Behavioral/Cognitive

Cortico-Striatal Activity Characterizes Human Safety Learning via Pavlovian Conditioned Inhibition

Patrick A.F. Laing,[1] Trevor Steward,[1,2] Christopher G. Davey,[1] Kim L. Felmingham,[2] Miguel Angel Fullana,[3,4] Bram Vervliet,[5,6] Matthew D. Greaves,[1] Bradford Moffat,[7] Rebecca K. Glarin,[7] and Ben J. Harrison[1]

[1]Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne, Melbourne, Victoria 3053, Australia, [2]Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Victoria 3052, Australia, [3]Adult Psychiatry and Psychology Department, Institute of Neurosciences, Hospital Clinic, Barcelona 08001, Spain, [4]Institut d'Investigacions Biomèdiques August Pi i Sunyer, Centro de Investigación Biomédia en Red de Salud Mental, Barcelona 08036, Spain, [5]Laboratory of Biological Psychology, Faculty of Psychology and Educational Sciences, KU Leuven 3000, Belgium, [6]Leuven Brain Institute, KU Leuven 3000, Belgium, and [7]The Melbourne Brain Centre Imaging Unit, Department of Radiology, University of Melbourne, Melbourne, Victoria 3052, Australia

Safety learning generates associative links between neutral stimuli and the absence of threat, promoting the inhibition of fear and security-seeking behaviors. Precisely how safety learning is mediated at the level of underlying brain systems, particularly in humans, remains unclear. Here, we integrated a novel Pavlovian conditioned inhibition task with ultra-high field (7 Tesla) fMRI to examine the neural basis of safety learning in 49 healthy participants. In our task, participants were conditioned to two safety signals: a conditioned inhibitor that predicted threat omission when paired with a known threat signal (A+/AX-), and a standard safety signal that generally predicted threat omission (BC-). Both safety signals evoked equivalent autonomic and subjective learning responses but diverged strongly in terms of underlying brain activation (P_{FDR} whole-brain corrected). The conditioned inhibitor was characterized by more prominent activation of the dorsal striatum, anterior insular, and dorsolateral PFC compared with the standard safety signal, whereas the latter evoked greater activation of the ventromedial PFC, posterior cingulate, and hippocampus, among other regions. Further analyses of the conditioned inhibitor indicated that its initial learning was characterized by consistent engagement of dorsal striatal, midbrain, thalamic, premotor, and prefrontal subregions. These findings suggest that safety learning via conditioned inhibition involves a distributed cortico-striatal circuitry, separable from broader cortical regions involved with processing standard safety signals (e.g., CS^{-}). This cortico-striatal system could represent a novel neural substrate of safety learning, underlying the initial generation of "stimulus–safety" associations, distinct from wider cortical correlates of safety processing, which facilitate the behavioral outcomes of learning.

Key words: conditioned inhibition; dorsal striatum; prediction error; safety learning; UHF fMRI; vmPFC

Significance Statement

Identifying safety is critical for maintaining adaptive levels of anxiety, but the neural mechanisms of human safety learning remain unclear. Using 7 Tesla fMRI, we compared learning-related brain activity for a conditioned inhibitor, which actively predicted threat omission, and a standard safety signal (CS^{-}), which was passively unpaired with threat. The inhibitor engaged an extended circuitry primarily featuring the dorsal striatum, along with thalamic, midbrain, and premotor/PFC regions. The CS^{-} exclusively involved cortical safety-related regions observed in basic safety conditioning, such as the vmPFC. These findings extend current models to include learning-specific mechanisms for encoding stimulus–safety associations, which might be distinguished from expression-related cortical mechanisms. These insights may suggest novel avenues for targeting dysfunctional safety learning in psychopathology.

Received Oct. 31, 2021; revised May 3, 2022; accepted May 6, 2022.

Author contributions: P.A.F.L., B.M., R.K.G., and B.J.H. designed research; P.A.F.L. performed research; P.A.F.L., T.S., M.D.G., and B.J.H. analyzed data; P.A.F.L. wrote the first draft of the paper; P.A.F.L., T.S., C.G.D., K.L.F., M.A.F., B.V., M.D.G., and B.J.H. edited the paper; P.A.F.L. and B.J.H. wrote the paper; B.M. and R.K.G. contributed unpublished reagents/analytic tools.

This work was supported by National Health and Medical Research Council of Australia Project Grant 1161897 to B.J.H. T.S. was supported by National Health and Medical Research Council of Australia-MRFF Investigator Grant MRF1193736, a BBRF Young Investigator Grant, and the University of Melbourne McKenzie Fellowship. We acknowledge the scientific and technical assistance of the Australian National Imaging Facility, a National Collaborative Research Infrastructure Strategy capability at the Melbourne Brain Center Imaging Unit, University of Melbourne. The multiband fMRI sequence was generously supported by a research collaboration agreement with CMRR, the University of Minnesota; the MP2RAGE works in progress sequence was provided by Siemens Healthineers (Germany). We thank Lisa Incerti for contributions to data collection.

P. A. F. Laing's and B. J. Harrison's present address: Melbourne Neuropsychiatry Center, Level 3, Alan Gilbert Building, 161 Barry Street, Carlton, 3053, Victoria, Australia.

The authors declare no competing financial interests.

Correspondence should be addressed to Patrick A.F. Laing at plaingresearch@gmail.com or Ben J. Harrison at habj@unimelb.edu.au.

https://doi.org/10.1523/JNEUROSCI.2181-21.2022

Introduction

Safety learning builds associations between neutral stimuli and the absence of threat, facilitating the inhibition of fear in safe situations. Impaired safety learning is thought to contribute to the pathophysiology of anxiety-related disorders (van Rooij and Jovanovic, 2019; Grasser and Jovanovic, 2021), leading to a renewed interest in its brain-behavioral basis across species (Fendt et al., 2021). Despite its compelling clinical relevance, safety learning remains understudied in humans, and presents several conceptual and empirical challenges (Laing and Harrison, 2021). For instance, while safety signals elicit fewer fear responses compared with threat signals, the same occurs for neutral stimuli, which predict neither the presence nor absence of threat (Rescorla, 1969). Neurobiological studies suggest, however, that safety is not a neutral state but one that conveys information critical to survival and well-being (Tashjian et al., 2021).

Prevailing neuroimaging evidence for human safety processing stems from differential fear conditioning studies, where an unreinforced stimulus (CS^-) is compared with a conditioned threat stimulus (CS^+). These studies consistently identify increased activity of various regions, such as the ventromedial prefrontal cortex (vmPFC), in discriminating CS^- from CS^+ (Fullana et al., 2016). These activations are distinct from baseline activity or deactivations to threat (Harrison et al., 2017), but it is unclear whether they reflect safety learning *per se*. For example, animal models demonstrate a role for the vmPFC in expressing fear inhibition at test, but less involvement during initial learning (Sarlitto et al., 2018; Kreutzmann et al., 2020). Further, while medial prefrontal cortical pathways from the ventral tegmentum and hippocampus facilitate the postconditioning use of safety information (Meyer et al., 2019; Yan et al., 2019), regions such as the insular cortex and dorsal striatum (caudate and putamen) show learning-specific involvement during conditioning (Rogan et al., 2005; Christianson et al., 2008, 2011; Foilb et al., 2016). These findings indicate important differences between brain systems that acquire safety information via conditioning and those that subsequently express this information in the form of fear inhibition and affective appraisal (Battaglia et al., 2022).

In order to isolate learning-specific mechanisms in the brain, human studies could integrate paradigms more directly informed by fundamental principles from associative learning theory. We have proposed that the "Pavlovian-conditioned inhibition" paradigm leverages these principles to provide an optimal experimental model of safety learning (Laing and Harrison, 2021). In this paradigm, a CS is reinforced alone (A^+), but not reinforced when combined with a second CS (AX^-). The "conditioned inhibitor" thereby indicates threat omission in proximity to threat signals. Nonreinforcement of AX^- evokes a salient mismatch between expected threat-delivery and actual threat omission, inducing a prediction error (Wagner and Rescorla, 1972). Established learning theories predict that this mechanism generates robust links between the inhibitor and threat omission (Schultz and Dickinson, 2000), providing an operational definition of safety learning (Laing and Harrison, 2021). The paradigm has been shown to produce conditioned safety in human behavioral studies (Neumann et al., 1997; Laing et al., 2021) but is yet to be investigated in neuroimaging.

The current study aimed to examine the neural basis of human safety learning via Pavlovian conditioned inhibition combined with 7 Tesla fMRI. First, we compared the neural correlates of a conditioned inhibitor, which actively signaled threat omission (A^+/AX^-), and a standard safety signal, which was passively unreinforced (BC^-). Second, we investigated regions underlying initial learning of "stimulus–safety" associations by contrasting early and late conditioning trials, and comparing responses during conditioning with those during a subsequent test phase. We hypothesized that the conditioned inhibitor would engage subcortical circuitry, particularly striatal and midbrain regions, which have well-established roles in prediction error-based learning (Papalini et al., 2020), whereas the standard safety signal would involve distributed cortical regions, including the vmPFC, linked to the cognitive evaluation of safety. Beyond general gains in spatiotemporal sensitivity, 7 Tesla fMRI was used to assess more stimuli over fewer trials, and to identify small subcortical regions that often evade characterization in standard fMRI.

Materials and Methods

Participants. Forty-nine participants were recruited to the study. All participants met the following eligibility criteria: (1) they were between 18 and 35 years of age, (2) had no current or past diagnosis of mental illness as per screening via the Mini-International Neuropsychiatric Interview (Sheehan et al., 1998), (3) were fluent in English, (4) were not taking any psychoactive medications, and (5) had no contraindications to MRI, including pregnancy. All participants had normal or corrected-to-normal vision and provided written informed consent, following a complete description of study protocol, which was approved by the University of Melbourne Research Ethics Committee. Of the initial sample, two participants did not complete scanning (one because of technical failure; one who discontinued), and a further four were excluded because of excessive head motion (see image preprocessing). The final sample consisted of 43 participants (20 female) with a mean age of 24.35 years (±4.35).

