NeuroImage 229 (2021) 117709

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A ventral striatal prediction error signal in human fear extinction learning

M. Thiele[a,b], K.S.L. Yuen[a,b], A.V.M. Gerlicher[a,c], R. Kalisch[a,b,*]

[a] Neuroimaging Center (NIC), Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Center, Langenbeckstr. 1, 55131 Mainz, Germany
[b] Leibniz Institute for Resilience Research (LIR), Wallstr. 7, 55122 Mainz, Germany
[c] Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WS Amsterdam, The Netherlands

ABSTRACT

Animal studies have shown that the prediction error (PE) signal that drives fear extinction learning is encoded by phasic activity of midbrain dopamine (DA) neurons. Thus, the extinction PE resembles the appetitive PE that drives reward learning. In humans, fear extinction learning is less well understood. Using computational neuroimaging, a previous study from our group reported hemodynamic activity in the left ventral putamen, a subregion of the ventral striatum (VS), to correlate with a PE function derived from a formal associative learning model. The activity was modulated by genetic variation in a DA-related gene. To conceptually replicate and extend this finding, we here asked whether an extinction PE (EPE) signal in the left ventral putamen can also be observed when genotype information is not taken into account. Using an optimized experimental design for model estimation, we again observed EPE-related activity in the same striatal region, indicating that activation of this region is a feature of human extinction learning. We further observed significant EPE signals across wider parts of the VS as well as in frontal cortical areas. These results may suggest that the prediction errors during extinction learning are available to larger parts of the brain, as has also been observed in human neuroimaging studies of reward PE signaling. Conclusive evidence that the human EPE signal is of DAergic nature is still outstanding.

1. Significance statement

When one repeatedly experiences a feared situation in the absence of the anticipated negative consequences, one usually learns that one's fears were unfounded, and fear subsides. In the laboratory, this 'extinction learning' is modeled by first presenting some stimulus A together with an aversive stimulus, to produce fear of A, and then presenting A many times, but without the aversive stimulus. In mice, not receiving the aversive stimulus leads to firing of midbrain dopamine neurons, and this activity drives the subsequent reduction of fear. We here provide indirect evidence using neuroimaging that a similar learning mechanism may be operating during extinction in humans. This sheds light on the neurobiological processes putatively underlying the treatment of fear-related disorders with exposure-based therapy.

2. Introduction

According to associative learning theories (Pearce and Hall, 1980; Recorla and Wagner, 1972), the accumulation of new information is driven by a prediction error (PE) – the difference between expected and observed outcomes. In his seminal studies, Schultz (1998 for review) found that phasic burst-firing of midbrain dopamine (DA) neurons corresponds to a reward prediction error (RPE), tracking unexpected reward. Subsequently, O'Doherty et al. (2003) translated this finding to humans by showing that the blood oxygenation-level dependent (BOLD) signal measured with functional magnetic resonance imaging (fMRI) in a target region of midbrain DA neurons, the left ventral putamen, correlates with a theoretically derived RPE function. Recently, two imaging meta-analyses demonstrated that the RPE signal is widely scattered across the whole striatum and also found an association between BOLD activity and RPE signaling in many cortical areas, especially in cingulate and medial and lateral frontal areas (Chase et al., 2015; Garrison et al., 2013).

Fear extinction is a case of associative learning where a stimulus (conditioned stimulus, CS), previously coupled with an aversive event (unconditioned stimulus, UCS), is no longer accompanied by the feared outcome. As a consequence, the conditioned fear reaction (CR) towards the CS diminishes (Pavlov, 1927). The unexpected omission of an expected UCS during fear extinction is a better-than-expected outcome and may be experienced as a pleasant surprise or relief. It has therefore been hypothesized that the extinction prediction error (EPE) is analogous to the appetitive RPE and is mediated by the same neural substrate (Raczka et al., 2011; Abraham et al., 2014). In line with this hypothesis, a study by our group (Raczka et al., 2011) found an EPE signal in an identical subregion of the ventral striatum (VS) as O'Doherty et al. (2003) that was, furthermore, modulated by a polymorphism of the DA transporter gene *DAT1*. Specifically, subjects with a less effective transporter variant (and therefore theoretically higher phasic DA peaks) learned extinction more quickly and exhibited larger EPE-correlated BOLD activity.

Meanwhile, studies in the fruit fly have shown that the same DA neuron population that mediates reward learning also mediates extinction, but not fear, learning (Felsenberg et al., 2018). Further,

[*] Corresponding author.
E-mail address: rkalisch@uni-mainz.de (R. Kalisch).

https://doi.org/10.1016/j.neuroimage.2020.117709
Received 7 October 2020; Received in revised form 30 November 2020; Accepted 19 December 2020
Available online 15 January 2021

Salinas-Hernández et al. (2018) showed in the mouse that midbrain DA neuron firing also tracks EPEs, and optogenetic inhibition of DA neurons precisely at the time of UCS omission was found to inhibit extinction learning (Luo et al., 2018; Salinas-Hernández et al., 2018), suggesting the DAergic EPE signal is necessary for fear extinction. Finally, optogenetic activation of the neurons accelerated learning (Salinas-Hernández et al., 2018), indicating the signal is also sufficient to drive extinction (Kalisch et al., 2019).

While a DAergic EPE can now be considered established in rodents, to date there is only one human imaging study (Raczka et al., 2011) linking the EPE with activity in the major output region of the mesolimbic DA system, the VS. Confirming EPE-related activity in this region, however, is a prerequisite for further pursuing the DAergic EPE hypothesis in humans. This criterion should also be fulfilled independently from the potential contribution of individual genetic differences to the signal, as were taken into account by Raczka et al. (2011). We therefore here tried to conceptually replicate Raczka et al. (2011), by predicting a simple EPE main effect (ignoring genotype) in the same VS subregion (specifically, the left ventral putamen) that was found by them as well as by the earlier RPE study by O'Doherty et al. (2003) (*main analysis*). If confirmed, such a result would indicate consistency of the current findings with one RPE study (O'Doherty et al., 2003) and one EPE study where the EPE is considered of DAergic nature (modulated by *DAT1* genotype; Raczka et al.). In order to compensate for the large sample sized used in Raczka et al., we optimized our EPE modeling procedures.

To extend these findings, we also searched for significant activation in the wider striatal region meta-analytically linked with RPE signaling (Garrison et al., 2013) (*secondary analysis*). Lastly, we explored potential extra-striatal EPE signals (additional *exploratory analyses*).

