15.6)	(15.5)	15.5)	(15.6)	(15.5)	(15.5)
0.22	0.21	0.25	0.23	0.21	0.24	0.23	0.23	0.24	0.24
(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
0.07	(0.07	(0.10	0.15 ((0.12	0.12	(0.10	0.19	(0.03)	0.14
(0.01)	(0.02)	-0.25	-0.27	-0.23	-0.21	(0.02)	-0.77	-0.78	-0.26
		(0.06)	(0.07)	(0.06)	(0.06)		(0.05)	(0.05)	(0.07)
		(0.00)	(0.07)	(0.00)	(0.00)	0.21	0.05	0.03	0.26
						(0.06)	(0.08)	(0.08)	(0.12)
		0.70	0.81	0.87	0.76	· ,	0.80	0.74	0.82
		(0.06)	(0.06)	(0.04)	(0.07)		(0.06)	(0.06)	(0.06)
				0.37					
				(0.06)					
-0.40	-0.23								
(0.08)	(0.07)								
-0.25	-0.15								
(0.08)	(0.06)					0.05			
-0.26	-0.25					-0.25			
(0.08)	(0.07)	0.21			0.27	(0.08)			
(0.02)		0.21			0.27				
0.00)		0.05)			(0.07)				
(0.20		(0.20							
0.16		0.22							
(0.07)		(0.05)							
0.59	0.50	(0.00)							
	12φ 110.8 (7.8) 303.9 (15.6) 0.22 (0.04) 0.07 (0.01) -0.25 (0.08) -0.26 (0.08) -0.26 (0.08) 0.15 (0.03) 0.20 (0.06) 0.16 (0.07) 0.59	12φ 9φ 110.8 113.6 (7.8) (7.8) 303.9 291.9 (15.6) (15.6) 0.22 0.21 (0.04) (0.04) 0.07 0.07 (0.01) (0.02) -0.40 -0.23 (0.08) (0.07) -0.25 -0.15 (0.08) (0.07) -0.26 -0.25 (0.08) (0.07) 0.15 (0.03) 0.20 (0.06) 0.16 (0.07) 0.59 0.50	12φ 9φ $6φ_0$ 110.8 113.6 113.9 (7.8) (7.8) (7.8) 303.9 291.9 286.8 (15.6) (15.5) 0.22 0.21 0.25 (0.04) (0.04) (0.04) (0.04) 0.07 0.07 0.10 (0.01) (0.02) -0.25 (0.06) -0.25 (0.06) -0.25 -0.15 (0.06) -0.25 -0.15 (0.08) (0.08) (0.07) - -0.26 -0.25 (0.06) -0.26 -0.25 (0.06) -0.26 -0.25 (0.06) -0.26 -0.25 (0.05) 0.08 (0.07) (0.05) 0.15 0.21 (0.05) 0.20 0.26 (0.06) 0.16 0.22 (0.07) 0.16 0.22 (0.07) 0.59 0.50 (0.05)	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	Models 12φ 9φ 6φ₀ 3φ₀ φοrespective φοrespective φοrespective φοrespective φorespective φorespective	Models 12φ 9φ 6φ₀ 3φ₀ φorf,	Models 12φ 9φ 6φ₀ 3φ₀ φ₀rered φ₀ree <td>Models 12φ 9φ 6φ0 3φ0 φ0FFG φ0FFG φ1H 3φ, 3ε εFF, 110.8 113.6 113.9 117.8 117.8 117.8 117.8 117.7 118.5((7.8) (7.8) (7.8) (7.8) (7.8) 7.7) 7.8) 7.8) 303.9 291.9 286.8 276.9 272.8 276.8 275.9 282.4 278.3 115.6) (15.5) (15.5) 15.6) 15.5) 15.5) 15.5 15.5) 0.22 0.21 0.25 0.23 0.21 0.24 0.23 0.23 0.24 (0.04) (0.04) (0.04) (0.04) (0.04) (0.04) 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.03 0.04 0.05 0.03 0.05 0.21 0.05 0.05<</td>	Models 12φ 9φ 6φ0 3φ0 φ0FFG φ0FFG φ1H 3φ, 3ε εFF, 110.8 113.6 113.9 117.8 117.8 117.8 117.8 117.7 118.5((7.8) (7.8) (7.8) (7.8) (7.8) 7.7) 7.8) 7.8) 303.9 291.9 286.8 276.9 272.8 276.8 275.9 282.4 278.3 115.6) (15.5) (15.5) 15.6) 15.5) 15.5) 15.5 15.5) 0.22 0.21 0.25 0.23 0.21 0.24 0.23 0.23 0.24 (0.04) (0.04) (0.04) (0.04) (0.04) (0.04) 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.04 0.03 0.04 0.05 0.03 0.05 0.21 0.05 0.05<

