Cortico-striatal activity characterizes human safety learning via Pavlovian conditioned inhibition

Abbreviated title: Safety learning and cortico-striatal activity

Patrick A.F. Laing ^{a*}, Trevor Steward ^{a,b}, Christopher G. Davey ^a, Kim L. Felmingham ^b, Miguel Angel Fullana ^{c, d}, Bram Vervliet ^{e, f}, Matthew D. Greaves ^a, Bradford Moffat ^g, Rebecca K. Glarin ^g, Ben J. Harrison ^{a*}

- ^a. Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne
- ^b. Melbourne School of Psychological Sciences, The University of Melbourne
- ^c. Adult Psychiatry and Psychology Department, Institute of Neurosciences, Hospital Clinic, Barcelona, Spain
- ^d. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Barcelona, Spain
- ^e. Laboratory of Biological Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
- ^f. Leuven Brain Institute, KU Leuven, Belgium
- ^g. The Melbourne Brain Centre Imaging Unit, Department of Radiology, The University of Melbourne

Conflicting or financial interests: The authors declare no conflicts of interest

Acknowledgments: This work was funded by a National Health and Medical Research Council of Australia (NHMRC) Project Grant (1161897) to BJH. We acknowledge the scientific and technical assistance of the Australian National Imaging Facility – a National Collaborative Research Infrastructure Strategy (NCRIS) capability at the Melbourne Brain Centre Imaging Unit (MBCIU), The University of Melbourne. The multiband fMRI sequence was generously supported by a research collaboration agreement with CMRR, The University of Minnesota and the MP2RAGE works in progress sequence was provided by Siemens Healthineers (Germany). The authors gratefully acknowledge Ms Lisa Incerti for her contributions to data collection.

***Corresponding authors:** plaingresearch@gmail.com; habj@unimelb.edu.au

Present address: Melbourne Neuropsychiatry Centre, Level 3, Alan Gilbert Building, 161 Barry Street, Carlton, 3053, Victoria, Australia

Abstract

Safety learning generates associative links between neutral stimuli and the absence of threat, promoting the inhibition of fear and security-seeking behaviours. 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 (UHF) fMRI to examine the neural basis of inhibitory 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. The conditioned inhibitor was characterized by more prominent activation of the dorsal striatum, anterior insular and dorsolateral prefrontal cortex compared to the standard safety signal, whereas the latter evoked greater activation of the ventromedial prefrontal cortex, 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.

Keywords: *conditioned inhibition, safety learning, prediction error, dorsal striatum, vmPFC, 7-Tesla fMRI*

Significance statement:

Identifying safety is critical for maintaining adaptive levels of anxiety, but the neural mechanisms of human safety learning remain unclear. Using ultra-high field 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/prefrontal cortex 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.

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 (Grasser and Jovanovic, 2021; van Rooij and Jovanovic, 2019), 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 threat responses compared to fear signals, the same occurs for neutral or ambiguous stimuli, which predict neither the presence or 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 (Grupe and Nitschke, 2013; Tashjian et al., 2021).

Prevailing neuroimaging evidence for human safety learning stems from differential fear conditioning studies, where a standard safety signal (CS-) is compared to a conditioned threat stimulus (CS+). These studies consistently identify increased activity of the ventromedial prefrontal cortex (vmPFC), posterior cingulate cortex, and hippocampus in discriminating CS- from CS+ (Fullana et al., 2015). Though robust, it is unclear to what extent these findings directly 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 (Gewirtz et al., 1997; Kreutzmann et al., 2020; Sarlitto et al., 2018). Further, while medial prefrontal cortical pathways from the ventral tegmentum and hippocampus facilitate the post-conditioning 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 (Christianson et al., 2008; Christianson et al., 2011; Foilb et al., 2016; Rogan et al., 2005). These findings suggest that

important differences exist 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., 2021; Tashjian et al., 2021).

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 (Yau and McNally, 2018). Established learning theories predict that this mechanism generates robust links between the inhibitor and threat-omission (Schultz and Dickinson, 2000; Wagner and Rescorla, 1972), providing an operational definition of safety learning (Laing and Harrison, 2021). The paradigm has been shown to produce conditioned safety in human behavioral studies (Laing et al., 2021; Neumann et al., 1997), but is yet to be investigated in neuroimaging.

The aim of the current study was to examine the neural basis of human safety learning via Pavlovian conditioned inhibition combined with ultra-high field fMRI. Our first aim was to directly compare 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 aimed to identify regions underlying the initial learning of ‘stimulus-safety’ associations by contrasting early and late conditioning trials, and comparing responses

during conditioning to those during a subsequent test phase. We hypothesized that the conditioned inhibitor would more distinctly engage subcortical circuitry, particularly striatal and midbrain regions, which have well-established roles in prediction error-based reinforcement learning (Garrison et al., 2013; Papalini et al., 2020), whereas the standard safety signal would involve distributed cortical regions, including the vmPFC, linked to the cognitive evaluation of learned safety.