3. Materials and methods

3.1. Experimental design

3.1.1. Participants

32 participants were recruited for the study. Inclusion criteria were: (a) age between 18 and 39 years, (b) no self-reported physical, neurological or psychiatric illness, (c) no use of illicit drugs, (d) no former participation in fear conditioning experiments. Participants were screened for psychiatric illnesses and drug consumption in a telephone interview and with an extensive neuropsychiatric interview. We excluded one participant from all analyses because of technical issues with the scanner, leading to early termination and an incomplete dataset. The resulting sample size for the analysis of fear rating data (see 2.1.5) was n=31 (average age: 23.3 years, range: 19-39 yrs, 19 female). From this subsample, one additional participant was excluded specifically from the SCR analysis because he was a non-responder (see 2.1.4), reducing the final sample size for SCR analysis to n=30 (23.4 yrs, 19-39 yrs, 19 fem.). From the same subsample of n=31 participants, two participants were excluded specifically from the fMRI analysis because of excessive head movement during fMRI acquisition (see 2.1.6) (final sample size for fMRI analysis: n=29, 23.2 yrs, 19-39 yrs, 18 fem.). Supplementary Table S1 gives an overview.

The Ethics Committee of the State Medical Board in Rheinland-Pfalz, Germany, approved the study, and all participants gave written informed consent.

3.1.2. Experimental task

The experimental task was an fMRI task adapted from Raczka et al. (2011) and was performed using Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). It consisted of three phases that directly followed each other without pause or announcement: fear acquisition, fear extinction and fear reacquisition (see Figure S1 for schematic diagram). Two geometric symbols (square and diamond) were each presented 12 times per phase (8s trial duration, inter-trial intervals (ITIs) were jittered between 11s and 14s with a mean of 12.5s) on a black background. During fear acquisition and reacquisition, one of the two symbols (CS+) was paired with a painful, but tolerable electric stimulation at the right ankle (UCS) in 66% of trials. We thus lowered the reinforcement ratio compared to Rackza et al., who had used an 80% ratio, with the intention to thereby reduce initial EPEs and accordingly prolong the time window in early extinction in which a substantial EPE signal can be detected. The UCS was applied 50ms before the offset of the CS+. The other symbol (CS-) was never paired with the UCS. Assignment of symbols to CS+ and CS- was counter-balanced across participants. We used two fixed trial orders, whereupon 50% of all participants were assigned to the first and the other 50% to the second order. In both trial orders, one symbol never occurred more than two times in a row; the first two trials always contained one CS+ and one CS-; and the UCS was always omitted at the 1st, 3rd, 5th and 8th CS+ trial of fear acquisition and the 1st, 2nd, 5th and 6th CS+ trial of fear reacquisition.

Participants were instructed to learn about the relationship between the symbols and the pain stimulus. In Raczka et al.'s paradigm, participants had provided a rating of their fear of the last experienced CS+ and CS- after every 12th trial. Thus, ratings in that previous study were retrospective and infrequent. By contrast, in the present study, participants continuously indicated their level of fear in a moment-to-moment manner throughout the entire experiment (during Cs and ITIs). For this purpose, they used a button box that moved a cursor along a visual analog scale (poles 0 = no fear and 100 = highest level of fear) that was always present at the bottom of the screen. We reasoned that, by providing more frequent and momentaneous fear ratings as a basis for PE model fitting in the task (see below), we would obtain more reliable and accurate PE estimates. The cursor's starting position at the first trial was 50.

Taken together, the adapted version of the Raczka et al. task contained potential improvements, which we expected to increase the sensitivity of PE signal detection in fMRI.

3.1.3. UCS calibration

The UCS was adjusted individually before the experiment to ensure maximum tolerable pain. Participants were given a series of stimuli of increasing intensity, starting from the perceptual threshold. After every stimulation, participants rated their individual pain level on a scale from 1 (very low level of pain) to 10 (the strongest pain one can imagine given the applied electrode). The stimulation amplitude was increased until participants did not want to receive higher stimulation (average stimulation intensity: 13.6μS, SD: 9.7μS; average pain level rating: 9.6, SD: 0.9). Before the experiment, all participants agreed explicitly to the stimulation amplitude.

3.1.4. Acquisition and preprocessing of skin conductance data

Skin conductance was recorded from self-adhesive Ag/AgACI electrodes (EL-507, BIOPAC® Systems Inc., Goleta, California, USA) attached to the palm of the left hand using a BIOPAC MP150 with EDA100C device (BIOPAC® Systems Inc., Goleta, California, USA). The raw signal was amplified and low-pass filtered with a cut-off frequency of 1Hz. Custom-made scripts running on MATLAB2015b and MATLAB 2017b (MathWorks, Inc., Natick, Massachusetts, United States, www.matlab.com) were used for analysis. In deviation from Raczka et al. (2011), we preprocessed SCRs following the updated analysis recommendations by Lonsdorf et al. (2019). SCRs were scored manually with a custom-made script running on MATLAB 2015b as the trough-to-peak difference between the first response onset (between 1s to 4s following stimulus onset) and the successive response maximum. Responses smaller than 0.01µ were scored as zero. Non-responders were defined as participants showing more than 50% zero-scored SCR responses towards the UCS. Using these criteria, one scored participant had to be excluded from the SCR analysis. SCRs were log-transformed and range-corrected before statistical testing.

3.1.5. Preprocessing of fear rating data

Average fear ratings per trial were extracted with a custom-made script running on MATLAB 2015b from the continuous rating time courses (from stimulus onset until 50 ms before stimulus offset), resulting in trial-by-trial fear rating values.

3.1.6. Acquisition and preprocessing of imaging data

Imaging was conducted on a Siemens MAGNETOM Trio 3 Tesla MRI system with a 32-channel head coil. FMRI data was acquired using gradient echo, echo planar imaging (EPI) with a multiband sequence covering the whole brain (TR: 1850ms, TE: 34,6ms, multi-band acceleration factor: 4, voxel size: 2.1 isotropic, flip angle: 90°, field of view: 210mm). A high-resolution T1 weighted image was acquired after the end of the functional sequence for anatomical visualization and normalization of the EPI data (TR: 1900ms, TE: 2540ms, voxel size: 0.8mm isotropic, flip angle: 9°, field of view: 260mm).

Preprocessing was carried out in SPM12 (www.fil.ion.ucl.ac.uk/spm) running on MATLAB 2015b (MathWorks, Inc., Natick, Massachusetts, United States, www.matlab.com) and was identical to Raczka et al. (2011). The first 5 initial EPI images were discarded to account for the equilibration effect. Images were realigned to the 6th volume. Two participants with head motion exceeding a threshold of 3mm in translation respectively 2° in rotation were excluded. Realigned EPI volumes were co-registered to the anatomical image. The anatomical image was segmented and normalized to MNI space. The normalization parameters were then applied to the EPI volumes. Images were spatially smoothed using a 4mm with a full-width-at-half-maximum Gaussian kernel. Based on a reviewer recommendation, we used the ArtRepair toolbox (Mazaika et al., 2009) to scrub EPI images that correlate with vigorous head movements before calculating activations. A frame-to-frame displacement threshold of 1mm was used as criterion to determine if a brain volume requires scrubbing or not. The identified brain volumes were replaced by interpolating neighboring volumes to remove spikes.

