INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus

Biologies

Cyril Herry and Daniel Jercog

Stable coding of aversive associations in medial prefrontal populations

Volume 346 (2023), p. 127-138

Published online: 20 December 2023

https://doi.org/10.5802/crbiol.126

Les Comptes Rendus. Biologies *sont membres du*
Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 1768-3238

Comptes Rendus
Biologies
2023, Vol. 346, p. 127-138
https://doi.org/10.5802/crbiol.126

Research article

Stable coding of aversive associations in medial prefrontal populations

Cyril Herry[*, a, b] **and Daniel Jercog**[*, a, b]

[a] INSERM, Neurocentre Magendie, U1215, 146 Rue Léo-Saignat, 33077 Bordeaux,
France

[b] Univ. Bordeaux, Neurocentre Magendie, U1215, 146 Rue Léo-Saignat,
33077 Bordeaux, France

E-mails: cyril.herry@inserm.fr (C. Herry), daniel.jercog@gmail.com (D. Jercog)

Abstract. The medial prefrontal cortex (mPFC) is at the core of numerous psychiatric conditions, including fear and anxiety-related disorders. Whereas an abundance of evidence suggests a crucial role of the mPFC in regulating fear behaviour, the precise role of the mPFC in this process is not yet entirely clear. While studies at the single-cell level have demonstrated the involvement of this area in various aspects of fear processing, such as the encoding of threat-related cues and fear expression, an increasingly prevalent idea in the systems neuroscience field is that populations of neurons are, in fact, the essential unit of computation in many integrative brain regions such as prefrontal areas. What mPFC neuronal populations represent when we face threats? To address this question, we performed electrophysiological single-unit population recordings in the dorsal mPFC while mice faced threat-predicting cues eliciting defensive behaviours, and performed pharmacological and optogenetic inactivations of this area and the amygdala. Our data indicated that the presence of threat-predicting cues induces a stable coding dynamics of internally driven representations in the dorsal mPFC, necessary to drive learned defensive behaviours. Moreover, these neural population representations primary reflect learned associations rather than specific defensive behaviours, and the construct of such representations relies on the functional integrity of the amygdala.

Keywords. Associative learning, Medial prefrontal cortex, Population coding, Machine learning.

Note. Daniel Jercog received the Grandes Avancées en Biologie 2022 award.

Funding. French National Research Agency (Grant agreement nos. ANR-DOPAFEAR, ANR-PRELONGIN, ANR-10-EQPX-08 OPTOPATH), Conseil Regional d'Aquitaine, and Fondation pour la Recherche Médicale (Grant agreement no. FRM-Equipes FRM 2017).

Published online: 20 December 2023

1. Introduction

Associative learning is the process by which relationships among various stimuli, behaviours, and outcomes are acquired [1]. The study of the neural basis of aversive associative learning has played a major role in explaining the development and treatment of fear-related disorders [2]. Indeed, such conditions are often thought to result from deficits in associative learning processes [3]. From a neuronal standpoint, forming aversive associative memories is intrinsically linked to the representation of different

[*] Corresponding authors.

ISSN (electronic) : 1768-3238

elements that constitute them, including threat-predicting cues, aversive outcomes, behavioural responses as well as the contextual environment, which combination can be conceived as the occurrence of a specific threatening situation [4]. Previous studies have provided invaluable insights about the brain circuits and functions involved in defensive behaviour and fear learning by studying single neuron responses in isolation or estimating cell assemblies on specific defined timescales [5]. A number of such studies investigated how medial prefrontal cortex neurons and dedicated projections could regulate threat-related behaviour. While prefrontal neurons have been shown to exhibit tone selective or freezing selective responses during fear expression and extinction learning [6, 7], prefrontal-amygdala projections have been shown to exert control over the expression of threat-related behaviours [8]. In addition, theta-related oscillations and prefrontal-amygdala synchrony have been associated as a mechanism for signalling safety or fearful states [9–12]. Beside the study of the role of single neurons in threat-related behaviour, an idea that it is becoming increasingly prevalent in systems neuroscience is that neuronal populations are in fact the essential unit of computation in many integrative brain regions [4, 13]. A central idea behind population coding is that a downstream area integrates the heterogeneous activity from multiple neurons to determine some value about the inputs (Figure 1A). Such coding strategy has numerous advantages including an increased information and sensitivity, accurate representations, robustness to the effects of noise in individual neuron representations, and the ability to represent complex stimuli [14]. But what information is conveyed by prefrontal populations during fear behaviour? To address this question, we performed electrophysiological single-unit population recordings in the dorsomedial prefrontal cortex (dmPFC) of mice during fear behaviours during learning, extinction and reversal of instrumental aversive conditioning. We found that threat-predicting cues, but not neutral cues, induced sustained neuronal representations consistent with a stable coding of information in dmPFC networks that primarily reflects the aversive associations rather than specific defensive behaviours. Pharmacological inactivation experiments showed that such dmPFC threat representations require the functional integrity of the amygdala, whereas temporal specific optogenetical inhibition of the dmPFC showed the critical involvement of these representations in the initiation of defensive actions.