Material and Methods

Participants. Forty-nine participants were recruited to the study. All participants met the following eligibility criteria: (i) they were aged between 18-35 years, (ii) had no current or past diagnosis of mental illness as per screening via the Mini-International Neuropsychiatric Interview (MINI, Sheehan et al., 1998), (iii) were fluent in English, (iv) were not taking any psychoactive medications, and (v) 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 due to technical failure; one who discontinued), and a further four were excluded due to excessive head motion (see image pre-processing). 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 (Figure 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 500ms duration. Inter-trial-intervals (ITI) were jittered between 8-12s ($M = 10$ s) and featured a white fixation cross in the center of a black

screen. Each CS had a duration of 5s total, firstly presented for 1s alone, then joined by a threat-expectancy rating scale for a further 4s, 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. 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' upon 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 (Harris et al., 2014; Laing and Harrison, 2021; Lysle and Fowler, 1985). While partial CS+ reinforcement is typically used to avoid the US confounding of CS+ responses (Fullana et al., 2015), this was unnecessary in the current study, which was designed to target safety signals exclusively.

The experiment featured conditioning and test phases. Six CS configurations were presented in conditioning: A+, AX-, BC-, D+, AD+, and X- (Figure 1). 'AX-' was the conditioned inhibitor, which was compared to the control safety signal 'BC-': hereafter referred to as the *standard safety signal*. Each CS was presented for 6 trials each, in pseudorandomized order, resulting in 36 trials. The test phase consisted of three CS configurations (AD+, DC-, DX-) also for 6 trials each, resulting in 18 trials total. To ensure that test responses were not the product of conscious instruction (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, Pittsburgh, PA) and was presented on a 32" 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 Corp.), which also provided passive noise cancellation (~30dB). Participants' responses were registered with a 2-button LS-PAIR Lumina response pad (Cedrus Corporation USA), which they were familiarized with prior to scanning.

--- Insert Figure 1 ---

Skin conductance responses (SCR). Skin conductance was recorded using MRI-compatible finger electrodes (Ag/AgCl) 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 prior to the attachment of electrodes. The signal was amplified and sampled at 1000 Hz using PowerLab v8.0 (ADInstruments, Dunedin, NZ) and recording was triggered concurrently with the beginning of the experiment and the functional imaging sequence. The Psychophysiological Modelling Toolbox (PsPM, Bach et al., 2018; Bach and Melinscak, 2020) in MATLAB (version 9.4, The MathWorks Inc.) was used for pre-processing and modelling of SCRs. A custom program was used to identify artefacts in the skin conductance timeseries, which were removed in PsPM prior to further analysis. After artefact removal, SC data were filtered with a 10ms median filter followed by a first order bidirectional band-pass Butterworth filter (cut-off frequencies 0.0159 - 5Hz) and down-sampled to 10Hz. Dynamic Causal Modelling was implemented via the PsPM toolbox to

provide trial-by-trial estimates of sympathetic nervous system activity. SCR artefacts were detected via a semi-automated process. A custom program identified time series containing: a) signal increases greater than 20% per second, b) signal decreases greater than 10% per second, or c) absolute changes greater than 0.075 μ S per millisecond. 17 subjects' SC data were flagged for review based on these criteria and excluded after manual review. A further 7 subjects' SCR data was comprised by technical issues. This resulted in a final SCR sample of $N=24$.

Image acquisition. Imaging was performed on a 7T research scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 32-channel head-coil (Nova Medical Inc., Wilmington MA, USA). The functional sequence consisted of a multi-band (6 times) and grappa (2 times) accelerated GE-EPI sequence in the steady state (TR, 800ms; TE, 22.2 ms; pulse/flip angle, 45°; field-of-view, 20.8 cm; slice thickness (no gap), 1.6 mm; 130 x 130-pixel matrix; 84 interleaved axial slices aligned to AC-PC line; Setsompop et al., 2012). The total sequence time was 16 minutes and 10 seconds, corresponding to 1202 whole-brain EPI volumes. A T1-weighted high-resolution anatomical image (MP2RAGE; Marques et al. 2010) was acquired for each participant to assist with functional time series co-registration (TR = 5000 ms; TE, 3.0 ms; inversion times, 700/2700ms; pulse/flip angles, 4/5°; field-of-view, 24 cm; slice thickness (no gap), 0.73 mm; 330 x 330-pixel matrix; 84 sagittal slices aligned parallel to the midline). The total sequence time was 7 minutes and 12 seconds. 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 (see further).