3.1.7. Data and Code Accessibility

The computational modeling and the analysis of the fear ratings and SCRs and the fMRI analysis were conducted with custom-made Matlab code optimized for MATLAB 2015b and MATLAB 2017b (MathWorks, Inc., Natick, Massachusetts, United States, www.matlab.com) and (in case of fMRI analysis) SPM12 (www.fil.ion.ucl.ac.uk/spm). All multivariate statistics were run with SPSS 23 (IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp). All code was run on a Linux Remote-desktop Server (Intel® Xenon® CPU E5-2690 0, 2.90 GHz processor, 128GB RAM) running on Debian 8. The code is freely available online (osf.io/qf4vj/; see Table S2 for detailed explanations). Raw fMRI data files are not accessible due to data protection rules (GDPR). Statistical images from the single-subject level fMRI models are available on request. The anatomical masks used for region of interest (ROI) analyses (see below) are available at osf.io/qf4vj (see Table S3).

3.2. Statistical analyses

3.2.1. Fear rating and skin conductance data

Fear ratings and skin conductance responses (SCRs) were used to assess conditioned responding. While fear ratings are subjective and explicit, SCRs have the advantage of being objective and implicit. Although modeling was done on fear ratings only (see 2.2.2), SCRs were acquired to provide important corroborating evidence for successful learning and to comply with our laboratory standards. Separate repeated-measures ANOVAs were used for the analysis of trial-wise fear rating as well as SCR data, testing for effects of stimulus (CS+ vs. CS-), time (early trials 1-6 vs. late trials 7-12) and their interactions, separately per experimental phase. All univariate tests were performed with SPSS 23 (IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp).

3.2.2. Computational modeling of fear rating data

The same computational modeling procedure as in Raczka et al. (2011) was applied with custom-made scripts running on MATLAB 2015b. The only exception was that we modeled trial-by-trial fear ratings instead of retrospective ratings provided intermittently (see above). Note that, in order to replicate Raczka et al., we did not model SCRs. Further, we consistently find in our studies (including the Raczka et al. study as well as the current one) that SCRs measured in the MRI scanner show too much trial-to-trial variability to permit trial-wise modeling.

The modeling procedure is based on a simplified Rescorla-Wagner (RW) model as an established associative learning model (Rescorla and Wagner, 1972):

$$V_{t+1} = V_t + \alpha^*(R - V_t)$$

The model formalizes the expected aversive value V of a given CS over trials t (V_t) (see Fig. 1A). The expected value of a stimulus at the next trial t+1 is a function of its current expected value modified by the expectancy violation at the time of occurrence or non-occurrence of the UCS (i.e., at CS offset) – the aversive prediction error (PE). In fear conditioning, PE equals the difference between the actual aversive outcome (R) and the current expectancy (V_t) and is positive if R is unexpected or worse than expected. If the UCS is unexpectedly omitted (such as in unpaired CS+ trials during acquisition or during early extinction), the aversive PE is negative. The extinction prediction error EPE according to Raczka et al. is an appetitive signal mediated by the reward system and opposing, or inhibiting, the aversive learning system that mediates fear conditioning. It therefore carries an opposite sign to the aversive PE and, consequently, is positive when a UCS is unexpectedly omitted (see Fig. 1B, red dots). To update V for the next trial, the PE is weighted by a learning rate (α) that the model assumes to be constant across any phase of learning. Because, in accordance with Raczka et al., we assume different learning mechanisms for acquisition and extinction of fear, we also assume different learning rates for the three experimental phases (separate free parameters α_{acq}, α_{ext}, α_{reacq}).

Following Raczka et al., fear ratings were normalized to the average of the first CS+ and CS- ratings and range-corrected to the minimum and maximum of all ratings in the sample across all experimental phases. As a result, the first CS+ and CS- ratings of all subjects became 0.41 (V_1) (see Fig. 1A; cf. Raczka et al.: 0.36). R on reinforced CS+ trials was operationalized on an individual basis as the participant's last CS+ fear rating during fear acquisition scaled by the CS+ reinforcement ratio (rating/0.66). This value accordingly represents the aversiveness of receiving a UCS for each participant individually. For unreinforced trials (unpaired CS+, all CS-), R was set as the participant's last CS- fear rating during acquisition. Thereby, the R term signaled whether a UCS was delivered or not, providing the critical outcome information whose deviation from the expected aversive value V drives learning in the model. We used individual R values (rather than always setting R at 1 for paired and 0 for unpaired trials) to factor out any potential inter-individual differences in learning that might in fact merely result from differences in UCS processing.

To estimate the optimal individual learning rates for all phases in one go, an exhaustive grid-search was performed on all combinations of α values, ranging from 0.01 to 1 in steps of 0.01. The best individual model was selected based on a sum of least-squares (SLS) approach (best SLS: 0.05). For a sample average of the resulting model, see Fig. 1A.

3.2.3. Computational modeling of fMRI data

Analogous to Raczka et al. (2011), we set up separate fMRI models with SPM12 running on MATLAB 2015 for acquisition, extinction and