Safety learning task. Participants completed a novel Pavlovian -conditioned inhibition task (Fig. 1), adapted from a prior behavioral paradigm (Laing et al., 2021). A series of geometric figures were used as conditioned stimuli (CS) and the aversive unconditioned stimulus (US) was a 95 dB white noise of 500 ms duration. Intertrial intervals (ITIs) were jittered between 8 and 12 s (mean = 10 s) and featured a white fixation cross in the center of a black screen. Each CS had a duration of 5 s total, first presented for 1 s alone, then joined by a threat expectancy rating scale for a further 4 s, after which the US was delivered or omitted. Threat expectancy ratings were made on a 9 point scale, using a button box in participants' right hand. Concurrent acquisition of expectancy ratings has been shown to enhance physiological measures of learning (Warren et al., 2014). The scale displayed the following labels: "DEFINITELY NO" at the left-most end, "NOT SURE" in the center, and "DEFINITELY YES" at the right-most end, and was automatically centered at "NOT SURE" on each new presentation. To enhance inhibitory learning and support the analysis of safety signal responses, the task included a 100% reinforcement rate for all CS^+. This choice ensured that A^+ maintained a robust threat association, such that US omission following AX^- would reliably violate participants' threat expectation (Lysle and Fowler, 1985; Harris et al., 2014; Laing and Harrison, 2021). While partial CS^+ reinforcement is typically used to avoid the US confounding CS^+ responses (Fullana et al., 2016), this was unnecessary in the current study, which was designed to target safety signals exclusively. Overlap between brain response to A^+ and AX^- are illustrated alongside each of the main contrasts (see fMRI analyses and Results).

The experiment featured conditioning and test phases. Six CS configurations were presented in conditioning: A^+, AX^-, BC^-, D^+, AD^+, and X^- (Fig. 1). AX^- was the conditioned inhibitor, which was compared with the control safety signal BC^-, hereafter referred to as the standard safety signal. Each CS was presented for 6 trials, in pseudorandomized order, resulting in 36 trials. The test phase consisted of three CS configurations (AD^+, DC^-, DX^-) also for 6 trials, resulting in 18 trials total. To ensure that test responses were not the product of conscious instruction

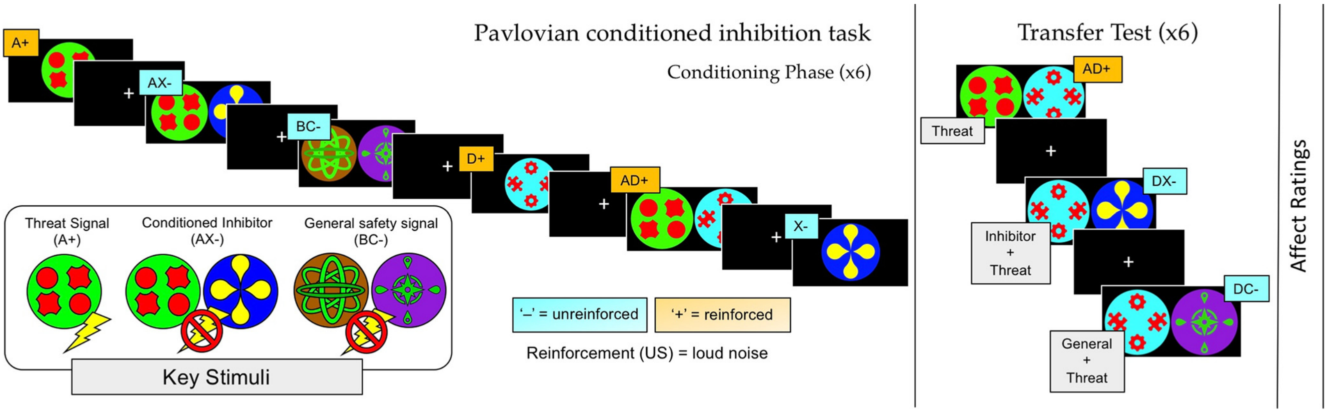

Figure 1. Pavlovian-conditioned inhibition fMRI task. Adapted from Laing et al. (2021). Six stimulus configurations were presented during conditioning. The transfer test occurred immediately following conditioning and featured the two safety signals (inhibitor X and standard C) combined with a conditioned threat cue (D$^+$). Between each CS trial, the ITI (8-12 s) featured a fixation cross. Participants assigned affective ratings for stimuli X, C, D, and the fixation cross immediately following the test phase, providing measures of subjective positive affect and anxious arousal.

(see Mechias et al., 2010), the test phase commenced immediately after conditioning with no signaled interlude. At test, the inhibitor (X) and standard safety signal (C) were combined with the threat signal D$^+$ and not reinforced (DX$^-$, DC$^-$), while AD$^+$ was reinforced throughout. To control for the influence of presentation order, approximately half of participants were presented with DX$^-$ on their first test trial, with the others presented with DC$^-$ (Laing et al., 2021). Following the task's final phase, participants rated the degree to which the X, C, D, and the fixation cross stimuli, respectively, evoked changes in positive-negative affect and anxious arousal via standardized self-assessment manikins (Bradley and Lang, 1994). The task was programmed in E-Prime (Psychology Software Tools) and was presented on a 32 inch LCD BOLD screen (Cambridge Research Systems) visible via a reverse mirror mounted on the head coil. Noise bursts (US) were delivered via Sensimetrics Insert Earphones (S15 model, Sensimetrics), which also provided passive noise cancellation ($\sim$30 dB). Participants' responses were registered with a 2-button LS-PAIR Lumina response pad (Cedrus), which they were familiarized with before scanning.

Skin conductance responses (SCRs). Skin conductance was recorded using MRI-compatible finger electrodes (Ag/AgCI) fitted with conductance gel (0.5% saline) to the intermediate phalanges of the index and middle finger of participants' left hands. Fingers were cleaned with alcohol wipes before the attachment of electrodes. The signal was amplified and sampled at 1000 Hz using PowerLab version 8.0 (ADInstruments), and recording was triggered concurrently with the beginning of the experiment and the functional imaging sequence. The Psychophysiological Modeling Toolbox (PsPM) (Bach et al., 2018; Bach and Melinscak, 2020) in MATLAB (version 9.4, The MathWorks) was used for preprocessing and modeling of SCRs. SCR artifacts were detected via a semiautomated process. A custom program identified time series containing: (1) signal increases $>$20% per second, (2) signal decreases $>$10% per second, or (3) absolute changes $>$0.075 µS per millisecond. The SC data of 21 subjects were flagged for review based on these criteria. Seventeen were excluded after manual review, in cases where artifacts reflected pervasive signal distortion, and four had artifacts removed in PsPM before further analysis. A further 7 subjects' SCR data were comprised by technical issues, resulting in a final SCR sample of $N = 24$. Following artifact removal, SC data were filtered with a 10 ms median filter followed by a first-order bidirectional bandpass Butterworth filer (cutoff frequencies 0.0159-5 Hz) and downsampled to 10 Hz. Dynamic causal modeling was implemented via the PsPM toolbox to provide trial-by-trial estimates of sympathetic nervous system activity, which were represented by the flexible CS-evoked SCR extracted from each subject's model.

Image acquisition. Imaging was performed on a 7T research scanner (Siemens Healthcare) equipped with a 32 channel head coil (Nova Medical). The functional sequence consisted of a multiband (6 times) and GRAPPA (2 times) accelerated GE-EPI sequence in the steady state (TR, 800 ms; TE, 22.2 ms; pulse/flip angle, 45°; FOV, 20.8 cm; slice thickness [no gap], 1.6 mm; 130 × 130-pixel matrix; 84 interleaved axial slices aligned to AC-PC line) (Setsompop et al., 2012). The total sequence time was 16 min and 10 s, corresponding to 1202 whole-brain EPI volumes. A T1-weighted high-resolution anatomic image (MP2RAGE) (Marques et al., 2010) was acquired for each participant to assist with functional time series coregistration (TR = 5000 ms; TE, 3.0 ms; inversion times, 700/2700 ms; pulse/flip angles, 4/5°; FOV, 24 cm; slice thickness [no gap], 0.73 mm; 330 × 330 pixel matrix; 84 sagittal slices aligned parallel to the midline). The total sequence time was 7 min and 12 s. To assist with head immobility, foam-padding inserts were placed on either side of the participants' head. Cardiac and respiratory recordings were sampled at 50 Hertz (Hz) using a Siemens (Bluetooth) pulse-oximeter and respiratory belt. Information derived from these recordings were used for physiological noise correction.