Image pre-processing. Imaging data was pre-processed using Statistical Parametric Mapping (SPM) 12 (v7771, Wellcome Trust Centre for Neuroimaging, London, UK) within a MATLAB 2019b environment (The MathWorks Inc., Natick, MA). Motion correction was performed by realigning each subject's time series to the mean image, and all images were resampled using 4th Degree B-Spline interpolation. Individualized motion parameters were estimated via Motion Fingerprint (Wilke, 2012) to account for head motion, and participant data was 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 anatomical image was co-registered 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 3.2mm^3 full-width-at-half-maximum (FWHM) gaussian kernel to increase anatomical precision. Physiological noise was modelled 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 blood-oxygen level dependent (BOLD) signal sensitivity and temporal signal-to-noise ratio (tSNR) at 7T (Reynaud et al., 2017). The Retrospective Image-based Correction function (RETROICOR, Glover et al., 2000) was applied to model periodic effects of heartbeat and respiration on BOLD signals. The Respiratory Response Function (RRF, Birn et al., 2008), convolved with respiration volume per time (RVT), was used to model low frequency signal fluctuations arising from changes in depth and rate of breath. Heart rate variability (HRV) was convolved with a Cardiac Response Function (CRF, Chang et al., 2009) to account for BOLD variances due to heartrate-dependent changes in blood oxygenation. Individualized DARTEL tissue maps segmented from each participant's respective anatomical 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 pre-processed time-series was included in a first-level SPM general linear model (GLM) analysis, which specified the onsets of each CS event-type (grouped as separate conditions for 'early'/first 3 trials and 'late'/last 3 trials) in each task phase to be convolved with canonical hemodynamic response function. 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 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-). Secondly, to identify learning-specific responses, we compared early versus late trials for each safety signal, respectively (e.g., AX-*early* vs. AX-*late*). Thirdly, to distinguish brain responses to each safety signal during learning versus expression, we compared responses to each stimulus type across conditioning and test phases (e.g., AX- vs. DX-), and differences between safety signals when combined with a CS+ at test (e.g., DX- vs. DC-). For each of the main conditioned inhibitor contrasts, we provide extended results that illustrate its overlap with brain responses to the simple conditioned threat (A+). Statistical significance was estimated with a whole-brain false-discovery rate corrected threshold ($P_{FDR} < 0.05$) and a 5-voxel cluster-extent threshold ($K_E \geq 5$). Analyses of threat-expectancy ratings and SCR were performed by separating responses into early and late phases (3 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 conditioned inhibitor and control cue showed a main effect of stimulus ($F_{1,45} = 35.22, p < 0.001, \eta^2_p = 0.44$) and phase (early, late, $F_{1,45} = 234.45, p < 0.001, \eta^2_p = 0.84$), but no interaction. Post-hoc test showed higher threat-expectancy for the inhibitor (AX- > BC-; $M = 10, t = 5.93, d = 0.87, p < 0.001$), driven by significant differences in early ($M = 14.85, t = 6.47, p < 0.001$), but not late ($M = 5.15, t = 2.24, p = 0.17$) conditioning trials (Figure 2A). SCRs (Figure 2B) showed no effect of stimulus (AX- > BC-; $p = 0.93$), but a significant effect of phase (early, late, $F_{1,23} = 4.58, p < 0.05, \eta^2_p = 0.17$), indicating elevated SCRs in early conditioning which were decreased by later trials (early > late: $M = 0.13, t = 2.14, d = 0.44, p = 0.04$). Transfer test responses showed mixed results. Threat-expectancy showed main effects of stimulus ($F_{1,45} = 894.58, p < 0.001, \eta^2_p = 0.95$), phase ($F_{1,45} = 28.25, p < 0.001, \eta^2_p = 0.39$), and their interaction ($F_{1,45} = 16.48, p < 0.001, \eta^2_p = 0.27$). Expectancy was robustly decreased for the conditioned inhibitor (AD+ > DX-; $M = 88.08, t = 36.74, d = 5.42, p < 0.001$) and standard safety signal (AD+ > DC-; $M = 87.57, t = 36.52, d = 5.38, p < 0.001$) 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 behavioural threat response during learning, but did not differ in responses evoked at test.

Subjective ratings: positive affect and anxious arousal. Compared to a CS+, the conditioned inhibitor and standard safety signal elicited greater positive affect ($X > D$: $M = 62.5$, $t = 7.58$, $d = 1.12$, $p < 0.001$; $C > D$: $M = 60.33$, $t = 7.32$, $d = 1.08$, $p < 0.001$; Figure 2E) and lower anxious arousal ($X > D$: $M = 51.09$, $t = 10.38$, $d = 1.53$, $p < 0.001$; $C > D$: $M = 50.54$, $t = 10.26$, $d = 1.51$, $p < 0.001$; Figure 2F). Further, both safety signals also evoked greater positive affect ($X >$ fixation cross: $M = 26.09$, $t = 3.17$, $d = 0.47$, $p = 0.006$; $C >$ fixation cross: $M = 23.92$, $t = 2.90$, $d = 0.43$, $p = 0.009$; Figure 2E) and lower arousal (fixation cross $>$ X: $M = 14.67$, $t = 2.98$, $d = 0.44$, $p = 0.01$; fixation cross $>$ C: $M = 14.13$, $t = 2.87$, $d = 0.42$, $p = 0.01$; Figure 2F) 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).