2. Results

2.1. *Associated threats drive sustained population representations in dmPFC networks*

To investigate what are medial prefrontal neuronal populations collectively reflecting about defensive behaviours, mice were initially implanted with electrode bundles targeting the dmPFC and submitted to a differential active avoidance task [15] (Figure 1B). After a habituation session to the CS^+ and CS^- sounds (50 ms sound-pips at 1 Hz, see Section 4), mice were trained in the task during 4 consecutive days. In this task, mice learned to avoid a conditioned stimulus (CS^+) that predicts the delivery of a mild foot shock (US) by shuttling between two symmetric compartments separated by a small hurdle. A second conditioned stimulus not associated with the foot shock (CS^-) was used as an internal control. After CS onset, shuttling from the current compartment within a 7 s period was defined as an avoided CS response, leading to the termination of the CS and preventing US delivery for CS^+ trials. In contrast, remaining in the same compartment led to US delivery for CS^+ trials after 7 s. After a habituation session where both CS are presented but not reinforced, mice were trained in our task and learned to selectively avoid CS^+ while CS^- avoidance remained low and similar to inter-trial shuttling levels.

CS^+ presentation induced a modulation of a large fraction of dmPFC neurons compared to CS^- when comparing their averaged activity (Figure 1C). To address how stimuli information is represented in dmPFC populations, we used a decoding approach based on the collective recorded dmPFC neuronal activity across animals [15, 16]. Here, the firing activity from the ensemble of recorded cells at a given time point t (so-called pseudo-population vector) is compared for CS^+ and CS^- trials versus the spontaneous baseline activity preceding each stimuli (see Section 4). For each time point, we trained a set of linear classifiers designed to maximally separate the dmPFC evoked population patterns for both CS/baseline conditions, while quantifying their