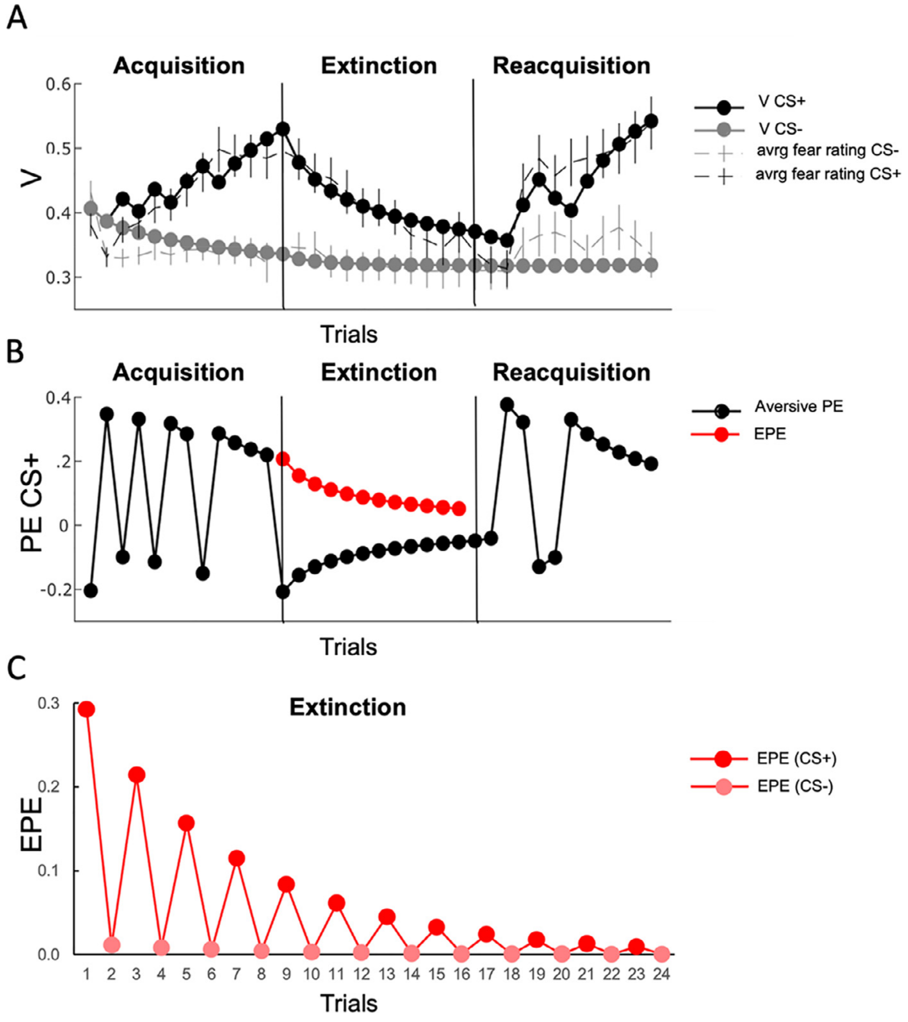

Fig. 1. Computational modeling results. A) Sample average of the aversive values (V) of CS+ (black dots) and CS- (grey dots) derived from the computational model. Crossed broken lines depict the measured sample-average CS+ and CS- fear ratings. B) Sample average of the aversive PEs towards the CS+ derived from the computational modeling (black dots). Red dots illustrate the oppositely signed EPE, as employed for fMRI analysis. PEs to the CS- are always minimal and are not shown. C) Parametric modulator for the categorical CS (CS+ and CS- combined) offset regressor from one trial sequence, as used in the single-subject level fMRI analysis to predict EPE-related activity during extinction.

reacquisition. All results reported below are based on the same SPM model. At the single-subject level of analysis, separate event-type regressors (stick functions) for CS onsets, CS offsets (both irrespective of CS+ or CS-), UCS presentation, and key presses on the button box modeled the time course of events. The categorical CS onset regressor was parametrically modulated by the individual's trial-by-trial V estimate, whereas the corresponding EPE estimate was used as parametric modulator of the CS offset regressor (see Fig. 1C). V and EPE estimates were derived from individual RW models using sample-averaged learning rates per phase (α_{acq}: M = 0.16 (SD = .24), α_{ext}: M = 0.27 (SD = 0.31), α_{reacq}: M = 0.13 (SD = 0.2); cf. Raczka et al.: means 0.16 / 0.21 / 0.19). By using sample-averaged learning rates, individual activation maps become comparable between individuals. The correlation between the critical PE regressor and the V regressor was sufficiently small (R=0.17; cf. Raczka et al.: 0.21) to permit robust estimation. The group-level design was a flexible factorial design including the parameter estimate images of the CS onset and offset regressors with their parametric modulators, separately per phase. Given the focus of this paper, we here only report results from the extinction phase.

3.2.4. *Anatomical hypotheses*

Our *main hypothesis* was EPE-related activation (EPE parametric modulator of CS offsets) in an a priori region of interest (ROI) consisting in a 6mm sphere centered around the peak voxel in the left ventral putamen identified by Raczka et al. (Montreal Neurological Institute (MNI) coordinates x,y,z = -32,8,-6; their Fig. 3b) in their own ROI analysis based on the RPE results by O'Doherty et al. (2003). See Fig. 2A, green voxels. To correct for multiple comparisons, we applied small-volume correction (SVC) using the family-wise error method (FWE) at an α threshold of 5% (voxel level).

For alternative, external validation of striatal EPE signaling in a *secondary analysis* (same SVC threshold), we used an ROI based on the whole-brain meta-analysis of RPE imaging studies by Garrison et al. (2013; their Fig. 4, yellow voxels). To restrict the ROI to striatal voxels, the whole-brain RPE mask provided by the authors was overlapped with the Harvard-Oxford masks for caudate, putamen, pallidum, and nucleus accumbens (Desikan et al., 2006; Frazier et al., 2005; Goldstein et al., 2007; Makris et al., 2006). This generated one single mask covering the wider striatum (Fig. 2A, yellow voxels).

In their exploratory analyses, Raczka et al. had also reported EPE-related BOLD activity in several extra-striatal regions (their supplement). In additional *exploratory analyses*, we therefore also tested replicability of major extra-striatal activations in 6mm spherical ROIs centered around voxels in left posterior superior frontal gyrus (-16,2,50), left lateral anterior cingulate sulcus (-16,40,18), and left intra-parietal sulcus (-16,-62,44). For one prominent midline activation reported by Raczka et al. (left basal forebrain; -2,2,-10), the ROI was a midline-centered box of dimensions 20mm x 16mm x 16mm around coordinates 0,2,-10, as practiced in previous work by our group (e.g., Lonsdorf et al., 2014). See Fig. 2B. This *exploratory analysis* was complemented by an ROI analysis using the extra-striatal voxels of Garrison et al.'s RPE mask. See Fig. 2C. Extra-striatal activations from both sources do not overlap.

All ROI masks are freely available (Table S3).

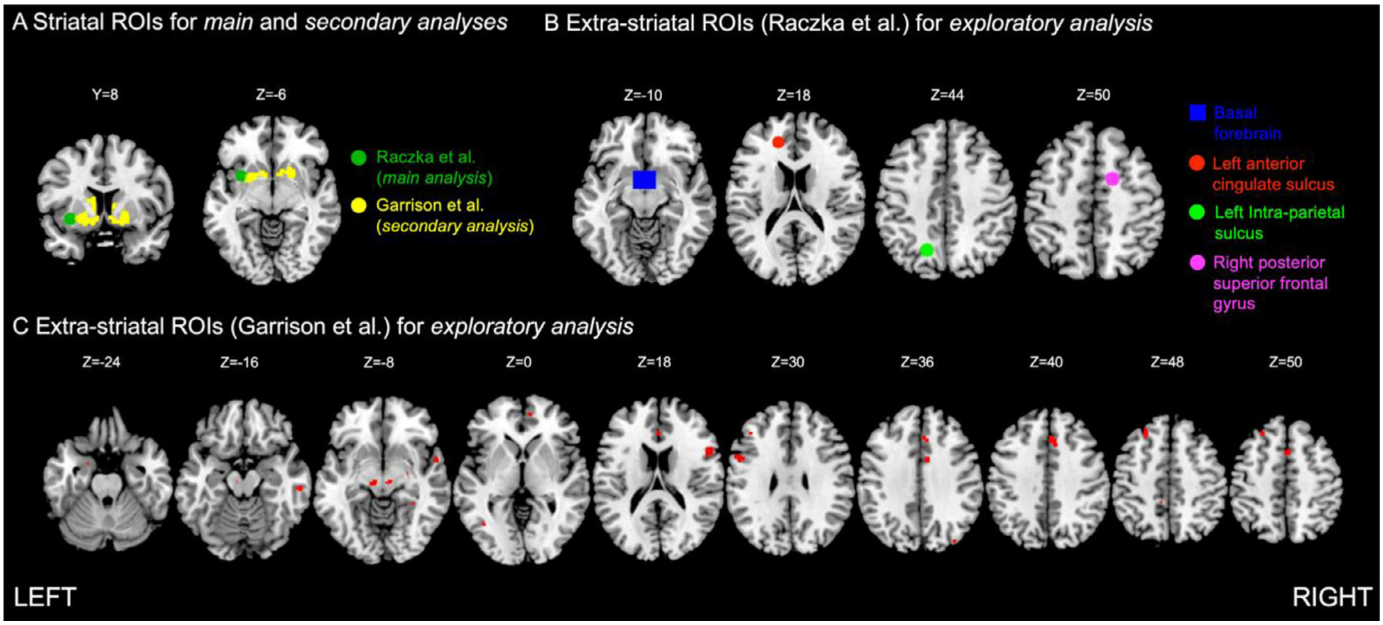
A Striatal ROIs for main and secondary analyses
B Extra-striatal ROIs (Raczka et al.) for exploratory analysis
Y=8
Z=-6
Raczka et al. (main analysis)
Garrison et al. (secondary analysis)
Z=-10
Z=18
Z=44
Z=50
Basal forebrain
Left anterior cingulate sulcus
Left Intra-parietal sulcus
Right posterior superior frontal gyrus
C Extra-striatal ROIs (Garrison et al.) for exploratory analysis
Z=-24
Z=-16
Z=-8
Z=0
Z=18
Z=30
Z=36
Z=40
Z=48
Z=50
LEFT
RIGHT