--- Insert Figure 2 ---

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 ($AX- > BC-$). 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 prefrontal cortex (Figure 3). Figure 3-1 (extended data) highlights the overlap between the differential response to the conditioned inhibitor versus standard safety signal and the response to conditioned threat alone (A+). Partial overlap was noted with regards to activation of the anterior insular cortex. The standard safety signal compared to conditioned inhibitor (BC- > AX-) evoked significantly greater activation of the ventromedial prefrontal cortex, 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’ (500um, MNI space) *ex vivo* template (Edlow et al., 2019).

--- Insert Figure 3 ---

--- Insert Table 1 ---

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 vs early trials (AX- *late* > AX- *early*). Figure 4-1 (extended data) 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+).

--- Insert Figure 4 ---

--- Insert Table 2 ---

For the standard safety signal ($BC_{-early} > BC_{-late}$), early trials evoked significantly greater activation of posterior dorsolateral prefrontal cortex, superior- and intraparietal cortex, precuneus (dorsal and ventral subareas), and medial and lateral visual association cortex, spanning the fusiform, lingual gyrus and occipital pole. There was no significantly greater activation identified for the standard safety signal during late vs 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 grey, bilateral dorsal premotor cortex, and precuneus. Complete results are provided in Table 3. Figure 5-1 (extended data) 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 to conditioning ($DX_{-} > AX_{-}$). For the standard safety signal, significantly greater activation of the anterior fusiform and lingual gyrus was identified during conditioning compared to test ($BC_{-} > DC_{-}$). There was no significantly greater activation identified for the standard safety signal at test compared to 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 $P_{\text{FDR}} < 0.05$ threshold ($K_E \geq 5$ voxels).

--- Insert Figure 5 ---

--- Insert Table 3 ---

Discussion

This study combined ultra-high field fMRI with Pavlovian conditioned inhibition to characterize the neural basis of safety learning in humans. We compared two kinds of 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 to the standard safety signal, the conditioned inhibitor selectively engaged the dorsal anterior striatum, anterior insular and posterior dorsolateral prefrontal cortex, with further analyses implicating extended activation of midbrain, thalamic and premotor cortex subregions during early safety learning and when comparing conditioning to 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 non-overlapping. 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 (Harris et al., 2014; Lysle and Fowler, 1985), 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 non-overlapping 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 receives inputs from premotor cortical regions, which relays information to the ventrolateral thalamus by way of the pallidum and SNc (Alexander and Crutcher, 1990; Alexander et al., 1986; Haber and Knutson, 2010). Anatomical characteristics of this circuit are also consistent with intrinsic premotor cortex connectivity with the dorsal striatum and SNc observed in humans via fMRI (Choi et al., 2012; Di Martino et al., 2008; Morris et al., 2016).

The engagement of these cortico-striatal regions by the inhibitor, but not the standard safety signal, is consistent with the primary psychological difference between these stimuli: the manipulation of expectancy violation (Laing et al., 2021). In our task, early trials of the conditioned inhibitor (AX-) are expected 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 the 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 theory (Pauli et al., 2015; Schultz and Dickinson, 2000; Schultz et al., 2003). For instance, dorsal striatal nuclei show elevated prediction error responses in early conditioning, which decreases as contingencies are learned (Cooper et al., 2012; Oyama et al., 2015; Valentin and O'Doherty, 2009), while also updating conditioned responses to the CS (Yoshizawa et al., 2018). The dorsal striatum receives direct projections from midbrain dopaminergic neurons of the SNc (Lanciego et al., 2012), which provide a prediction error signal by firing in response to unexpected stimulus-delivery (or omission), and decreasing in magnitude as expectations and outcomes are aligned (Eshel et al., 2016; Salinas-Hernández et al., 2018; Waelti et al., 2001). As a result, safety learning via conditioned inhibition appears to invoke a more direct instantiation of these active error-corrective learning systems compared to the standard safety signal, which is passively unreinforced, and 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., 2015). 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 (Foilb et al., 2016; Rogan et al., 2005), and expression-specific contributions of hippocampal and prefrontal cortical regions (Kreutzmann and Fendt, 2020; Kreutzmann et al., 2020; Meyer et al., 2019; Yan et al., 2019). 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 (Harrison et al., 2017; Hennings et al., 2020; Hermann et al., 2020; Merz et al., 2018; Roy et al., 2012). The vmPFC has a notably multifaceted functional role, modulating both fear and safety expression, rather than unidirectional threat-inhibition (Battaglia et al., 2021; Tashjian et al., 2021). 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 (i) a cortico-striatal system that encodes initial safety learning associations, and (ii) 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 both prediction error-based learning and safety-retention across time (Esser et al., 2021; Gerlicher et al., 2018; Raczka et al., 2011; Thiele et al., 2021; Yan et al., 2019).