Image preprocessing. Imaging data were preprocessed using Statistical Parametric Mapping (SPM) 12 (version 7771, Welcome Trust Center for Neuroimaging) within a MATLAB 2019b environment (The MathWorks). Motion correction was performed by realigning each subject's time series to the mean image, and all images were resampled using fourth Degree B-Spline interpolation. Individualized motion parameters were estimated via Motion Fingerprint (Wilke, 2012) to account for head motion, and participant data were excluded if mean total scan-to-scan voxel displacement exceeded 1.6 mm (one voxel), resulting in the exclusion of four participants. Each participant's anatomic image was coregistered to their respective mean functional image, segmented, and normalized to the International Consortium of Brain Mapping template using the unified segmentation plus DARTEL approach. Smoothing was applied with a 3.2 mm^3 FWHM Gaussian kernel to increase anatomic precision. Physiologic noise was modeled at the first level using the PhysIO Toolbox (Kasper et al., 2017). This toolbox applies noise correction to fMRI sequences using physiological recordings and is shown to enhance BOLD signal sensitivity and temporal signal-to-noise ratio at 7T (Reynaud et al., 2017). The Retrospective Image-based Correction function (Glover et al., 2000) was applied to model periodic effects of heartbeat and respiration on BOLD signals. The respiratory response function (Birn et al., 2008), convolved with respiration volume per time, was used to model low-frequency signal fluctuations arising from changes in depth and rate of breath. Heart rate variability was convolved with a cardiac response function (Chang et al., 2009) to account for BOLD variances because of heart rate-dependent changes in blood oxygenation. Individualized DARTEL tissue maps segmented from each participant's respective anatomic scan were used to apply aCompCor, which models negative BOLD signals using principal components derived from white matter and CSF (Behzadi et al., 2007).

fMRI analyses. Each participant's preprocessed time-series was included in a first-level SPM GLM analysis, which specified the onsets of each CS event type (grouped as separate conditions for "early"/first three trials and "late"/last 3 trials) in each task phase to be convolved with canonical HRF. The fixation-cross ITI periods throughout whole task served as the implicit baseline. A high-pass filter (1/128 s) accounted for low-frequency noise, while temporal autocorrelations were estimated with an autoregressive model. Contrast images were estimated for each CS condition against the implicit baseline and were carried forward to the group level using the summary statistics approach to random-effects analyses.

As our first aim was to characterize overall differences in brain response to inhibitory versus standard safety signals, we compared all trials of the conditioned inhibitor and standard safety signal during the conditioning phase (AX$^-$ vs BC$^-$). Second, to identify brain regions where activity differed as a function of learning-specific mechanisms, we compared early versus late trials for each safety signal, respectively (e.g., AX$^-_{early}$ vs AX$^-_{late}$). Third, to further separate learning-specific activity from general stimulus processing, or the expression of safety, we compared brain responses for each of the safe signals during the conditioning phase relative to the test phase (e.g., AX$^-$ vs DX$^-$). We also analyzed differences between the two safety signals within the test phase, when they were each combined with a CS$^+$ (e.g., DX$^-$ vs DC$^-$). For contrasts involving the conditioned inhibitor (AX$^-$), we have provided extended results (see Figs. 3*B*, 4*B*, 5*B*) to illustrate its overlap with brain responses to the simple conditioned threat stimulus (A$^+$), which are included to demonstrate to what degree AX$^-$ activations reflected the threat value of A$^+$ relative to learning processes attributable to AX$^-$. To confirm that safety signal responses were not confounded by diverging prediction error magnitudes for AX$^-$ and BC$^-$, US omission onsets were modeled and compared with the main (CS onset) analyses. It was found that, for each main contrast, modeling of US omission resulted in findings equivalent to the original analyses. In other words, brain activations for the main contrasts in this report reflected CS evoked responses, which were not confounded by differences in US omission response.

Statistical significance was estimated with a whole-brain false-discovery rate (FDR)-corrected threshold ($P_{\mathrm{FDR}} < 0.05$) and a 5-voxel cluster-extent threshold ($K_E \geq 5$). As noted in past 7 Tesla studies (Sclocco et al., 2018), family-wise error (FWE) correction on high-resolution data can approach a Bonferroni correction (i.e., be overly conservative), inflating the risk of Type II error (false negatives). For these reasons, we adopted the FDR thresholding approach in combination with a reduced smoothing kernel. Analyses of threat expectancy ratings and SCR were performed by separating responses into early and late phases (three trials each) consistent with the imaging analyses, and subjected to repeated-measures ANOVAs and *post hoc t* tests. One-sample *t* tests were used to compare differences in subjective ratings (valence and arousal) that were collected at the end of the task.

Results

Behavioral results

Safety learning: conditioned inhibitor versus standard safety signal

Conditioned response measures (threat expectancy ratings and SCR) were analyzed to assess changes in threat response across the conditioning phase. Threat expectancy ratings for the threat cue (A$^+$), conditioned inhibitor (AX$^-$), and standard safety signal (BC$^-$) showed a main effect of stimulus ($F_{(1,45)}$ = 421.99, $p < 0.001$, $\eta_p^2 = 0.90$) and phase (early, late, $F_{(1,45)}$ = 69.04.45, $p < 0.001$, η_p^2 = 0.61), and a stimulus × phase interaction ($F_{(1,45)}$ = 162.69, $p < 0.001$, η_p^2 = 0.78). *Post hoc* tests showed higher threat expectancy for the threat cue compared with the inhibitor (A$^+$ > AX$^-$: mean = 59.46, t = 23, d = 3.39, $p < 0.001$, 95% CI [59.46, 65.76]) and standard safe signal (A$^+$ > BC$^-$: mean = 69.48, t = 23, d = 3.96, $p < 0.001$, 95% CI [69.46, 75.76]). The inhibitor showed higher averaged threat expectancy ratings than the standard safety signal (AX$^-$ > BC$^-$: mean = 10, t = 5.93, d = 0.87, $p < 0.001$, 95% CI [3.69, 16.31]), driven by significant differences in early (mean = 14.85, t = 4.87, $p < 0.001$, 95% CI [5.76, 23.95]), but not late (mean = 5.15, t = 1.69, p = 0.17) conditioning trials (Fig. 2*A*). SCRs (Fig. 2*B*) showed main effects of stimulus ($F_{(1,23)}$ = 6.14, p = 0.004, η_p^2 = 0.21), phase (early, late, $F_{(1,23)}$ = 5.55, p = 0.027, η_p^2 = 0.19), but no interaction (p = 0.84). *Post hoc* tests showed no overall difference in SCRs for AX$^-$ and BC$^-$ (p = 0.95), but significant differences with A$^+$ (A$^+$ > AX$^-$: mean = 0.272, t = 3.00, d = 6.12, p = 0.013, 95% CI [0.047, 0.49]; A$^+$ > BC$^-$: mean = 0.278, t = 3.07, d = 6.26, p = 0.011, 95% CI [0.053, 0.50]), and an overall difference between early and late phase SCRs (mean = 0.129, t = 2.36, d = 0.48, p = 0.027, 95% CI [0.016, 0.242]). Subjects also showed significant threat-safety discrimination when averaging ratings (mean = 52.73, d = 2.451, $p < 0.001$, 95% CI [46.34, 59.12]) and SCRs (mean = 0.10, d = 0.592, $p < 0.008$, 95% CI [0.027, 0.16]) for all CS$^+$ and CS$^-$ stimuli. Transfer test responses showed mixed results. Threat expectancy showed main effects of stimulus ($F_{(1,45)}$ = 894.58, $p < 0.001$, η_p^2 = 0.95), phase ($F_{(1,45)}$ = 28.25, $p < 0.001$, η_p^2 = 0.39), and their interaction ($F_{(1,45)}$ = 16.48, $p < 0.001$, η_p^2 = 0.27). Expectancy was robustly decreased for the conditioned inhibitor (AD$^+$ > DX$^-$; mean = 88.08, t = 36.74, d = 5.42, $p < 0.001$, 95% CI [82.23, 93.93]) and standard safety signal (AD$^+$ > DC$^-$; mean = 87.57, t = 36.52, d = 5.38, $p < 0.001$, 95% CI [81.72, 93.42]) relative to the threat compound, but showed no significant differences between them (DX$^-$ vs DC$^-$). In SCRs, no main effects were identified for stimuli (AD$^+$, DX$^-$, DC$^-$; p = 0.27) or phase (early, late, p = 0.96). In sum, behavioral measures indicated that each safety signal evoked equivalent decreases in behavioral threat response during learning, but did not differ in responses evoked at test.

Subjective ratings: positive affect and anxious arousal

Compared with a CS$^+$, the conditioned inhibitor and standard safety signal elicited greater positive affect (X > D: mean = 62.5, t = 7.58, d = 1.12, $p < 0.001$, 95% CI [40.43, 84.57]; C > D: mean = 60.33, t = 7.32, d = 1.08, $p < 0.001$, 95% CI [38.26, 82.39]; Fig. 2*E*) and lower anxious arousal (X > D: mean = 51.09, t = 10.38, d = 1.53, $p < 0.001$, 95% CI [37.90, 64.27]; C > D: mean = 50.54, t = 10.26, d = 1.51, $p < 0.001$, 95% CI [37.36, 63.73]; Fig. 2*F*). Further, both safety signals also evoked greater positive affect (X > fixation cross: mean = 26.09, t = 3.17, d = 0.47, p = 0.006, 95% CI [4.02, 48.15]; C > fixation cross: mean = 23.92, t = 2.90, d = 0.43, p = 0.009, 95% CI [1.87, 45.98]; Fig. 2*E*) and lower arousal (fixation cross > X: mean = 14.67, t = 2.98, d = 0.44, p = 0.01, 95% CI [1.49, 27.86]; fixation cross > C: mean = 14.13, t = 2.87, d = 0.42, p = 0.01, 95% CI [0.95, 27.31]; Fig. 2*F*) relative to ratings of the fixation cross, which served as a putative baseline comparison. The inhibitor and standard safety signal did not differ from one another on either affective measure. In summary, both safety signals accumulated high positive affect and low anxious arousal following conditioning, consistent with previous findings (Harrison et al., 2017; Laing et al., 2021).

Imaging results

Differential safety responses: conditioned inhibitor versus standard safety signal

We first estimated differential neural responses to the conditioned inhibitor versus standard safety signal (AX$^-$ > BC$^-$) across all conditioning trials. As shown in Figure 3, this direct comparison identified significantly greater activation to the conditioned inhibitor in regions including the anterior insular cortex bilaterally, extending to the dorsal putamen in the right hemisphere; the left caudate body extending to dorsal putamen, the left ventrolateral cerebellum (Crus II), and right posterior dorsolateral PFC (Fig. 3). Figure 3*B* highlights the overlap between the differential response to the conditioned inhibitor versus standard safety signal and the response to conditioned threat alone (A$^+$).