	(0.06)	(0.06)								
	0.69	0.59								
ϕ_{+1NG}	(0.06)	(0.07)								
	0.62	0.53 (0.71			
ϕ_{+1HG}	(0.06)	0.07)					(0.07)			
	0.08	•	0.15			0.14	. ,			
ϕ_{OFFG}	(0.09)		(0.09)			(0.12)				
	0.10		0.14			、 ,				
ϕ_{ONG}	(0.09)		(0.09)							
	0.26		0.30							
ϕ_{OHG}	(0.08)		(0.09)							
	(0.00)	0.09	(0.05)	0 29	0.35					
ϕ_{OFF}		(0.05)		(0.07)	(0.06)					
		0.15		0 33	(0.00)	0.29				
ϕ_{ON}		(0.04)		(0.06)		(0.06)				
		0.14		0.32		(0.00)				
ф _{он}		(0.05)		(0.07)						
		(0.00)		(0.07)			-0 18			
$\phi_{\text{-1FF,NL}}$							(0.07)			
							0.59			
$\phi_{\text{+1FF,NL}}$							(0.07)			
							(0.07)	0.40	0.45	
ε _{FF}								(0.09)	(0.06)	
								0.48	(0.00)	
ε _N								(0.40 (0.08)		
								0.00)		
ε _H								(0.00)		
								(0.09)	0.40	
ε _{NH}									(0.45)	
									(0.00)	0.04
ε _{off}										(0.12)
										(0.12)
ε _{onh}										0.07
										(0.11)

LLE is the log-likelihood estimate. BIC is the Bayes Information Criterion. α denotes the 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 $-1\square$ feedback, φ_{+1} is the subjective value for $+1\square$ feedback, φ_{0FF} is the subjective value for the feedback combining $0\square$ and fearful faces, and φ_{0NH} is the subjective value for $0\square$ + happy or fearful face. ε_X is the bias added for feedback combination X. For example, ε_{0FF} is the bias term for neutral $0\square$ feedback presented together with fearful faces, and ε_H is the bias term for happy faces.

947 Table 6. Repeated measures ANOVA. R DLPFC beta parameter estimates for the 'Basic' prediction948 error term.

Drodictor	Sum of	-16	Mean of	r		m ²
Predictor	Squares	uj	Squares	Г	þ	Чр

(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	<0.0001
TrAnx x GaLo	0.003	1	0.003	0.008	0.927	0.0003
Error(GaLo)	10.02	25	0.401			
Feedback	0.024	1	0.024	0.185	0.671	0.007
TrAnx x Feedback	0.227	1	0.227	1.776	0.195	0.066
Error(Feedback)	3.196	25	0.128			
GaLo x Feedback	0.26061	1	0.261	1.357	0.255	0.051
TrAnx x GaLo x Feedback	0.001	1	0.001	0.004	0.951	0.0002
Error(GaLo x Feedback)	4.801	25	0.192			

TrAnx = Continuous covariate Trait anxiety; GaLo = Factor Gain/Loss (Gain or Loss pair); Feedback 949

950 = Factor Feedback (Good, Bad). p_{GG} = Greenhouse-Geisser corrected p-value. η_p^2 = Partial eta-

squared. * p < 0.05. 951

952 Table 7. Repeated measures ANOVA. R DLPFC beta parameter estimates for the 'Boost prediction

95

53	error term.		