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 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 (Grill et al., 2021; Klucken et al., 2009; Li et al., 2011; Richter et al., 2020; Schiller et al., 2008; 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., ≥ 24hr intervals post-learning, Lonsdorf et al., 2017).

Conclusion

Overall, our mapping of safety learning via conditioned inhibition aligns with evidence from non-human 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, though anxiety-related neuropathology is often described in terms of prefrontal threat-processing (Alexandra Kredlow et al., 2021), recent studies have shifted focus towards reinforcement learning systems similar to those observed here (Ney et al., 2021; Seidemann et al., 2021; Zilcha-Mano et al., 2020), which could provide new insights for characterizing pathophysiology and neural mechanisms of treatment response.

References

- Alexander, G.E., Crutcher, M.D., 1990. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences* 13, 266-271.
- Alexander, G.E., DeLong, M.R., Strick, P.L., 1986. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience* 9, 357-381.
- Alexandra Kredlow, M., Fenster, R.J., Laurent, E.S., Ressler, K.J., Phelps, E.A., 2021. Prefrontal cortex, amygdala, and threat processing: implications for PTSD. *Neuropsychopharmacology*.
- Bach, D.R., Castegnetti, G., Korn, C.W., Gerster, S., Melinscak, F., Moser, T., 2018. Psychophysiological modeling: Current state and future directions. *Psychophysiology* 0, e13214.
- Bach, D.R., Melinscak, F., 2020. Psychophysiological modelling and the measurement of fear conditioning. *Behaviour Research and Therapy* 127, 103576.
- Battaglia, S., Garofalo, S., di Pellegrino, G., Starita, F., 2020. Revaluing the Role of vmPFC in the Acquisition of Pavlovian Threat Conditioning in Humans. *The Journal of Neuroscience* 40, 8491.
- Battaglia, S., Harrison, B.J., Fullana, M.A., 2021. Does the human ventromedial prefrontal cortex support fear learning, fear extinction or both? A commentary on subregional contributions. *Molecular Psychiatry*.
- Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* 37, 90-101.
- Birn, R.M., Smith, M.A., Jones, T.B., Bandettini, P.A., 2008. The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage* 40, 644-654.
- Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 49-59.
- Chang, C., Cunningham, J.P., Glover, G.H., 2009. Influence of heart rate on the BOLD signal: the cardiac response function. *Neuroimage* 44, 857-869.

- Choi, E.Y., Yeo, B.T.T., Buckner, R.L., 2012. The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology* 108, 2242-2263.
- Christianson, J.P., Benison, A.M., Jennings, J., Sandsmark, E.K., Amat, J., Kaufman, R.D., Baratta, M.V., Paul, E.D., Campeau, S., Watkins, L.R., Barth, D.S., Maier, S.F., 2008. The Sensory Insular Cortex Mediates the Stress-Buffering Effects of Safety Signals But Not Behavioral Control. *The Journal of Neuroscience* 28, 13703.
- Christianson, J.P., Jennings, J.H., Ragole, T., Flyer, J.G.N., Benison, A.M., Barth, D.S., Watkins, L.R., Maier, S.F., 2011. Safety Signals Mitigate the Consequences of Uncontrollable Stress Via a Circuit Involving the Sensory Insular Cortex and Bed Nucleus of the Stria Terminalis. *Biological Psychiatry* 70, 458-464.
- Cooper, J.C., Dunne, S., Furey, T., O'Doherty, J.P., 2012. Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of cognitive neuroscience* 24, 106-118.
- Di Martino, A., Scheres, A., Margulies, D.S., Kelly, A.M.C., Uddin, L.Q., Shehzad, Z., Biswal, B., Walters, J.R., Castellanos, F.X., Milham, M.P., 2008. Functional connectivity of human striatum: a resting state FMRI study. *Cerebral cortex* 18, 2735-2747.
- Edlow, B.L., Mareyam, A., Horn, A., Polimeni, J.R., Witzel, T., Tisdall, M.D., Augustinack, J.C., Stockmann, J.P., Diamond, B.R., Stevens, A., Tirrell, L.S., Folkerth, R.D., Wald, L.L., Fischl, B., van der Kouwe, A., 2019. 7 Tesla MRI of the ex vivo human brain at 100 micron resolution. *Scientific Data* 6, 244.
- Eisenberger, N.I., Master, S.L., Inagaki, T.K., Taylor, S.E., Shirinyan, D., Lieberman, M.D., Naliboff, B.D., 2011. Attachment figures activate a safety signal-related neural region and reduce pain experience. *Proceedings of the National Academy of Sciences* 108, 11721.
- Eshel, N., Tian, J., Bukwich, M., Uchida, N., 2016. Dopamine neurons share common response function for reward prediction error. *Nature neuroscience* 19, 479.
- Esser, R., Korn, C.W., 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, J.C., Jovanovic, T., 2021. Learning safety to reduce fear: Recent insights and potential implications. *Behavioural Brain Research* 411, 113402.