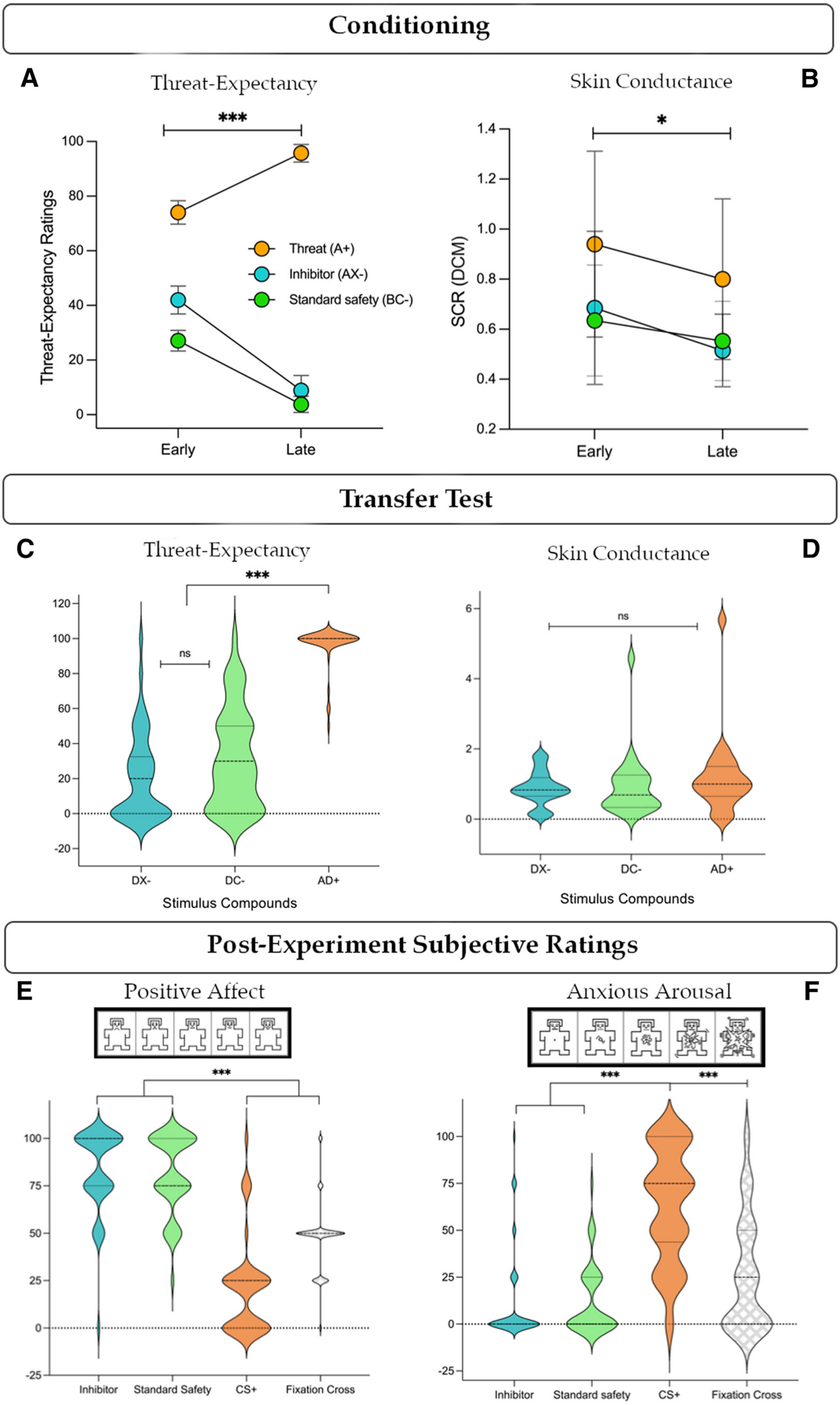

Figure 2. Behavioral responses. ***A***, Threat expectancy ratings showed a significant decrease across conditioning for each safety signal, with higher ratings during early trials for the inhibitor versus standard, and no difference during late conditioning. ***B***, SCRs showed no difference between stimuli (inhibitor vs standard), but a significant decrease from early to late conditioning for each stimulus. ***C***, Both safety signals inhibited threat expectancy at test compared with the threat compound AD^{+} but did not differ from one another. ***D***, SCRs showed no differences at test.

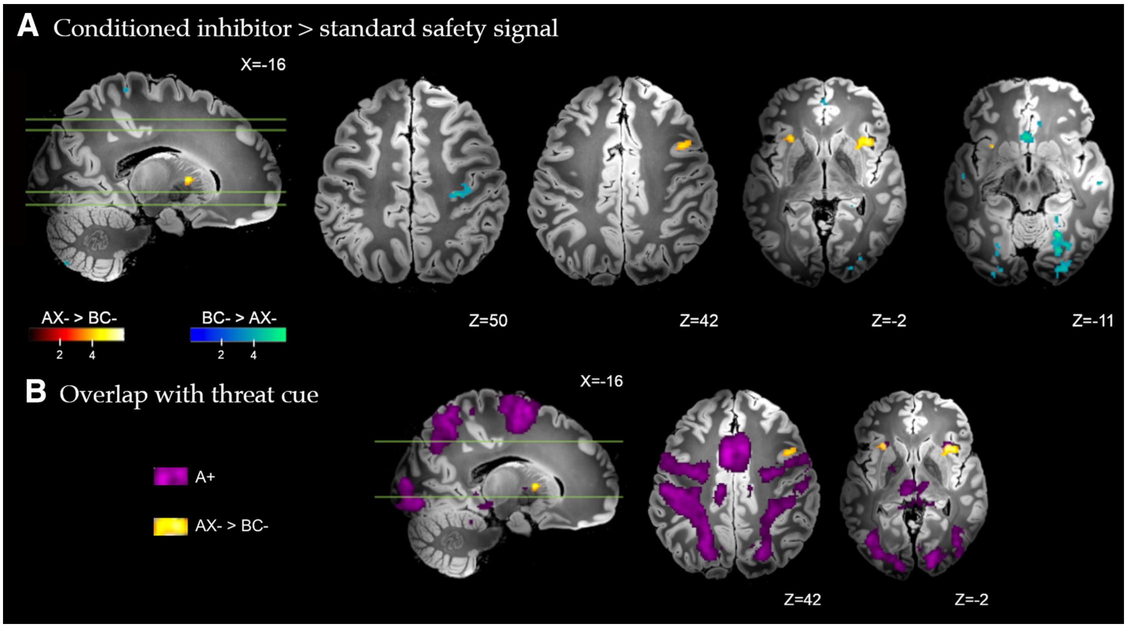
A Conditioned inhibitor > standard safety signal
X=-16
AX- > BC-
2 4
BC- > AX-
2 4
Z=50
Z=42
Z=-2
Z=-11
B Overlap with threat cue
X=-16
A+
AX- > BC-
Z=42
Z=-2

Figure 3. ***A***, Brain regions with significant differential response to the conditioned inhibitor (AX$^-$ > BC$^-$) versus the standard safety signal (BC$^-$ > AX$^-$) across all trials of conditioning. Whole-brain FDR-corrected ($p < 0.05$) results are displayed on a high-resolution anatomic template in MNI space. ***B***, Partial overlap between the differential response to the conditioned inhibitor versus standard safety signal (AX$^-$ > BC$^-$, yellow), and the response to the conditioned threat alone (A$^+$, purple) in activation of the anterior insular cortex.

Partial overlap was noted with regards to activation of the anterior insular cortex. The standard safety signal compared with conditioned inhibitor (BC$^-$ > AX$^-$) evoked significantly greater activation of the ventromedial PFC, spanning posterior (subgenual) and anterior (frontopolar) subregions, retrosplenial posterior cingulate cortex, right posterior hippocampus, basal forebrain (~medial forebrain bundle), right posterior primary motor cortex, cerebellum (VI and VIIb) medial and lateral visual association cortex, spanning the fusiform, lingual gyrus, and occipital pole. Complete results are provided in Table 1. All GLM results are presented on the Synthesized_FLASH25 (500 μm, MNI space) *ex vivo* template (Edlow et al., 2019).

Safety learning dynamics during conditioning

Safety learning dynamics were analyzed via contrasts of early versus late trials within each safety stimulus during conditioning. As shown in Figure 4, early versus late trials for the conditioned inhibitor (AX$^-_{early}$ > AX$^-_{late}$) were associated with significantly greater activation of the caudate body extending to dorsal anterior putamen and globus pallidus (external), the midbrain substantia nigra, ventral lateral and intralaminar thalamic nuclei, dorsal-mid cingulate and dorsal premotor cortex, and the cerebellum (IX). There was no significantly greater activation identified for the conditioned inhibitor during late versus early trials (AX$^-_{late}$ > AX$^-_{early}$). Figure 4*B* illustrates that there was minimal overlap between the differential response to the early versus late conditioned inhibitor trials and the response to the conditioned threat alone (A$^+$).

←

E, Subjective affective ratings reliably discriminated safety signals with high positive affect from the CS$^+$ and neutral fixation cross. ***F***, Similarly, safety signals elicited low appraisal of anxious arousal relative to high ratings of CS$^+$, and moderate ratings of the fixation cross. AX$^-$, conditioned inhibitor; BC$^-$, standard safety signal; DX$^-$, inhibitor + threat cue; DC$^-$, standard safety signal + threat cue; AD$^+$, threat compound; DCM, dynamic causal model for SCR (via PsPM). Error bars indicate 95% CIs. *$p < 0.05$. ***$p < 0.001$. ns, $p > 0.05$.