Fig. 2. A priori ROIs based on Raczka et al. (2011) and Garrison et al. (2013). A) Striatal ROIs. The green sphere is the left ventral putamen ROI based on Raczka et al. for the *main analysis*, yellow voxels are the wider striatal ROI based on Garrison et al. for the *secondary analysis*. B) Extra-striatal ROIs based on Raczka et al. (their supplement) for the additional *exploratory analysis*. C) Extra-striatal ROIs based on Garrison et al. for the additional *exploratory analysis*.

4. Results

4.1. *Fear ratings and skin conductance*

Fear rating data (Fig. 3A) showed successful acquisition, extinction and reacquisition of fear, as evidenced by significant main effects of stimulus (CS+>CS-) and significant stimulus by time interactions in all three experimental phases (acq: stimulus: $F(1,30) = 17.65$, $p < 0.001$; time: $F(1,30) = 7.55$, $p = 0.010$; stimulus by time interaction: $F(1,30) = 27.67$, $p < 0.001$; ext: stimulus: $F(1,30) = 18.89$, $p < 0.001$; time: $F(1,30) = 9.64$, $p = 0.004$; interaction: $F(1,30) = 17.51$, $p < 0.001$; reacq: stimulus: $F(1,30) = 18.76$, $p < 0.001$; time: $F(1,30) = 16.41$, $p < 0.001$; interaction: $F(1,3) = 25.6$, $p < 0.001$; Fig. 3B). SCR data (Fig. 3C) also showed significant stimulus effects in all three phases, indicating presence of conditioned fear, but these differential CRs did not decline over the course of extinction (no stimulus by time interaction). Instead, there was a decline of SCRs to both CS+ and CS- (main effect of time) (acq: stimulus: $F(1,29) = 24.35$, $p < 0.001$; time: $F(1,29) = 16.91$, $p < 0.001$; ext: stimulus: $F(1,29) = 14.74$, $p = 0.001$; time: $F(1,29) = 5.68$, $p = 0.024$; reacq: stimulus: $F(1,29) = 16.35$, $p < 0.001$; Fig. 3D).

4.2. *fMRI*

Main analysis: conceptual replication of Raczka et al. (2011). The EPE time course shown in Fig. 1B (red curve) was used to parametrically modulate the categorical CS offset regressor, as shown in Fig. 1C. As predicted, we observed significant BOLD activity related to this regressor in our predefined left ventral putamen ROI based on the Raczka et al. findings (see Methods and Fig. 2A) at peak coordinate x,y,z = -28,4,-8 ($Z = 3.77$; $p_{SVC} = 0.01$). The peak was part of a cluster located in the ventral putamen (Fig. 4A left). Another significant peak at -36, 10, -6 ($Z = 3.77$; $p_{SVC} = 0.01$) was located on the lateral border of the ROI and was part of a cluster situated in the insula rather than the ventral putamen. It will not be considered in the further (Fig. 4A right). The *main analysis* shows that the left ventral putamen is consistently activated to both RPE (as in the original study by O'Doherty et al., 2003) and EPE (Raczka et al., 2011, and present study).

Table 1
Results from the *secondary analysis* applying the striatal RPE ROI based on Garrison et al. (see Fig. 2A). L/R, left/right; x,y,z, MNI coordinates of peak activation; kE, number of voxels in cluster; Z, Z-statistics; psvc, p-value with FWE correction at the voxel level, using SPM's SVC.

L/R	Brain region	x,y,z	k_E	Z	p_{svc}
L	Pallidum	-14 6 -8	8	5.50	p<0.001
L	Ventral putamen	-24 6 -2	21	4.60	0.001
L		-26 6 -6		4.42	0.003
L		-28 2 -8		4.15	0.009
R	Ventral putamen	24 10 -4	4	4.19	0.007
R		22 8 0		3.68	0.049
R	Caudate	14 6 12	2	3.70	0.046

Secondary analysis: wider striatal EPE signal. Using a striatal ROI derived from a meta-analysis of RPE activation (Garrison et al., 2013; see Methods and Fig. 2A), we found significant EPE-associated BOLD activity in the left and right ventral putamen, the left pallidum and the right caudate (Fig. 4B, Table 1). This indicates that wider parts of the striatum, and in particular its ventral aspects, encode an EPE signal, as has also been observed previously for the RPE (Garrison et al., 2013; Chase et al., 2015). Note that the results of this and the previous analysis were highly similar when calculating different models that included either only odd or even CS+s (not shown).

Exploratory analysis: extra-striatal EPE signal. Using four extra-striatal ROIs derived from Raczka et al. (see Methods and Fig. 2B), we observed significant activity in the left anterior cingulate sulcus (-10,40,18; $Z = 3.93$; $p_{SVC} = 0.005$; and -12,38,14; $Z = 3.79$; $p_{svc} = 0.008$) only. Due to the exploratory nature of the analysis, we did not correct for testing multiple ROIs (four), but note that the observed activation would have survived conservative Bonferroni correction at $\alpha = 0.05/4 = 0.0125$. Applying the extra-striatal ROI derived from the RPE meta-analysis by Garrison et al. (2013) (see Methods and Fig. 2C) revealed significant activation in midline and right frontal cortices (Table 2).