- Foillb, A.R., Flyer-Adams, J.G., Maier, S.F., Christianson, J.P., 2016. Posterior insular cortex is necessary for conditioned inhibition of fear. *Neurobiology of Learning and Memory* 134, 317-327.
- Fullana, M.A., Albajes-Eizaguirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., Harrison, B.J., 2018. Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neuroscience & Biobehavioral Reviews* 88, 16-25.
- Fullana, M.A., Harrison, B.J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., Radua, J., 2015. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry* 21, 500.
- Garrison, J., Erdeniz, B., Done, J., 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews* 37, 1297-1310.
- Gerlicher, A.M.V., Tüscher, O., Kalisch, R., 2018. Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nature Communications* 9, 4294.
- Gewirtz, J.C., Falls, W.A., Davis, M., 1997. Normal conditioning inhibition and extinction of freezing and fear-potentiated startle following electrolytic lesions of medial prefrontal cortex in rats. *Behavioral neuroscience* 111, 712.
- Glover, G.H., Li, T.-Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine* 44, 162-167.
- Grasser, L.R., Jovanovic, T., 2021. Safety learning during development: Implications for development of psychopathology. *Behavioural Brain Research* 408, 113297.
- Grill, F., Nyberg, L., Rieckmann, A., 2021. Neural correlates of reward processing: Functional dissociation of two components within the ventral striatum. *Brain and Behavior* 11, e01987.
- Grupe, D.W., Nitschke, J.B., 2013. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat Rev Neurosci* 14.
- Haber, S.N., Knutson, B., 2010. The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology* 35, 4-26.

- Harris, J.A., Kwok, D.W.S., Andrew, B.J., 2014. Conditioned inhibition and reinforcement rate. *Journal of Experimental Psychology: Animal Learning and Cognition* 40, 335.
- Harrison, B.J., Fullana, M.A., Via, E., Soriano-Mas, C., Vervliet, B., Martínez-Zalacaín, I., Pujol, J., Davey, C.G., Kircher, T., Straube, B., Cardoner, N., 2017. Human ventromedial prefrontal cortex and the positive affective processing of safety signals. *NeuroImage* 152, 12-18.
- Hennings, A.C., McClay, M., Lewis-Peacock, J.A., Dunsmoor, J.E., 2020. Contextual reinstatement promotes extinction generalization in healthy adults but not PTSD. *Neuropsychologia* 147, 107573.
- Hermann, A., Stark, R., Müller, E.A., Kruse, O., Wolf, O.T., Merz, C.J., 2020. Multiple extinction contexts modulate the neural correlates of context-dependent extinction learning and retrieval. *Neurobiology of Learning and Memory* 168, 107150.
- Josselyn, S.A., Falls, W.A., Gewirtz, J.C., 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, A.O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T.U., Sebold, M., Manjaly, Z.-M., Pruessmann, K.P., Stephan, K.E., 2017. The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods* 276, 56-72.
- Klucken, T., Tabbert, K., Schweckendiek, J., Merz, C.J., Kagerer, S., Vaitl, D., Stark, R., 2009. Contingency learning in human fear conditioning involves the ventral striatum. *Human brain mapping* 30, 3636-3644.
- Kreutzmann, J.C., Fendt, M., 2020. Chronic inhibition of GABA synthesis in the infralimbic cortex facilitates conditioned safety memory and reduces contextual fear. *Translational Psychiatry* 10, 120.
- Kreutzmann, J.C., Jovanovic, T., Fendt, M., 2020. Infralimbic cortex activity is required for the expression but not the acquisition of conditioned safety. *Psychopharmacology* 237, 2161-2172.
- Kutlu, M.G., Zachry, J.E., Melugin, P.R., Cajigas, S.A., Chevee, M.F., Kelley, S.J., Kutlu, B., Tian, L., Siciliano, C.A., Calipari, E.S., 2021. Dopamine release in the nucleus accumbens core signals perceived saliency. *Current Biology*.