Table 1. Differential safety responses: conditioned inhibitor versus standard safety signal[a]

	MNI coordinates				
Regions	*x*	*y*	*z*	K$_E$	*Z*
Inhibitor > standard safety (AX$^-$ > BC$^-$)					
Anterior insular cortex	40	18	−3	85	5.43
Caudate body	−16	6	6	11	5.02
Posterior cerebellum (Crus II)	−37	−75	−46	9	4.59
Dorsal anterior putamen	−27	13	10	5	4.51
Posterior middle frontal gyrus	42	8	40	6	4.51
Anterior insular cortex	−29	21	0	20	4.49
Standard safety > inhibitor (BC$^-$ > AX$^-$)					
Fusiform gyrus	24	−59	−11	276	5.97
Postcentral gyrus	29	−26	48	40	4.93
Subgenual cingulate cortex	−3	22	−11	50	4.92
Basal forebrain	2	0	−8	7	4.61
Middle temporal gyrus	58	−18	−13	17	4.59
Fusiform gyrus	40	−40	−21	11	4.55
Lateral occipital cortex	29	−83	2	30	4.54
Occipital pole	24	−91	2	11	4.45
Fusiform gyrus	27	−86	−13	43	4.37
Precentral gyrus	37	−21	53	16	4.36
Hippocampus	24	−38	−3	5	4.31
Occipital pole	27	−90	13	17	4.31
Posterior cingulate cortex	−3	−51	10	6	4.28
Cerebellum (VI)	26	−48	−21	5	4.26
Occipital pole	−26	−93	3	12	4.25
Subgenual cingulate cortex	2	29	−6	6	4.21
Fusiform gyrus	−27	−83	−16	17	4.17
Cerebellum (VIIb)	−16	−80	−53	7	4.17
Lingual gyrus	22	−45	−11	8	4.08
Ventromedial frontal cortex	0	50	−5	16	3.95

[a] K$_E$, cluster size in number of voxels; *Z*, SPM *Z* scores. Coordinates reported in MNI space. Whole-brain contrasts estimated at $P_{FDR} < 0.05$.

For the standard safety signal (BC$^-_{early}$ > BC$^-_{late}$), early trials evoked significantly greater activation of posterior dorsolateral PFC, superior and intraparietal cortex, precuneus (dorsal and ventral subareas), and medial and lateral visual association

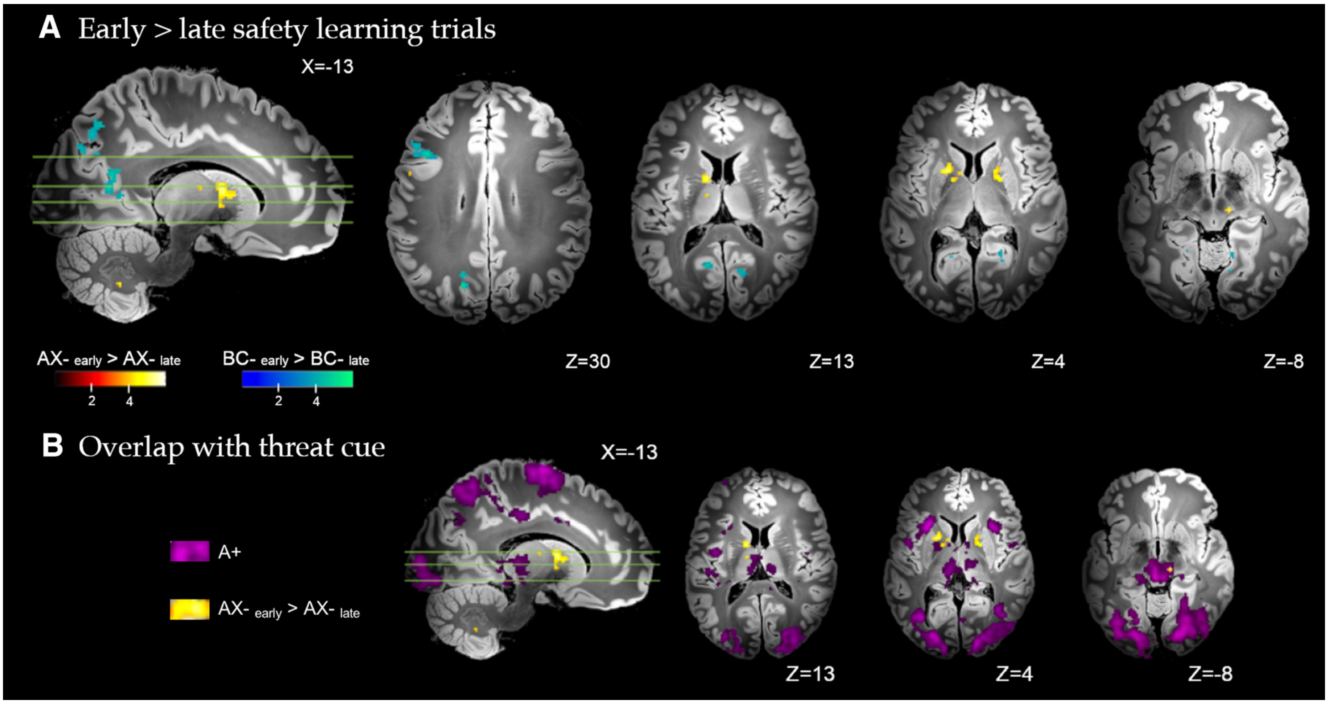

Figure 4. ***A***, Brain regions with significant differential response to early versus late conditioning trials, for the conditioned inhibitor ($AX^{-}_{early} > AX^{-}_{late}$) and standard safety signal ($BC^{-}_{early} > BC^{-}_{late}$). Whole-brain FDR-corrected ($p < 0.05$) results are displayed on a high-resolution anatomic template in MNI space. ***B***, Minimal overlap observed between the differential response to the early versus late conditioned inhibitor trials ($AX^{-}_{early} > AX^{-}_{late}$, yellow), and the response to the conditioned threat alone (A^{+}, purple).

cortex, spanning the fusiform, lingual gyrus, and occipital pole. There was no significantly greater activation identified for the standard safety signal during late versus early trials ($BC^{-}_{late} > BC^{-}_{early}$). Complete results are provided in Table 2.

Conditioning versus test

As shown in Figure 5, direct comparison of the conditioned inhibitor across the conditioning and test phases ($AX^{-} > DX^{-}$) identified significantly greater activation of the bilateral dorsal anterior putamen and globus pallidus (external), the left caudate (tail), midbrain substantia nigra and periaqueductal gray, bilateral dorsal premotor cortex, and precuneus. Complete results are provided in Table 3. Figure 5*B* highlights the overlap between the differential response to the conditioned inhibitor at training versus test and the response to the conditioned threat alone (A^{+}). Partial overlap was noted with regards to activation of the left premotor cortex. There was no significantly greater activation identified for the conditioned inhibitor at test compared with conditioning ($DX^{-} > AX^{-}$). For the standard safety signal, significantly greater activation of the anterior fusiform and lingual gyrus was identified during conditioning compared with test ($BC^{-} > DC^{-}$). There was no significantly greater activation identified for the standard safety signal at test compared with conditioning ($DC^{-} > BC^{-}$). As a further test of safety expression, we contrasted test-phase responses to the inhibitor and standard safety signal directly (DX^{-} vs DC^{-}). No clusters exceeded the $P_{FDR} < 0.05$ threshold ($K_E \geq 5$ voxels).

Discussion

This study combined 7 Tesla fMRI with Pavlovian-conditioned inhibition to characterize the neural basis of safety learning in humans. We compared two conditioned safety signals: a conditioned inhibitor, which preceded threat omission in compound with a CS^{+} (A^{+}, AX^{-}); and a standard safety signal, a stimulus compound that was unreinforced (BC^{-}) without threat proximity. Supporting our general hypotheses, safety learning via conditioned inhibition evoked prominent subcortical activations spanning the dorsal striatum, thalamus, and midbrain, together with dorsal prefrontal and premotor cortex regions. Conversely, standard safety signal processing was associated with greater engagement of distributed frontal and occipital-parietal regions, which have previously been linked to the cognitive appraisal of safety information.

Compared with the standard safety signal, the conditioned inhibitor selectively engaged the dorsal anterior striatum, anterior insular, and posterior dorsolateral PFC, with further analyses implicating extended activation of midbrain, thalamic, and premotor cortex subregions during early safety learning and when comparing conditioning with test phase responses. We observed that activation of the anterior insular, premotor, and midbrain subregions by the conditioned inhibitor overlapped with responses to the conditioned threat signal (A^{+}), whereas most subcortical regions, particularly the dorsal striatum, were nonoverlapping. Consequently, the conditioned inhibitor's neural response encompassed both threat-responsive regions and learning-related cortico-striatal activation. The former may reflect a key component of inhibitory learning: the direct conflict between threat and safety information. Conditioned inhibition depends on a reasonable level of threat expectation (Lysle and Fowler, 1985; Harris et al., 2014), requiring that inhibitory stimuli predict safety when threat is otherwise expected (Sosa and Ramírez, 2019). Insular activity may reflect modulation of threat value (Sharvit et al., 2018; Teckentrup et al., 2019), analogous to its role in fear extinction (Fullana et al., 2018). Interestingly, the nonoverlapping cortico-striatal regions associated with conditioned inhibition are constituents of well-established cortico-basal ganglia pathways, which hold deep intrinsic connectivity in primate neuroanatomy. For instance, the dorsal striatum

Table 2. Safety learning dynamics: early versus late conditioning[a]