Fig. S2 and Table S4 report results of an exploratory whole-brain analysis corrected at p<0.05 FWE, performed for the purpose of hypothesis generation and to facilitate potential meta-analysis. Note that

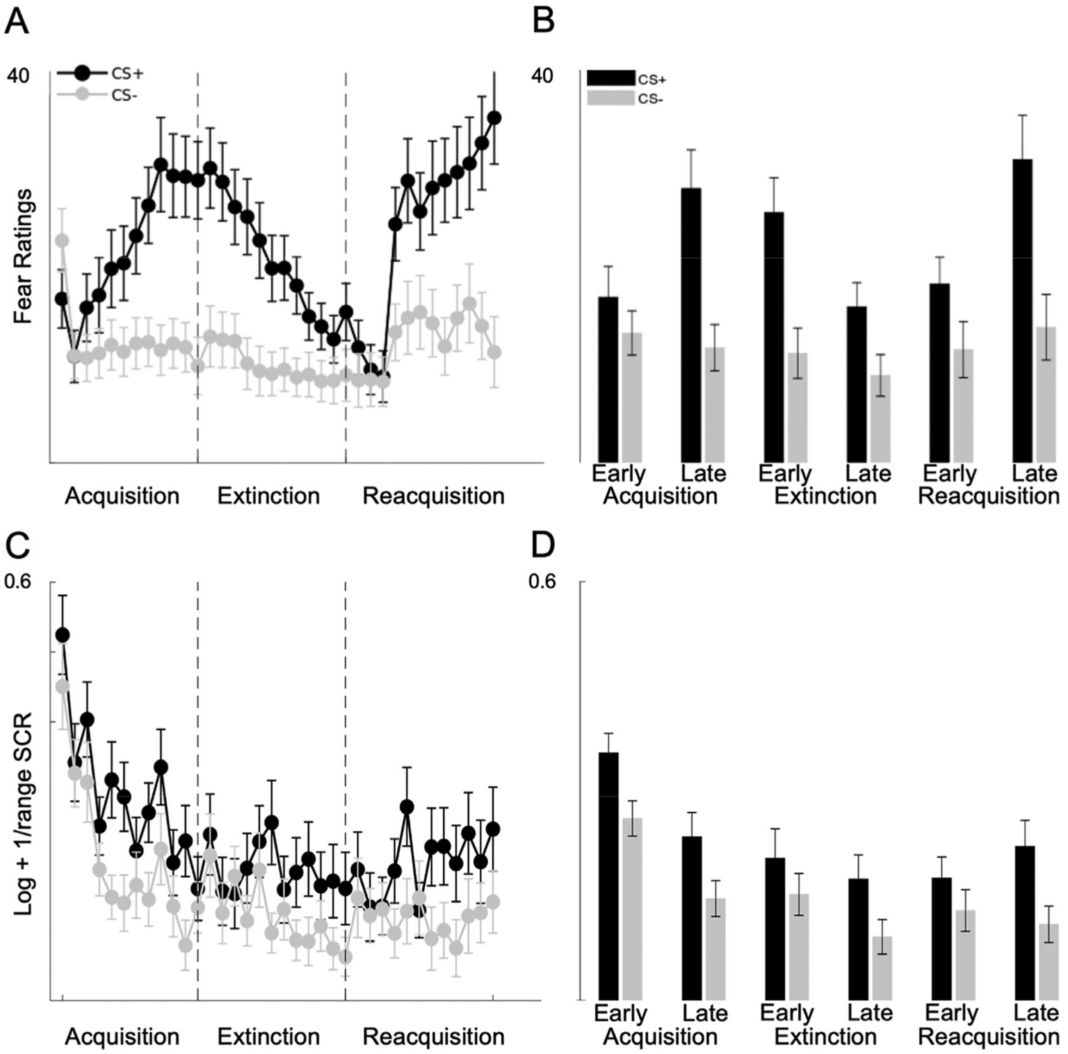

Fig. 3. Fear ratings and SCRs. Conditioned responses were assessed as fear ratings (A, trial-by-trial time course; B, early vs. late averages) and SCRs (C, time course; D, averages). Both measures show stimulus main effects (CS+>CS-) in all phases; fear ratings also show stimulus by time interactions in the three phases. Black: CS+; grey: CS-. Early: trials 1-6; late: trials 7-12 (as used for ANOVA). Error bars indicate SEM.

Table 2
Results from the *exploratory analysis* applying the extra-striatal RPE ROI based on Garrison et al. (see Fig. 2C). L/R, left/right; x,y,z, MNI coordinates of peak activation; kE, number of voxels in cluster; Z, Z-statistics; psvc, p-value with FWE correction at the voxel level, using SPM's SVC.

L/R	Brain region	x,y,z	k_E	Z	p_{svc}
R	Supplementary motor area	4 12 54	14	4.67	0.001
L	Supplementary motor area	-14 -2 68	5	4.05	0.010
R	Rostral dorsal ACC	8 26 38	8	3.91	0.017
		4 28 40		3.54	0.064
R	Inferior frontal gyrus	56 14 16	5	3.79	0.028

this analysis yielded significant activation at the whole-brain level in the left pallidum peak at -14,6,-8 identified in above *secondary analysis* (see Fig. 4B and Table 1), in a peak in the left putamen (-26,8,14), as well as in numerous frontal, parietal and temporal areas.

5. Discussion

The aim of the present study was to investigate the neural correlates of the extinction prediction error (EPE) in a task optimized for detecting EPE signals in fMRI by employing continuous momentary instead of intermittent retrospective fear ratings. Using EPE estimates based on these ratings, we observed that activity in the left ventral putamen follows an EPE signal over the course of extinction learning. Thus, we conceptually replicate Raczka et al.'s (2011) finding of EPE-related hemodynamic activity in this region. Thereby, our results corroborate the existence of an EPE signal in this part of the VS. Note that the left ventral putamen is also the first region in which an RPE signal was detected using fMRI (O'Doherty et al., 2003) and that it has recently been confirmed as the area most consistently activated across fMRI studies modeling various types of PE signals (Chase et al., 2015; their Fig. 5). Our aim to make an analogy between neural RPE signals and a neural EPE signal correlating with the omission of an aversive event is based on studies showing that RPEs and relief-related PEs depend on the same neural signal in the VS, at least when primary reinforcing stimuli (e.g., pain or non-pleasant food) are applied (O'Doherty et al., 2004; Seymour et al., 2005). Our results are clearly in accordance with this line of evidence, which further confirms their robustness.

Our data further indicate that the EPE signal is not restricted to the left ventral putamen but rather spreads widely across the VS. As such, it resembles the RPE, for which a wide striatal representation has also been demonstrated (Garrison et al., 2013; Chase et al., 2015). The apparent anatomical overlap of EPE and RPE signals is consistent with the appetitive nature of EPE signals, claimed by Raczka et al. (2011).