- Laing, P.A.F., Harrison, B.J., 2021. Safety learning and the Pavlovian conditioned inhibition of fear in humans: Current state and future directions. *Neuroscience & Biobehavioral Reviews* 127, 659-674.
- Laing, P.A.F., Vervliet, B., Fullana, M.A., Savage, H.S., Davey, C.G., Felmingham, K.L., Harrison, B.J., 2021. Characterizing human safety learning via Pavlovian conditioned inhibition. *Behaviour Research and Therapy* 137, 103800.
- Lanciego, J.L., Luquin, N., Obeso, J.A., 2012. Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine* 2, a009621.
- Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D., 2011. Differential roles of human striatum and amygdala in associative learning. *Nature neuroscience* 14, 1250.
- Lonsdorf, T.B., Menz, M.M., Andreatta, M., Fullana, M.A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shiban, Y., Schmitz, A., Straube, B., Vervliet, B., Wendt, J., Baas, J.M.P., Merz, C.J., 2017. Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews* 77, 247-285.
- Lysle, D.T., Fowler, H., 1985. Inhibition as a "slave" process: Deactivation of conditioned inhibition through extinction of conditioned excitation. *Journal of Experimental Psychology: Animal Behavior Processes* 11, 71-94.
- Mechias, M.-L., Etkin, A., Kalisch, R., 2010. A meta-analysis of instructed fear studies: Implications for conscious appraisal of threat. *NeuroImage* 49, 1760-1768.
- Merz, C.J., Hamacher-Dang, T.C., Stark, R., Wolf, O.T., Hermann, A., 2018. Neural Underpinnings of Cortisol Effects on Fear Extinction. *Neuropsychopharmacology* 43, 384-392.
- Meyer, H.C., Odriozola, P., Cohodes, E.M., Mandell, J.D., Li, A., Yang, R., Hall, B.S., Haberman, J.T., Zacharek, S.J., Liston, C., Lee, F.S., Gee, D.G., 2019. Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. *Proceedings of the National Academy of Sciences*, 201910481.

- Mohammadi, M., Bergado-Acosta, J.R., Fendt, M., 2014. Relief learning is distinguished from safety learning by the requirement of the nucleus accumbens. *Behavioural Brain Research* 272, 40-45.
- Mollick, J.A., Chang, L.J., Krishnan, A., Hazy, T.E., Krueger, K.A., Frank, G.K.W., Wager, T.D., O'Reilly, R.C., 2021. The Neural Correlates of Cued Reward Omission. *Frontiers in Human Neuroscience* 15.
- Morris, L.S., Kundu, P., Dowell, N., Mechelmans, D.J., Favre, P., Irvine, M.A., Robbins, T.W., Daw, N., Bullmore, E.T., Harrison, N.A., 2016. Fronto-striatal organization: defining functional and microstructural substrates of behavioural flexibility. *cortex* 74, 118-133.
- Neumann, D.L., Lipp, O.V., Siddle, D.A.T., 1997. Conditioned inhibition of autonomic Pavlovian conditioning in humans. *Biological Psychology* 46, 223-233.
- Ney, L.J., Akhurst, J., Bruno, R., Laing, P.A.F., Matthews, A., Felmingham, K.L., 2021. Dopamine, endocannabinoids and their interaction in fear extinction and negative affect in PTSD. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 105, 110118.
- Oyama, K., Tateyama, Y., Hernádi, I., Tobler, P.N., Iijima, T., Tsutsui, K.-I., 2015. Discrete coding of stimulus value, reward expectation, and reward prediction error in the dorsal striatum. *Journal of neurophysiology* 114, 2600-2615.
- Papalini, S., Beckers, T., Vervliet, B., 2020. Dopamine: from prediction error to psychotherapy. *Translational Psychiatry* 10, 164.
- Pauli, W.M., Larsen, T., Collette, S., Tyszka, J.M., Seymour, B., Doherty, J.P., 2015. Distinct Contributions of Ventromedial and Dorsolateral Subregions of the Human Substantia Nigra to Appetitive and Aversive Learning. *The Journal of Neuroscience* 35, 14220.
- Raczka, K.A., 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. *Translational Psychiatry* 1, e12-e12.
- Rescorla, R.A., 1969. Pavlovian conditioned inhibition. *Psychological Bulletin* 72, 77-94.

- Reynaud, O., Jorge, J., Gruetter, R., Marques, J.P., van der Zwaag, W., 2017. Influence of physiological noise on accelerated 2D and 3D resting state functional MRI data at 7 T. *Magnetic resonance in medicine* 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. *European Neuropsychopharmacology* 36, 137-150.
- Roesch, M.R., Esber, G.R., Li, J., Daw, N.D., Schoenbaum, G., 2012. Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of Neuroscience* 35, 1190-1200.
- Rogan, M.T., Leon, K.S., Perez, D.L., Kandel, E.R., 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, T.D., 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in cognitive sciences* 16, 147-156.
- Salinas-Hernández, X.I., 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, M.C., Foilb, A.R., Christianson, J.P., 2018. Inactivation of the Ventrolateral Orbitofrontal Cortex Impairs Flexible Use of Safety Signals. *Neuroscience* 379, 350-358.
- Schiller, D., Levy, I., Niv, Y., LeDoux, J.E., Phelps, E.A., 2008. From Fear to Safety and Back: Reversal of Fear in the Human Brain. *The Journal of Neuroscience* 28, 11517-11525.
- Schultz, W., Dickinson, A., 2000. Neuronal coding of prediction errors. *Annual review of neuroscience* 23, 473-500.
- Schultz, W., Tremblay, L., Hollerman, J.R., 2003. Changes in behavior-related neuronal activity in the striatum during learning. *Trends in neurosciences* 26, 321-328.
- 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* 5, 2470547021996006.