Predictor	Sum of	df	Mean of	Mean of F		n ²
Tredictor	Squares	uj	Squares	I	Р	чр
(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 x 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 x Feedback	0.737	1	0.737	4.676	0.040	0.158
Error(Feedback)	3.939	25	0.158			
GaLo x Feedback	0.802	1	0.802	5.742	0.024	0.187
TrAnx x GaLo x Feedback	0.087	1	0.087	0.622	0.438	0.024
Error(GaLo x Feedback)	3.492	25	0.140			

TrAnx = Continuous covariate Trait anxiety; GaLo = Factor Gain/Loss (Gain or Loss pair); Feedback 954

= Factor Feedback (Good, Bad). p_{GG} = Greenhouse-Geisser corrected p-value. η_p^2 = Partial eta-955

956 squared. * p < 0.05.

957 Table 8. Repeated measures ANOVA. Amygdala beta parameter estimates for the 'Basic' 958 prediction error term.

Predictor	Sum of	df	Mean of	F	n	n^2
	Squares		Squares		μ	Чр
(Intercept)	0.16818	1	0.16818	2.3112	0.14099	
TrAnx	0.008879	1	0.008879	0.12202	0.72978	0.004857
Error	1.8192	25	0.072768			
GaLo	0.05135	1	0.05135	0.73074	0.40076	0.028399
TrAnx x GaLo	0.001819	1	0.001819	0.025884	0.87348	0.001034

Error(GaLo)	1.7568	25	0.070272			
Feedback	0.068436	1	0.068436	1.1218	0.29966	0.042946
TrAnx x Feedback	0.004216	1	0.004216	0.069115	0.79478	0.002757
Error(Feedback)	1.5251	25	0.061006			
GaLo x Feedback	0.013725	1	0.013725	0.25587	0.61741	0.01013
TrAnx x GaLo x Feedback	0.007149	1	0.007149	0.13327	0.71814	0.005302
Error(GaLo x Feedback)	1.3411	25	0.053643			

959 TrAnx = Continuous covariate Trait anxiety; GaLo = Factor Gain/Loss (Gain or Loss pair); Feedback

960 = Factor Feedback (Good, Bad). p_{GG} = Greenhouse-Geisser corrected p-value. η_p^2 = Partial eta-

961 squared. * p < 0.05.

962 **Table 9.** Repeated measures ANOVA. Amygdala beta parameter estimates for the 'Boost 963 prediction error term.

Predictor	Sum of	df	Mean of	E E	2	n ²	
Predictor	Squares	uj	Squares	Г	þ	ıp	
(Intercept)	0.012961	1	0.012961	0.18998	0.66668		
TrAnx	0.11583	1	0.11583	1.6977	0.20446	0.063589	
Error	1.7057	25	0.068226				
GaLo	0.11561	1	0.11561	1.9114	0.17903	0.071026	
TrAnx x GaLo	0.040593	1	0.040593	0.67114	0.4204	0.026144	
Error(GaLo)	1.5121	25	0.060483				
Feedback	0.011858	1	0.011858	0.274	0.60527	0.010842	
TrAnx x Feedback	0.026619	1	0.026619	0.6151	0.44024	0.024013	
Error(Feedback)	1.0819	25	0.043276				
GaLo x Feedback	0.044516	1	0.044516	0.54709	0.4664	0.021415	
TrAnx x GaLo x Feedback	0.076804	1	0.076804	0.94392	0.34058	0.036383	
Error(GaLo x Feedback)	2.0342	25	0.081367				

964

4 TrAnx = Continuous covariate Trait anxiety; GaLo = Factor Gain/Loss (Gain or Loss pair); Feedback

965 = Factor Feedback (Good, Bad). p_{GG} = Greenhouse-Geisser corrected p-value. η_p^2 = Partial eta-

966 squared. * p < 0.05.

968 **Table 10.** Brain regions showing significantly positive correlations between BOLD signal and the

969 basic prediction error term. p_{FWE} indicates familywise error rate (FWEr) corrected p-values for peak

970 voxel activities across the whole-brain. T-statistics were obtained from *t*-tests. Initial search

971 threshold: p=0.001, minimum cluster size: 5 voxels.