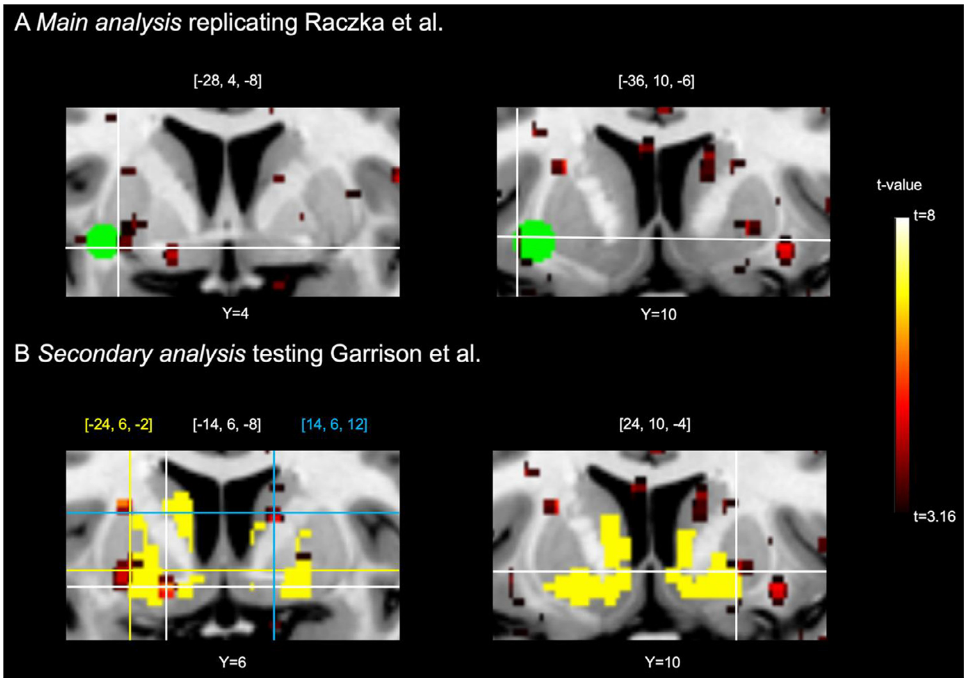

Fig. 4. EPE signals in the *main and secondary fMRI analyses*. A) *Main analysis* testing for activation in the left ventral putamen ROI (derived from Raczka et al., green mask). B) *Secondary analysis* testing for activation in a wider striatal ROI (derived from Garrison et al., yellow mask). All peaks identified (haircrosses) are corrected for multiple comparison using SPM's small volume correction (SVC), $p_{FWE}<0.05$ at voxel level. Display threshold serving to illustrate the anatomical distribution of the signal: p=0.001 unc.

There are no animal studies, to our knowledge, that have performed a comparable mapping of the spatial distribution of either EPE or RPE target regions. However, two recent studies in rats have mapped fMRI activity following electrical and/or optogenetic stimulation of the ventral tegmental area (VTA), the major source of DA neurons in the mesolimbic DA system (Lohani et al., 2017; Brocka et al., 2018). Both studies could not find the fMRI signal to be confined to any specific subregion of the striatum. This is consistent with the rather non-specific striatal distribution of RPE and EPE signals in the human fMRI studies.

Nevertheless, it must be noted that mapping RPE and EPE signals with BOLD-fMRI represents a methodological challenge, due to the vascular nature of the BOLD signal. So, Brocka et al. (2018) have argued that BOLD signal increases in DA target regions in the context of presumed DA neuron activity may not necessarily reflect local neuronal activity changes resulting from DA release, but may also be due to the effects of other neurotransmitters released from non-DA cells or to the effects of DA on the vasculature. This would mean that the BOLD signal is not well suited to determine the precise location of the striatal region that receives the DAergic RPE or EPE signal. In any case, the methods used in the present study cannot establish the DAergic nature of striatal, or other, BOLD signals related to the EPE. Future studies may use pharmacological manipulations to approach this question.

Another interesting finding of the current study is the existence of extra-striatal EPE-related BOLD response. This also corresponds to recent RPE meta-analyses (Garrison et al., 2013; Chase et al., 2015). The most prominent extra-striatal activation, replicated from Raczka et al. (2011), was found in the anterior cingulate cortex, another major output station of DA projections (Moore and Bloom, 1978). It should thus be considered possible that PE processing in appetitive learning also occurs outside the striatum. Areas like the anterior cingulate cortex may give the PE access to working memory space, perhaps in the form of a consciously perceived surprise signal, and thereby inform higher-order or model-based (as opposed to model-free reinforcement) learning systems presumably relying on neocortex (Dayan and Berridge, 2014). Extra-striatal EPE signaling may also help transform extinction learning experiences into memory traces of safety that can be retrieved at later encounters with an extinguished CS and then inhibit the return of conditioned fear responses (Bouton, 2004). It should be noted, though, that the ventromedial prefrontal cortex, which appears to be particularly important for consolidating and retrieving safety memories (Gerlicher et al., 2018), was not among the EPE target areas in our study.

Further methodological limitations of our study are worth mentioning. First, the monotonous decrease predicted by our computational model for the time course of CS+ EPE events (Fig. 1C) raises the possibility that the EPE fMRI regressor might also capture confounding time effects related, for instance, to fatigue or drift. This is counteracted, however, by including CS- events into the same regressor (Fig. 1C) and by fitting the model in one go across all experimental phases, such that a series of unpaired CS+s (extinction phase) is preceded and followed by a series of mostly paired CS+s (acquisition and reacquisition phases), thereby inducing some fluctuation in CS+ outcomes. Second, the presence of continuous moment-to-moment fear ratings probably improved detection of conditioned fear and prediction error modeling, compared to the infrequent retrospective ratings used in the Raczka et al. (2011) study. However, the ratings might also have introduced confounding fMRI signal related to movement that might not be fully captured by including the key presses in the fMRI model. This means that both studies have complementary strengths and weaknesses. The consistency of findings across both studies therefore is the main outcome of the present investigation and can be considered converging evidence for a ventral striatal EPE signal. We emphasize again that the present result is a conceptual, but not a direct replication. Third, we here limited our investigations on the EPE, for which we assumed an appetitive or reward-like nature, but did not address a potential contribution of the aversive learning system and of aversive PE signaling to extinction. We do not claim an exclusive role of the appetitive system and consider it possible that deactivations of the aversive system at the time of UCS omission in extinction (that is, aversive PEs; see, e.g., Kroes et al., 2016) may be a parallel driver of extinction learning.

Taken together, the present report advances the study of human extinction learning by replicating evidence for a role of the ventral putamen in extinction prediction error signaling. It thereby aligns human extinction learning more closely with recent findings in animal extinction learning (Kalisch et al., 2019). A key challenge for future research is to investigate the role of dopamine in these processes.

Data and code accessibility statement

The computational modeling and the analysis of the fear ratings and SCRs and the fMRI analysis were conducted with custom-made Matlab code optimized for MATLAB 2015b and MATLAB 2017b (MathWorks, Inc., Natick, Massachusetts, United States, www.matlab.com) and (in case of fMRI analysis) SPM12 (www.fil.ion.ucl.ac.uk/spm). All multivariate statistics were run with SPSS 23 (IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp). All code was run on a Linux Remote-desktop Server (Intel® Xenon® CPU E5-2690 0, 2.90 GHz processor, 128GB RAM) running on Debian 8. The code is freely available online (osf.io/qf4vj/; see Table S2 for detailed explanations). Raw fMRI data files are not accessible due to data protection rules (GDPR). Statistical images from the single-subject level fMRI models are available on request. The anatomical masks used for region of interest (ROI) analyses (see below) are available at osf.io/qf4vj (see Table S3).