	MNI coordinates				
Regions	*x*	*y*	*z*	K_E	*Z*
Conditioned inhibitor ($AX^-_{early} > AX^-_{late}$)					
Caudate body	−16	6	10	213	5.68
Midbrain (substantia nigra)	11	−22	−8	10	5.13
Pallidum	19	2	3	128	4.90
Thalamus (intralaminar)	8	−22	−2	9	4.61
Thalamus (ventral lateral nucleus)	−11	−10	13	10	4.44
Cerebellum (IX)	−13	−56	−43	7	4.35
Mid cingulate cortex	−6	19	43	8	4.30
Precentral gyrus	−53	5	29	9	4.30
Standard safety ($BC^-_{early} > BC^-_{late}$)					
Superior frontal gyrus	−21	8	51	78	5.12
Superior parietal lobule	−10	−69	50	179	5.10
Lingual gyrus	19	−45	−11	31	5.06
Precuneus	−18	−70	30	33	5.03
Ventral precuneus	−11	−62	11	93	4.87
Middle frontal gyrus	38	22	26	30	4.81
Middle frontal gyrus area	−42	18	30	120	4.78
Lingual gyrus	14	−67	−5	49	4.75
Precuneus	−18	−66	22	34	4.71
Ventral precuneus	21	−56	8	62	4.65
Precuneus	−18	−64	29	14	4.57
Superior parietal lobule	11	−69	48	33	4.55
Ventral precuneus	−16	−58	6	25	4.44
Supracalcarine cortex	6	−77	18	17	4.37
Intracalcarine sulcus	16	−69	11	75	4.36
Lateral occipital cortex	19	−82	27	12	4.24
Superior parietal lobule	13	−46	58	11	4.18
Fusiform gyrus	−24	−46	−13	7	4.15
Lingual gyrus	13	−56	−8	10	4.13
Fusiform gyrus	−30	−42	−16	5	4.12
Occipital pole	−11	−91	26	6	4.10
Precuneus	16	−64	38	9	4.06
Precuneus	16	−66	22	7	4.02
Cuneus	13	−85	29	6	4.00
Precuneus	11	−67	27	13	3.96
Lingual gyrus	24	−51	−10	5	3.92
Lingual gyrus	−19	−54	−11	6	3.88
Intraparietal sulcus	−21	−75	43	7	3.78

[a] K_E, cluster size in number of voxels; *Z*, SPM *Z* scores. Coordinates reported in MNI space. Whole-brain contrasts estimated at $P_{FDR} < 0.05$.

receives premotor cortical region inputs and relays information to the ventrolateral thalamus by way of the pallidum and SNc (Alexander et al., 1986; Alexander and Crutcher, 1990). Anatomical characteristics of this circuit are also consistent with intrinsic premotor cortex connectivity with the dorsal striatum and SNc in human fMRI (Di Martino et al., 2008; Choi et al., 2012).

The engagement of these cortico-striatal regions by the inhibitor, but not the standard safety signal, is consistent with their primary psychological difference: the manipulation of expectancy violation (Laing et al., 2021). In our task, early trials of the conditioned inhibitor (AX^-) ought to evoke prediction error-like mismatches, wherein threat-anticipation elicited by A^+ is violated by threat omission following AX^- (Laing and Harrison, 2021). These trials demand reconciliation of expected (threat) and actual (safety) outcomes, but as the stimulus–safety contingency is repeated, the need for error-processing and value-updating is minimized. Consequently, the observed temporal shift in dorsal striatal activity, in tandem with SNc, is consistent with the role of prediction error in associative learning (Schultz et al., 2003; Pauli et al., 2015). For instance, dorsal striatal nuclei show elevated prediction error responses in early conditioning, which decreases as contingencies are learned (Valentin and O'Doherty, 2009; Cooper et al., 2012; Oyama et al., 2015), while also updating conditioned responses to the CS (Yoshizawa et al., 2018). The dorsal striatum receives direct midbrain dopaminergic projections from the SNc (Lanciego et al., 2012), which signal prediction errors by firing in response to unexpected stimulus-delivery (or omission), and decreasing in magnitude as expectations and outcomes are aligned (Waelti et al., 2001; Eshel et al., 2016; Salinas-Hernández et al., 2018). As a result, safety learning via conditioned inhibition appears to evoke a more direct instantiation of these active error-corrective learning systems compared with the standard safety signal, which does not directly incur an expectancy-outcome mismatch. Our findings thereby implicate a specific role for these associative learning mechanisms in the domain of human safety learning, which has been primarily associated with prefrontal cortical mechanisms (Tashjian et al., 2021).

Brain regions associated with the standard safety signal (BC^-) overlap with the results of differential fear conditioning studies, which report consistent involvement of the vmPFC, hippocampus, and posterior cingulate cortex in safety versus threat discrimination (Fullana et al., 2016). However, as noted earlier, these experiments were not designed to discriminate mechanisms of safety learning from safety expression and have likely conflated these processes. An acquisition–expression dissociation is well supported in animal models, which demonstrate learning-specific roles for the striatum and insular (Rogan et al., 2005; Foilb et al., 2016), and expression-specific contributions of hippocampal and prefrontal cortical regions (Meyer et al., 2019; Yan et al., 2019; Kreutzmann and Fendt, 2020; Kreutzmann et al., 2020). In humans, learned safety should facilitate experiences of positive affect and inhibition of fear behaviors (Zhang et al., 2015), each requiring recall of safety information from memory. Affective valuation, response inhibition, and recall processes all converge on these standard safety-processing regions (Roy et al., 2012; Harrison et al., 2017; Hennings et al., 2020; Hermann et al., 2020). The vmPFC has a notably multifaceted functional role, modulating both fear and safety expression, rather than unidirectional threat inhibition (Tashjian et al., 2021; Battaglia et al., 2022). For instance, it is causally implicated in acquiring conditioned responses to a CS^+ (Battaglia et al., 2020), and in inhibiting responses in the face of unlearned safety signals (e.g., familial attachment figures) (Eisenberger et al., 2011). Thus, a "dual systems" perspective on human safety learning may be warranted, featuring (1) a cortico-striatal system that encodes initial safety associations, and (2) an expanded cortical system supporting the appraisal and expression (or "use") of learned safety information. Synchronization between these systems likely requires interactions extending beyond conventional vmPFC-oriented circuitry, involving midbrain, striatal, and prefrontal integration of associative learning and safety-retention across time (Raczka et al., 2011; Esser et al., 2021). More targeted research could characterize how the neural dynamics of prediction error informs CS evoked safety responses, and differentiate safety-specific prediction error signals from other learning signals (e.g., reward prediction error) (Corlett et al., 2022).

Although few, if any, existing fMRI studies have translated the Pavlovian conditioned inhibition model to human fear conditioning, it has recently been reported in a study of appetitive conditioning, where the inhibitor predicted reward omission, and evoked dorsal-striatal responses similar to those observed here (Mollick et al., 2021). Consistent with associative learning

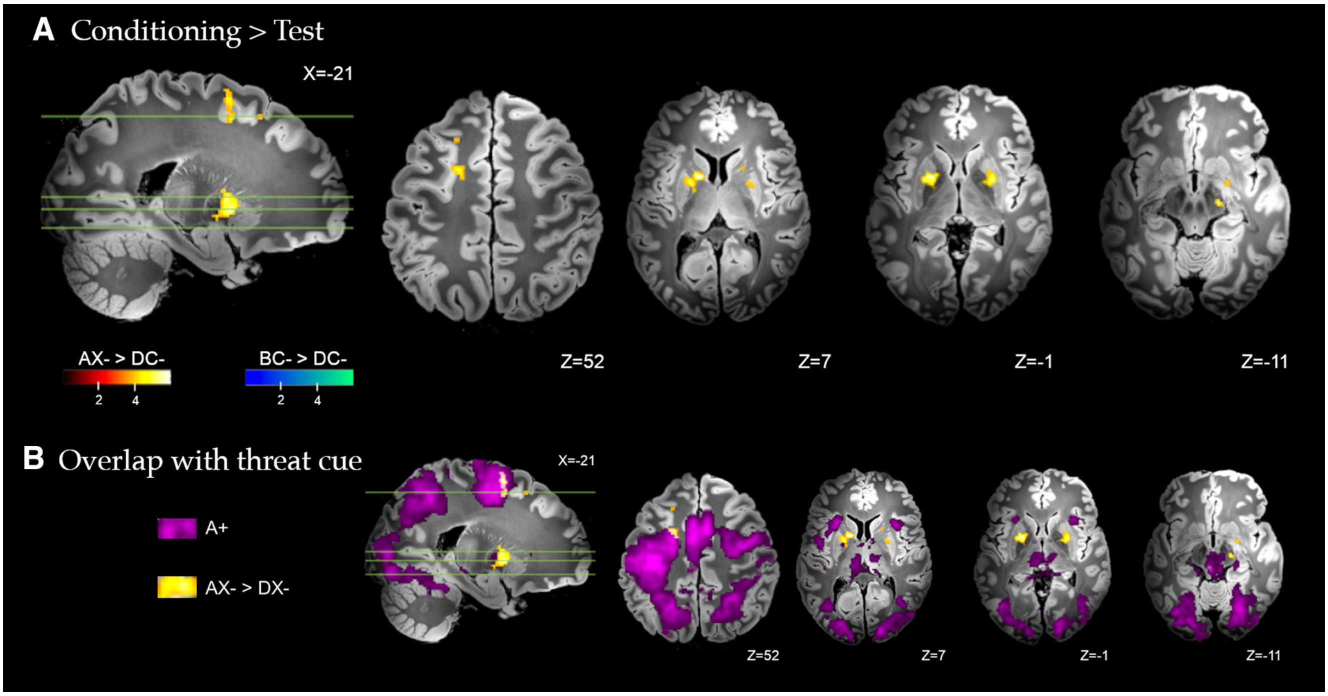

Figure 5. ***A***, Brain regions with significant differential response during conditioning versus transfer test, for the conditioned inhibitor (AX$^-$ > DX$^-$) and standard safety signal (BC$^-$ > DC$^-$). Whole-brain FDR-corrected ($p < 0.05$) results are displayed on a high-resolution anatomic template in MNI space. ***B***, Partial overlap between the differential response to the conditioned inhibitor at training versus test (AX$^-$ > DX$^-$, yellow) and the response to the conditioned threat alone (A$^+$, purple) with activation of the left premotor cortex.