- Setsompop, K., Gagoski, B.A., Polimeni, J.R., Witzel, T., Wedeen, V.J., Wald, L.L., 2012. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine* 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, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of clinical psychiatry* 59, 22-33.
- Sosa, R., Ramírez, M.N., 2019. Conditioned inhibition: Historical critiques and controversies in the light of recent advances. *Journal of Experimental Psychology: Animal Learning and Cognition* 45, 17.
- Stelly, C.E., Girven, K.S., Lefner, M.J., Fonzi, K.M., Wanat, M.J., 2021. Dopamine release and its control over early Pavlovian learning differs between the NAc core and medial NAc shell. *Neuropsychopharmacology*.
- Tashjian, S.M., Zbozinek, T.D., Mobbs, D., 2021. A Decision Architecture for Safety Computations. *Trends in Cognitive Sciences* 25, 342-354.
- Teckentrup, V., van der Meer, J.N., Borchardt, V., Fan, Y., Neuser, M.P., Tempelmann, C., Herrmann, L., Walter, M., Kroemer, N.B., 2019. The anterior insula channels prefrontal expectancy signals during affective processing. *NeuroImage* 200, 414-424.
- Thiele, M., Yuen, K.S.L., Gerlicher, A.V.M., Kalisch, R., 2021. A ventral striatal prediction error signal in human fear extinction learning. *NeuroImage* 229, 117709.
- Tobler, P.N., Dickinson, A., Schultz, W., 2003. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *Journal of Neuroscience* 23, 10402-10410.
- Valentin, V.V., O'Doherty, J.P., 2009. Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. *Journal of neurophysiology* 102, 3384-3391.

- van Rooij, S.J.H., Jovanovic, T., 2019. Impaired inhibition as an intermediate phenotype for PTSD risk and treatment response. *Progress in Neuro-Psychopharmacology and Biological 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, A.R., Rescorla, R.A., 1972. Inhibition in Pavlovian conditioning: Application of a theory. *Inhibition and learning*, 301-336.
- 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. *Proceedings of the National Academy of Sciences* 116, 13077.
- Yau, J.O.Y., McNally, G.P., 2018. Brain mechanisms controlling Pavlovian fear conditioning. *Journal of Experimental Psychology: Animal Learning and Cognition* 44, 341.
- Yoshizawa, T., Ito, M., Doya, K., 2018. Reward-Predictive Neural Activities in Striatal Striosome Compartments. *eneuro* 5, ENEURO.0367-0317.2018.
- Zhang, Z., Mendelsohn, A., Manson, K.F., 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.0072-0015.2015.
- 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, B.R., 2020. Diagnostic and Predictive Neuroimaging Biomarkers for Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 5, 688-696.

Figures

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 inter-trial interval (ITI, 8-12s) 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.

Figure 2. Behavioral measures. (A) Threat-expectancy ratings showed a significant decreased 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 to the threat compound AD+, but did not differ from one another. (D) SCRs showed no differences at test. (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 represent 95% confidence intervals. * = $p < 0.05$,

*** = $p < 0.001$, ns = $p > 0.05$.

Figure 3. Brain regions with significant differential response to the conditioned inhibitor (AX- > BC-) versus the standard safety signal (BC- > AX-). Whole-brain FDR-corrected ($P < 0.05$) results are displayed on a high-resolution anatomical template in MNI space.

Figure 4. 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 anatomical template in MNI space.

Figure 5. 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 anatomical template in MNI space.

Tables

Table 1. Differential safety responses – conditioned inhibitor versus standard safety signal. Whole brain contrasts estimated at $P_{FDR} < 0.05$.

Regions	MNI coordinates				
	x	y	z	K _E	Z
<i>Inhibitor > standard safety (AX- > BC-)</i>					
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
<i>Standard safety > inhibitor (BC- > AX-)</i>					
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

Note. ' K_E ' = cluster size in number of voxels. ' Z ' = SPM Z scores. Coordinates reported in MNI space.

Table 2. Safety learning dynamics – early versus late conditioning. Whole brain contrasts estimated at $P_{FDR} < 0.05$.

Regions	MNI coordinates			K_E	Z
	x	y	z		
<i>Conditioned inhibitor (AX- early > AX- late)</i>					
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
<i>Standard safety (BC- early > BC- late)</i>					
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

Note. ' K_E ' = cluster size in number of voxels. ' Z ' = SPM Z scores. Coordinates reported in MNI space.

Table 3. Safety conditioning versus test phase. Whole brain contrasts estimated at $P_{FDR} < 0.05$.

Regions	MNI coordinates				
	x	y	z	K _E	Z
<i>Conditioned inhibitor (AX- > DX-)</i>					
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 grey)	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
<i>Standard safety (BC- > DC-)</i>					
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

Note. ' K_E ' = cluster size in number of voxels. ' Z ' = SPM Z scores. Coordinates reported in MNI space.

Extended Data

Figure 3-1. 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.

Figure 4-1. 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).

Figure 5-1. 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.