Author contributions

MT, KY and RK designed research; MT performed research; KY, AG, MT and RT analyzed data; MT and RK wrote the paper.

Declaration of Competing interest

The authors declare no financial and non-financial conflicts of interest. R.K. receives advisory honoraria from JoyVentures, Herzlia, Israel.

Acknowledgments

This work was funded by Deutsche Forschungsgemeinschaft (DFG grant CRC 1193, subprojects B01, C01). We thank Jane Garrison and colleagues for providing their RPE mask, Benjamin Meyer for discussions about the experiment, and Anna Katharina Müller for help with data acquisition.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117709.

References

Abraham, AD, Neve, KA, Lattal, KM, 2014. Neurobiology of learning and memory dopamine and extinction : a convergence of theory with fear and reward circuitry. Neurobiol. Learn. Mem. 108, 65–77. Available at: http://dx.doi.org/10.1016/j.nlm.2013.11.007 .

Bouton, ME, 2004. Context and behavioral processes in extinction. Learn. Mem. 11, 485–494.

Chase, HW, Kumar, P, Eickhoff, SB, Dombrovski, AY, 2015. Reinforcement learning models and their neural correlates : an activation likelihood estimation meta-analysis. Soc. Cogn. Affect. Neurosci. 15, 435–459.

Dayan, P, Berridge, KC, 2014. Model-based and model-free Pavlovian reward learning : Revaluation, revision and revelation. Cogn. Affect. Behav. Neurosci..

Desikan, RS, Segonne, F, Fischl, B, Quinn, BT, Dickerson, BC, Blacker, D, Buckner, RL, Dale, AM, Maguire, RP, Hyman, BT, Albert, MS, Killiany, RJ, 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based on regions of interest. Neuroimage 31, 968–980.

Felsenberg, J, Jacob, PF, Walker, T, Barnstedt, O, Edmondson-Stait, AJ, Pleijzier, MW, Otto, N, Schlegel, P, Sharifi, N, Perisse, E, Smith, C, Lauritzen, JS, Costa, M, Jefferis, GSXE, Bock, DD, Waddell, S, 2018. Integration of parallel opposing memories underlies memory extinction. Cell 175, 709–722.

Frazier, JA, Chiu, S, Breeze, JL, Makris, N, Lange, N, Kennedy, DN, Herbert, MR, Bent, EK, Koneru, VK, Dieterich, ME, Hodge, SM, Rauch, SL, Grant, PE, Cohen, BM, Seidman, LJ, Caviness, VS, Biederman, J, 2005. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am. J. Psychiatry 162, 1256–1265. doi:10.1176/appi.ajp.162.7.1256, Available at:.

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

Gerlicher, AM V., Tüscher, O, Kalisch, R, 2018. Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. Nat. Commun. 9, 4294. Available at: http://www.nature.com/articles/s41467-018-06785-y .

Goldstein, JM, Seidman, LJ, Makris, N, Ahern, T, O'Brien, LM, Caviness Jr., VS, Kennedy, DN, Faraone, S V, Tsuang, MT, 2007. Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. Biol. Psychiatry 61, 935–945. doi:10.1016/j.biopsych.2006.06.027, Available at:.

Kalisch, R, Gerlicher, AM V, Duvarci, S, 2019. A dopaminergic basis for fear extinction. Trends Cogn. Sci. 23, 274–277.

Kroes, MCW, Tona, KD, den Ouden, HEM, Vogel, S, van Wingen, GA, Fernández, G., 2016. How administration of the beta-blocker propranolol before extinction can prevent the return of fear. Neuropsychopharmacology 41, 1569–2578.

Lonsdorf, TB, Haaker, J, Kalisch, R, 2014. Long-term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex : a reinstatement study in two independent samples. Soc. Cogn. Affect. Neurosci. 9, 1973–1983.

Lonsdorf, TB, Klingelhöfer-Jens, M, Andreatta, M, Beckers, T, Chalkia, A, Gerlicher, A, Jentsch, VL, Drexler, SM, Mertens, G, Richter, J, Sjouwerman, R, Wendt, J, Merz, C., 2019. Navigating the garden of forking paths for data exclusions in fear conditioning research. eLife 8, e52465.

Luo, R, Uematsu, A, Weitemier, A, Aquili, L, Koivumaa, J, McHugh, TJ, Johansen, JP, 2018. A dopaminergic switch for fear to safety transitions. Nat. Commun. 9, 1–11.

Makris, N, Goldstein, JM, Kennedy, D, Hodge, SM, Caviness, VS, Faraone, S V, Tsuang, MT, Seidman, LJ, 2006. Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr. Res. 83, 155–171. doi:10.1016/j.schres.2005.11.020, Available at:.

Mazaika, P, Hoeft, F, Glover, GH, Reiss, AL, 2009. Methods and software for fMRI analysis for clinical subjects. Neuroimage 47 (Suppl 1), S58.

Moore, RY, Bloom, FE, 1978. Central catecholamine neuron systems: anatomy and physiology of the dopamine systems. Annu. Rev. Neurosci. 1, 129–169.

O'Doherty, J, Dayan, P, Schultz, J, Deichmann, R, Friston, K, Dolan, RJ, 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304, 452–454.

O'Doherty, JP, Dayan, P, Friston, K, Critchley, H, Dolan, RJ, 2003. Temporal difference models and reward-related learning in the human brain. Neuron 28, 329–337.

Pavlov, IP, 1927. Conditioned Reflexes. Oxford University Press, London.

Pearce, JM, Hall, G, 1980. A model for pavlovian learning - variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. 87, 532–552.

Raczka, KA, Mechias, M-L, Gartmann, N, Reif, A, Deckert, J, Pessiglione, M, Kalisch, R, 2011. Empirical support for an involvement of the mesostriatal dopamine system in human fear extinction. Transl. Psychiatry 1, e12. Available at: http://dx.doi.org/10.1038/tp.2011.10 .

Recorla, R, Wagner, AR, 1972. In: Black, A.H., Prokasy, WF (Eds.). Appleton-Century-Crofts, New York.

Salinas-Hernández, XI, Vogel, P, Betz, S, Kalisch, R, Sigurdsson, T, Duvarci, S, 2018. Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. Elife 7, 1–25 e38818.

Schultz, W, 1998. Predictive reward signal of dopamine neurons. J. Neurophysiol. 80, 1–27. Available at: http://www.physiology.org/doi/10.1152/jn.1998.80.1.1 .

Seymour, B, O'Doherty, JP, Koltzenburg, M, Wiech, K, Frackowiak, R, Friston, K, Dolan, R, 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. 8, 1234–1240.