Table 3. Safety conditioning versus test phase[a]

	MNI coordinates				
Regions	*x*	*y*	*z*	K_E	*Z*
Conditioned inhibitor (AX$^-$ > DX$^-$)					
Putamen	−14	6	6	371	6.07
Putamen	26	2	0	269	5.54
Midbrain (substantia nigra)	18	−16	−13	19	5.54
Superior frontal gyrus	−19	3	51	138	4.97
Middle frontal gyrus	26	27	32	19	4.71
Brainstem (periaqueductal gray)	8	−35	−6	10	4.57
Superior frontal gyrus	−26	2	46	14	4.46
Caudate body	18	10	10	20	4.38
Precuneus	19	−54	16	8	4.34
Caudate tail	−19	−2	18	5	4.27
Superior frontal gyrus	−21	21	51	7	4.02
Middle frontal gyrus	30	37	29	7	3.98
Superior frontal gyrus	24	11	56	6	3.89
Standard safety (BC$^-$ > DC$^-$)					
Precentral gyrus	40	−13	56	41	5.93
Fusiform gyrus	22	−45	−13	48	5.48
Lingual gyrus	11	−62	2	9	4.65
Lingual gyrus	22	−59	−10	5	4.49

[a] K_E, cluster size in number of voxels; *Z*, SPM *Z* scores. Coordinates reported in MNI space. Whole-brain contrasts estimated at $P_{FDR} < 0.05$.

theory (Roesch et al., 2012), these overlapping findings suggest that US omission can function as a potent reinforcer when it contradicts expectation (Tobler et al., 2003), rather than being a meaningless or neutral event. The convergence of threat and reward omission on dorsal striatal systems thus points to a more fundamental role in inhibitory learning, such that the putamen, caudate, and pallidum encode learning of "CS → no US" associations whether the omitted US is pleasant or noxious. In comparison, other studies suggest an opposing role for the ventral striatum, which underlies the dynamic reorganization of aversive and rewarding excitatory (CS → US) associations (Klucken et al., 2009; Li et al., 2011; Richter et al., 2020; Grill et al., 2021; Stelly et al., 2021). Similarly, dopamine signals in the nucleus accumbens do not appear to encode prediction errors for threat omission (Kutlu et al., 2021), which may explain the ventral striatum's lack of involvement in safety learning (Josselyn et al., 2005; Mohammadi et al., 2014).

While comparisons between the training and test phases were instructive for discriminating safety learning from expression related neural activity, we did not observe robust behavioral differences between the safety signals during the test phase comparison, unlike our recent behavioral study (Laing et al., 2021). This null finding likely reflects adjustments to the task that were made for the fMRI environment, particularly the reduction of trial repetitions to reduce the overall length of the experiment. It may also be useful to examine test responses in future applications of the conditioned inhibition approach after imposing a delay between the conditioning and test phase (e.g., ≥24 h intervals after learning) (Lonsdorf et al., 2017). Further, our SCR analyses were impacted by the number of data exclusions resulting from artifacts, thereby limiting the scope of inferences drawn from those responses. Finally, the use of a 100% reinforcement rate limited the capacity to compare brain activity for safety and threat cues more directly, which presents a challenge for further implementations of this paradigm in fMRI.

In conclusion, our mapping of safety learning via conditioned inhibition aligns with evidence from nonhuman studies, provides novel evidence for the involvement of known reinforcement learning systems, and proposes an extension to current models of human safety learning. A distributed cortico-striatal system, centered on the dorsal striatum, showed elevated activity during early learning, when threat expectations and safety outcomes require greatest reconciliation, with decreased activity as trials proceeded. This may represent a neural substrate of safety learning, where initial "stimulus–safety" associations are formed, which can be distinguished from wider cortical correlates of safety expression that facilitate the behavioral outcomes of learning. These cortico-striatal systems could provide novel avenues for clinical translation. For instance, although anxiety-related neuropathology is often described in terms of prefrontal threat processing (Alexandra Kredlow et al., 2022), recent studies have shifted focus toward reinforcement learning systems similar to those observed here (Zilcha-Mano et al., 2020; Ney et al., 2021; Seidemann et al., 2021), which could provide new insights for characterizing pathophysiology and neural mechanisms of treatment response.

References

Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13:266–271.

Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9:357–381.

Alexandra Kredlow M, Fenster RJ, Laurent ES, Ressler KJ, Phelps EA (2022) Prefrontal cortex, amygdala, and threat processing: implications for PTSD. Neuropsychopharmacology 47:247–259.

Bach DR, Melinscak F (2020) Psychophysiological modelling and the measurement of fear conditioning. Behav Res Ther 127:103576.

Bach DR, Castegnetti G, Korn CW, Gerster S, Melinscak F, Moser T (2018) Psychophysiological modeling: current state and future directions. Psychophysiology 55:e13214.

Battaglia S, Garofalo S, di Pellegrino G, Starita F (2020) Revaluing the role of vmPFC in the acquisition of Pavlovian threat conditioning in humans. J Neurosci 40:8491–8500.

Battaglia S, Harrison BJ, Fullana MA (2022) Does the human ventromedial prefrontal cortex support fear learning, fear extinction or both? A commentary on subregional contributions. Mol Psychiatry 27:784–786.

Behzadi Y, Restom K, Liau J, Liu TT (2007) A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37:90–101.

Birn RM, Smith MA, Jones TB, Bandettini PA (2008) The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage 40:644–654.

Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25:49–59.

Chang C, Cunningham JP, Glover GH (2009) Influence of heart rate on the BOLD signal: the cardiac response function. Neuroimage 44:857–869.

Choi EY, Yeo BT, Buckner RL (2012) The organization of the human striatum estimated by intrinsic functional connectivity. J Neurophysiol 108:2242–2263.

Christianson JP, Benison AM, Jennings J, Sandsmark EK, Amat J, Kaufman RD, Baratta MV, Paul ED, Campeau S, Watkins LR, Barth DS, Maier SF (2008) The sensory insular cortex mediates the stress-buffering effects of safety signals but not behavioral control. J Neurosci 28:13703–13711.

Christianson JP, Jennings JH, Ragole T, Flyer JG, Benison AM, Barth DS, Watkins LR, Maier SF (2011) Safety signals mitigate the consequences of uncontrollable stress via a circuit involving the sensory insular cortex and bed nucleus of the stria terminalis. Biol Psychiatry 70:458–464.

Cooper JC, Dunne S, Furey T, O'Doherty JP (2012) Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. J Cogn Neurosci 24:106–118.

Corlett PR, Mollick JA, Kober H (2022) Meta-analysis of human prediction error for incentives, perception, cognition, and action. Neuropsychopharmacology 47:1339–1349.

Di Martino A, Scheres A, Margulies DS, Kelly AM, Uddin LQ, Shehzad Z, Biswal B, Walters JR, Castellanos FX, Milham MP (2008) Functional connectivity of human striatum: a resting state FMRI study. Cereb Cortex 18:2735–2747.

Edlow BL, Mareyam A, Horn A, Polimeni JR, Witzel T, Tisdall MD, Augustinack JC, Stockmann JP, Diamond BR, Stevens A, Tirrell LS, Folkerth RD, Wald LL, Fischl B, van der Kouwe A (2019) 7 Tesla MRI of the ex vivo human brain at 100 micron resolution. Sci Data 6:244.

Eisenberger NI, Master SL, Inagaki TK, Taylor SE, Shirinyan D, Lieberman MD, Naliboff BD (2011) Attachment figures activate a safety signal-related neural region and reduce pain experience. Proc Natl Acad Sci USA 108:11721–11726.

Eshel N, Tian J, Bukwich M, Uchida N (2016) Dopamine neurons share common response function for reward prediction error. Nat Neurosci 19:479–486.

Esser R, Korn CW, Ganzer F, Haaker J (2021) L-DOPA modulates activity in the vmPFC, nucleus accumbens, and VTA during threat extinction learning in humans. Elife 10:e65280.

Fendt M, Kreutzmann JC, Jovanovic T (2021) Learning safety to reduce fear: recent insights and potential implications. Behav Brain Res 411:113402.

Foilb AR, Flyer-Adams JG, Maier SF, Christianson JP (2016) Posterior insular cortex is necessary for conditioned inhibition of fear. Neurobiol Learn Mem 134:317–327.

Fullana MA, Harrison BJ, Soriano-Mas C, Vervliet B, Cardoner N, Àvila-Parcet A, Radua J (2016) Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol Psychiatry 21:500–508.

Fullana MA, Albajes-Eizagirre A, Soriano-Mas C, Vervliet B, Cardoner N, Benet O, Radua J, Harrison BJ (2018) Fear extinction in the human brain: a meta-analysis of fMRI studies in healthy participants. Neurosci Biobehav Rev 88:16–25.

Glover GH, Li TQ, Ress D (2000) Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn Reson Med 44:162–167.

Grasser LR, Jovanovic T (2021) Safety learning during development: implications for development of psychopathology. Behav Brain Res 408:113297.

Grill F, Nyberg L, Rieckmann A (2021) Neural correlates of reward processing: functional dissociation of two components within the ventral striatum. Brain Behav 11:e01987.

Harris JA, Kwok DW, Andrew BJ (2014) Conditioned inhibition and reinforcement rate. J Exp Psychol Anim Learn Cogn 40:335–354.

Harrison BJ, Fullana MA, Via E, Soriano-Mas C, Vervliet B, Martínez-Zalacaín I, Pujol J, Davey CG, Kircher T, Straube B, Cardoner N (2017) Human ventromedial prefrontal cortex and the positive affective processing of safety signals. Neuroimage 152:12–18.