Drain region	Homienhoro	MNI p	eak cooi	dinate	Typlup	р
Brainregion	nemisphere	х	У	z	I-value	PFWE
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

972

973

975 Figure legends

976 Figure 1. A. Principle of the learning task in the pilot study. In each trial, participants select one object in a pair of objects. 977 The best and worst object in each pair respectively provides correct feedback with a probability of 0.7 and 0.3. Each pair 978 is presented 30 times, allowing participants to learn which the best object is by trial-and-error. B. Illustration of a trial 979 progression. C. Schematic of the outcomes provided in each pair type in the pilot study. In total four different pairs types 980 were presented: Contradictory Loss, Affirmative Loss, Contradictory Gain, and Affirmative Gain. D,E. The average change 981 in performance across participants for the different pair types in Gain (D) and Loss (E) conditions. A hit is defined as the 982 selection of the best object in a pair. F-I. Correlations between Trait anxiety and the average learning performance for 983 trials 16-20 in Affirmative and Contradictory pair types. Trait anxiety correlated significantly with learning performance 984 in Contradictory Loss pairs only (I). *p<0.05, ns=not significant (p>0.05), r = Pearson's correlation coefficient, ρ = 985 Spearman's rank-order correlation coefficient.

986 Figure 2. A. Schematic of the outcomes provided in each pair type in the fMRI study. Neutral pairs acted as control 987 conditions by presenting Neutral faces for both Correct and Incorrect outcomes. In total six different pairs types were 988 presented: Contradictory Loss, Affirmative Loss, Neutral Loss, Contradictory Gain, Affirmative Gain, and Neutral Gain. B. 989 The average change in performance across participants for the different pair types in Gain and Loss conditions. A hit is 990 defined as the selection of the best object in a pair. Learning was statistically assessed via the average performance in 991 trials 16-20. C-F. Learning in Affirmative and Contradictory pair types relative the Neutral control condition. Trait anxiety 992 significantly improved/impaired learning in Affirmative Gain pairs (C)/Contradictory Loss pairs (F). G-J. Win-stay and 993 Lose-shift decisions in Affirmative and Contradictory pair types relative their Neutral counterparts. Trait anxiety 994 significantly increased behavioral switching in Affirmative Gain pairs (J) and in Contradictory Loss pairs (M). *p<0.05, 995 ***p<0.001, ns=not significant (p>0.05).

996 Figure 3. A. Difference in Bayesian Information Criterion relative the most parsimonious model (highlighted in red). The 997 inset shows the protected exceedance probability (XPp) for these models. The most parsimonious model was the most 998 likely model, as evidenced by an exceedance probability of 1.0. B. The average model-simulated change in performance 999 for the different conditions in Gain and Loss pairs. C-F. Model-simulated learning in Affirmative and Contradictory pair 1000 types relative their Neutral control conditions. Trait anxiety significantly improved/impaired learning in Affirmative Gain 1001 pairs (C)/Contradictory Loss pairs (F). G. Trait anxiety correlated negatively with the difference in the model-fitted 1002 subjective values of the neutral $0\square$ outcome paired with fearful faces (ϕ_{0FF}) versus neutral/happy faces (ϕ_{0NH}). H,I. 1003 Model-simulated performance improvements for gradual changes in the difference between ϕ_{OFF} and ϕ_{ONH} . Smaller 1004 values of ϕ_{OFF} (vs. ϕ_{ONH}) improves performance in Affirmative Gain pairs (H), but impairs performance in Contradictory 1005 Loss pairs (I). For illustration purposes, the performance improvements for when ϕ_{OFF} is equal to ϕ_{ONH} is subtracted from 1006 all data points, the x-axis shows $\phi_{0NH}-\phi_{0FF}$, and the separate lines for Neutral and Contradictory pairs (relative 1007 Affirmative pairs) in H were merged into one line, and similarly were the lines for Neutral and Affirmative pairs (relative 1008 Contradictory pairs) in I. J-K. For a model that estimates separate values for the neutral OP outcome paired with fearful 1009 faces in Gain (φ_{OFFH}) and Loss (φ_{OFFL}) conditions, trait anxiety correlated negatively with the difference in the fitted 1010 subjective value between the neutral 0⊡feedback paired with fearful faces in both Loss (J) and Gain (K) pairs, as 1011 compared to neutral/happy faces, L-M. For a model that estimates separate values for the neutral 0回 outcome paired 1012 with neutral (ϕ_{0N}) or happy (ϕ_{0H}) faces, trait anxiety correlated negatively with the difference in the fitted subjective 1013 value between the neutral Onleedback paired with fearful faces, as compared to both neutral (L) and happy (M) faces. 1014 *p<0.05, ** p<0.01, *** p<0.001.