Hennings AC, McClay M, Lewis-Peacock JA, Dunsmoor JE (2020) Contextual reinstatement promotes extinction generalization in healthy adults but not PTSD. Neuropsychologia 147:107573.

Hermann A, Stark R, Müller EA, Kruse O, Wolf OT, Merz CJ (2020) Multiple extinction contexts modulate the neural correlates of context-dependent extinction learning and retrieval. Neurobiol Learn Mem 168:107150.

Josselyn SA, Falls WA, Gewirtz JC, Pistell P, Davis M (2005) The nucleus accumbens is not critically involved in mediating the effects of a safety signal on behavior. Neuropsychopharmacology 30:17–26.

Kasper L, Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, Hauser TU, Sebold M, Manjaly ZM, Pruessmann KP, Stephan KE (2017) The PhysIO Toolbox for modeling physiological noise in fMRI data. J Neurosci Methods 276:56–72.

Klucken T, Tabbert K, Schweckendiek J, Merz CJ, Kagerer S, Vaitl D, Stark R (2009) Contingency learning in human fear conditioning involves the ventral striatum. Hum Brain Mapp 30:3636–3644.

Kreutzmann JC, Fendt M (2020) Chronic inhibition of GABA synthesis in the infralimbic cortex facilitates conditioned safety memory and reduces contextual fear. Transl Psychiatry 10:120.

Kreutzmann JC, Jovanovic T, Fendt M (2020) Infralimbic cortex activity is required for the expression but not the acquisition of conditioned safety. Psychopharmacology (Berl) 237:2161–2172.

Kutlu MG, Zachry JE, Melugin PR, Cajigas SA, Chevee MF, Kelley SJ, Kutlu B, Tian L, Siciliano CA, Calipari ES (2021) Dopamine release in the nucleus accumbens core signals perceived saliency. Curr Biol 31:4748–4761.e8.

Laing PA, Harrison BJ (2021) Safety learning and the Pavlovian-conditioned inhibition of fear in humans: current state and future directions. Neurosci Biobehav Rev 127:659–674.

Laing PA, Vervliet B, Fullana MA, Savage HS, Davey CG, Felmingham KL, Harrison BJ (2021) Characterizing human safety learning via Pavlovian-conditioned inhibition. Behav Res Ther 137:103800.

Lanciego JL, Luquin N, Obeso JA (2012) Functional neuroanatomy of the basal ganglia. Cold Spring Harb Perspect Med 2:a009621.

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

Lonsdorf TB, et al. (2017) Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neurosci Biobehav Rev 77:247–285.

Lysle DT, Fowler H (1985) Inhibition as a 'slave' process: deactivation of conditioned inhibition through extinction of conditioned excitation. J Exp Psychol Anim Behav Process 11:71–94.

Marques JP, Kober T, Krueger G, van der Zwaag W, Van de Moortele P-F, Gruetter R (2010) MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. NeuroImage 49:1271–1281.

Mechias ML, Etkin A, Kalisch R (2010) A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. Neuroimage 49:1760–1768.

Meyer HC, Odriozola P, Cohodes EM, Mandell JD, Li A, Yang R, Hall BS, Haberman JT, Zacharek SJ, Liston C, Lee FS, Gee DG (2019) Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. Proc Natl Acad Sci USA 116:26970–26979.

Mohammadi M, Bergado-Acosta JR, Fendt M (2014) Relief learning is distinguished from safety learning by the requirement of the nucleus accumbens. Behav Brain Res 272:40–45.

Mollick JA, Chang LJ, Krishnan A, Hazy TE, Krueger KA, Frank GK, Wager TD, O'Reilly RC (2021) The neural correlates of cued reward omission. Front Hum Neurosci 15:615313.

Neumann DL, Lipp OV, Siddle DA (1997) Conditioned inhibition of autonomic Pavlovian conditioning in humans. Biol Psychol 46:223–233.

Ney LJ, Akhurst J, Bruno R, Laing PA, Matthews A, Felmingham KL (2021) Dopamine, endocannabinoids and their interaction in fear extinction and negative affect in PTSD. Prog Neuropsychopharmacol Biol Psychiatry 105:110118.

Oyama K, Tateyama Y, Hernádi I, Tobler PN, Iijima T, Tsutsui KI (2015) Discrete coding of stimulus value, reward expectation, and reward prediction error in the dorsal striatum. J Neurophysiol 114:2600–2615.

Papalini S, Beckers T, Vervliet B (2020) Dopamine: from prediction error to psychotherapy. Transl Psychiatry 10:164.

Pauli WM, Larsen T, Collette S, Tyszka JM, Seymour B, Doherty JP (2015) Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. J Neurosci 35:14220–14233.

Raczka KA, Mechias ML, 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.

Rescorla RA (1969) Pavlovian-conditioned inhibition. Psychol Bull 72:77–94.

Reynaud O, Jorge J, Gruetter R, Marques JP, van der Zwaag W (2017) Influence of physiological noise on accelerated 2D and 3D resting state functional MRI data at 7 T. Magn Reson Med 78:888–896.

Richter A, Reinhard F, Kraemer B, Gruber O (2020) A high-resolution fMRI approach to characterize functionally distinct neural pathways within dopaminergic midbrain and nucleus accumbens during reward and salience processing. Eur Neuropsychopharmacol 36:137–150.

Roesch MR, Esber GR, Li J, Daw ND, Schoenbaum G (2012) Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. Eur J Neurosci 35:1190–1200.

Rogan MT, Leon KS, Perez DL, Kandel ER (2005) Distinct neural signatures for safety and danger in the amygdala and striatum of the mouse. Neuron 46:309–320.

Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci 16:147–156.

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:e38818.

Sarlitto MC, Foilb AR, Christianson JP (2018) Inactivation of the ventrolateral orbitofrontal cortex impairs flexible use of safety signals. Neuroscience 379:350–358.

Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. Annu Rev Neurosci 23:473–500.

Schultz W, Tremblay L, Hollerman JR (2003) Changes in behavior-related neuronal activity in the striatum during learning. Trends Neurosci 26:321–328.

Sclocco R, Beissner F, Bianciardi M, Polimeni JR, Napadow V (2018) Challenges and opportunities for brainstem neuroimaging with ultrahigh field MRI. Neuroimage 168:412–426.

Seidemann R, Duek O, Jia R, Levy I, Harpaz-Rotem I (2021) The reward system and post-traumatic stress disorder: does trauma affect the way we interact with positive stimuli? Chronic Stress (Thousand Oaks) 5:2470547021996006.

Setsompop K, Gagoski BA, Polimeni JR, Witzel T, Wedeen VJ, Wald LL (2012) Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. Magn Reson Med 67:1210–1224.

Sharvit G, Corradi-Dell'Acqua C, Vuilleumier P (2018) Modality-specific effects of aversive expectancy in the anterior insula and medial prefrontal cortex. Pain 159:1529–1542.

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC (1998) The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 59:22–33.

Sosa R, Ramírez MN (2019) Conditioned inhibition: historical critiques and controversies in the light of recent advances. J Exp Psychol Anim Learn Cogn 45:17–42.

Stelly CE, Girven KS, Lefner MJ, Fonzi KM, Wanat MJ (2021) Dopamine release and its control over early Pavlovian learning differs between the NAc core and medial NAc shell. Neuropsychopharmacology 46:1780–1787.

Tashjian SM, Zbozinek TD, Mobbs D (2021) A decision architecture for safety computations. Trends Cogn Sci 25:342–354.

Teckentrup V, van der Meer JN, Borchardt V, Fan Y, Neuser MP, Tempelmann C, Herrmann L, Walter M, Kroemer NB (2019) The anterior insula channels prefrontal expectancy signals during affective processing. Neuroimage 200:414–424.

Tobler PN, Dickinson A, Schultz W (2003) Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. J Neurosci 23:10402–10410.

Valentin VV, O'Doherty JP (2009) Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. J Neurophysiol 102:3384–3391.

van Rooij SJ, Jovanovic T (2019) Impaired inhibition as an intermediate phenotype for PTSD risk and treatment response. Prog Neuropsychopharmacol Biol Psychiatry 89:435–445.

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43–48.

Wagner AR, Rescorla RA (1972) Inhibition in Pavlovian conditioning: application of a theory. Inhibition and Learning, pp 301–336. Available at https://psycnet.apa.org/record/1973-06351-000

Warren VT, Anderson KM, Kwon C, Bosshardt L, Jovanovic T, Bradley B, Norrholm SD (2014) Human fear extinction and return of fear using reconsolidation update mechanisms: the contribution of on-line expectancy ratings. Neurobiol Learn Mem 113:165–173.

Wilke M (2012) An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. Neuroimage 59:2062–2072.

Yan R, Wang T, Zhou Q (2019) Elevated dopamine signaling from ventral tegmental area to prefrontal cortical parvalbumin neurons drives conditioned inhibition. Proc Natl Acad Sci USA 116:13077–13086.

Yoshizawa T, Ito M, Doya K (2018) Reward-predictive neural activities in striatal striosome compartments. eNeuro 5:ENEURO.0367-17.2018.

Zhang Z, Mendelsohn A, Manson KF, Schiller D, Levy I (2015) Dissociating value representation and inhibition of inappropriate affective response during reversal learning in the ventromedial prefrontal cortex. eNeuro 2: ENEURO.0072-15.2015.

Zilcha-Mano S, Zhu X, Suarez-Jimenez B, Pickover A, Tal S, Such S, Marohasy C, Chrisanthopoulos M, Salzman C, Lazarov A, Neria Y, Rutherford BR (2020) Diagnostic and predictive neuroimaging biomarkers for posttraumatic stress disorder. Biol Psychiatry Cogn Neurosci Neuroimaging 5:688–696.