1015 Figure 4 A. Schematic of the functional localizer task. In each trial a number (-1, 0, +1) or face (Fearful, Neutral, Happy) 1016 was presented for 2.5s. Participants indicated whether the stimulus was perceived as negative, neutral, or positive. No 1017 feedback was presented and participants were not given any particular instructions regarding how stimuli should be 1018 categorized. B. The contrast between Neutral and Fearful faces revealed a region in the a priori R DLPFC mask that was 1019 significantly more activated by Neutral versus Fearful faces. C. For visualization purposes, the average beta parameter 1020 estimates for Neutral and Fearful faces were extracted for all voxels within the R DLPFC cluster shown in B. D. Trait 1021 anxiety correlated negatively with the contrast between Fearful and Neutral faces for the R DLPFC cluster. E. The contrast 1022 between Faces and Numbers revealed bilateral regions in the a priori amygdala mask that was significantly more 1023 activated by Faces versus Numbers. F. For visualization purposes, the average beta parameter estimates for Faces and 1024 Numbers were extracted for all voxels within the bilateral amygdala cluster shown in E. G. Trait anxiety did not correlate 1025 significantly with the contrast between Faces and Numbers for the amygdala cluster.

1026Figure 5. A. The R DLPFC ROI used to analyze prediction error encoding in the learning task. B. The average (solid line)1027and individual (dots) beta parameters for the 'Basic' prediction error term, $\delta_{Basic'}$ averaged across voxels within the R1028DLPFC for the four different feedback types. On average, activity in the R DLPFC ROI correlated significantly with the1029'Basic' prediction error term independent of Feedback type and Trait anxiety. C. The average (solid line) and individual1030(dots) beta parameters for the prediction error 'Boost' term, δ_{Boost} , averaged across voxels within the R DLPFC for the1031four different feedback types. On average, activity in the R DLPFC ROI did not correlate with δ_{Boost} across the four1032feedback types, but showed significant interactions with trait anxiety and feedback types (see main text and D-G). D-G.

1034 with δ_{Boost} for Incorrect feedback in Gain pairs (D) and negatively with δ_{Boost} for Correct feedback in Loss pairs (G). The 1035 different feedbacks presented in each feedback type is shown above the corresponding plot. The fitted subjective values 1036 for the 0₪ feedbacks differed between fearful and happy/neutral faces. H. The amygdala ROI used to analyze prediction 1037 error encoding in the learning task. I. The average (solid line) and individual (dots) beta parameters for the 'Basic' 1038 prediction error term, δ_{Basic} , averaged across voxels within amygdala for the four different feedback types. On average, 1039 activity in the amygdala ROI did not correlate with the 'Basic' prediction error term nor was there any interaction with 1040 trait anxiety or feedback types. J. The average (solid line) and individual (dots) beta parameters for the prediction error 1041 'Boost' term, δ_{Boost} , averaged across voxels within the amygdala ROI for the four different feedback types. On average, 1042 activity in the amygdala ROI did not correlate with δ_{Boost} nor were there any interactions with trait anxiety or feedback 1043 types. K-N. For visualization purposes, correlations between trait anxiety and δ_{Boost} for the four different feedback types 1044 are displayed. **p<0.01, ***p<0.001, ns = not significant (p>0.05). 1045 Figure 6. A. Schematic of the ventral tegmental area (VTA) ROI. B. Average (solid line) and individual (dots) beta 1046 parameter estimates for the 'Basic' prediction error term, δ Basic, within the VTA ROI. On average, BOLD signal in the VTA 1047 ROI correlated significantly with δBasic (p=0.001, one-tailed t-test). C-H. BOLD signal in the Midbrain, Dorsomedial PFC, 1048 Bilateral Striatum, and Bilateral Anterior Insula correlated significantly with δ Basic after applying FWEr correction for the 1049 whole-brain. For visualization purposes, average (solid line) and individual (dots) beta parameter estimates were 1050 extracted from the peak voxels within each respective cluster. ** p<0.01.

Correlations between trait anxiety and δ_{Boost} within the four different feedback types. Trait anxiety correlated positively

1051













<u>JNeurosci Accepted Manuscript</u>