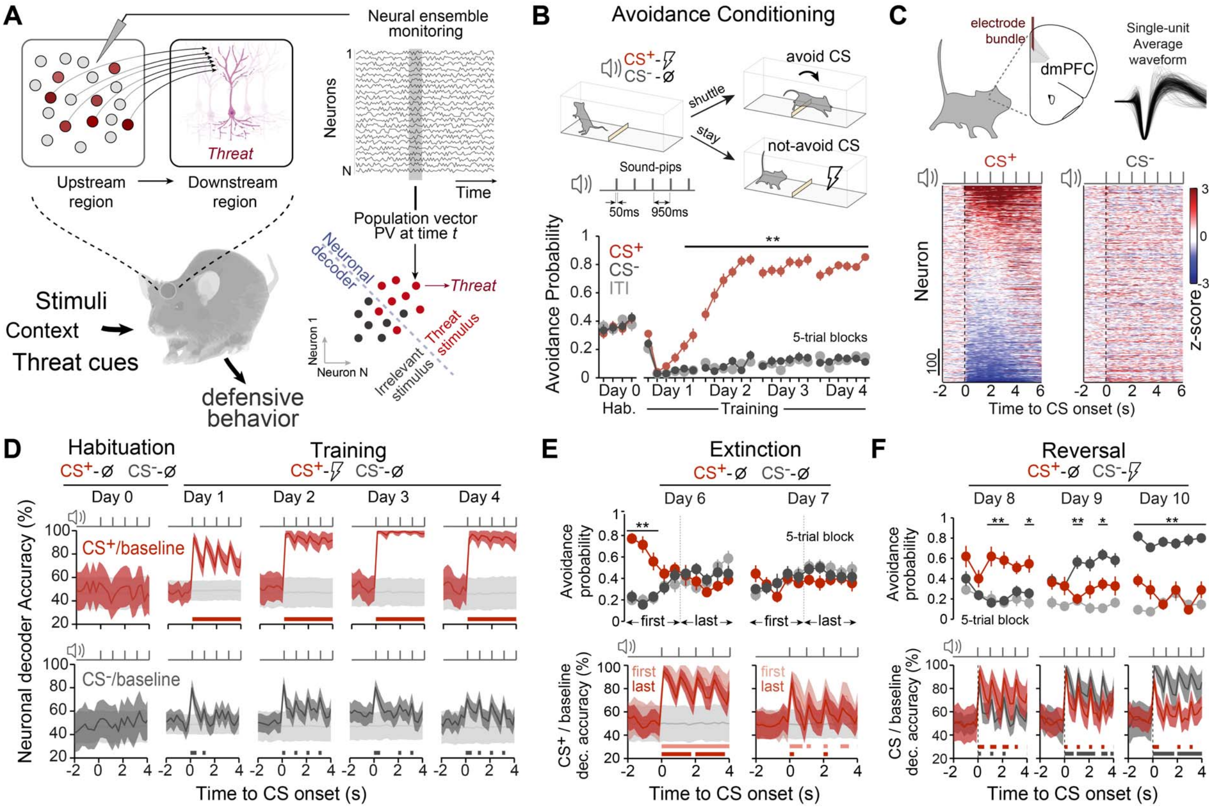

Figure 1. Aversive associative learning correlates in dorsomedial prefrontal neuronal populations. (A) Multiple neurons encode heterogeneous information about the external world, and the readout of this collective information from downstream neurons can be used to establish some precise value about external inputs. We monitored the activity of individual neurons from a certain region and look at the collective activity pattern across recorded cells for a given time point (so-called population vector, PV). We can assess the presence of information on those population patterns about a certain condition by comparing different patterns using neuronal decoders. (B) Top: Schematics of the cued differential avoidance task in mice and auditory cues structure. Bottom: learning curves (ITI, inter-trial-interval). Selective avoidance of CS$^+$ from the 5th 5-trial block on Day 1 ($^{**}P < 0.007$, two-way repeated-measures ANOVA. $N = 34$ mice). (C) Single-unit z-scored activity for CS$^+$ (left) and CS$^-$ (right) ordered by magnitude of CS$^+$ responses (day-3 and -4 data pooled: 34 mice, 68 sessions, 1261 units). (D) dmPFC population-activity-based decoding accuracies for CS$^+$ (top) or CS$^-$ (bottom) from baseline activity across days (shuffled trial label in grey; note the transient increases in accuracies during sound-pips). (E) Mean avoidance probability during extinction learning ($n = 14$ mice). Although CS$^+$ avoidance behaviour (top) was extinguished during the first extinction session ($^{**}P < 0.0110$ for 1st and 3rd block, two-way RM ANOVA), CS$^+$ from baseline decoding accuracy (bottom) was reduced only on the subsequent extinction session. (F) Mean avoidance probability during reversal learning ($n = 11$ mice). Previous CS$^+$ tone became neutral and the previous CS$^-$ tone was associated with the US. Avoidance behaviour was progressively reversed during training (top) ($^{**}P < 0.01$ $^{*}P < 0.05$, two-way RM ANOVA) and decoding accuracies changed accordingly (bottom). Accuracy data are mean ±1 s.d. Significant decoding accuracy periods over shuffle accuracies were represented by thick lines ($P < 0.05$, permutation test).

cross-validated decoding accuracy with the decoding accuracy obtained from shuffling the identity of the trial conditions. While during the habituation session to the auditory stimuli (before conditioning), decoding stimuli information from dmPFC populations was not significant from shuffle, training induced a progressive increase towards high and sustained (in between sound-pips) decoding accuracy values for CS$^+$, an observation in sharp contrast with CS$^-$ trials which displayed modest accuracy levels throughout training (Figure 1D).

In our task CS$^+$ aversiveness is acquired through associative learning, therefore we evaluated how avoidance extinction or reversal of discriminative avoidance learning could influence CS decoding accuracy. Extinction training in consecutive sessions progressively decrease CS$^+$ decoding accuracy towards non-significant values (Figure 1E). Following avoidance extinction, we performed a reversal experiment during which we paired the previous CS$^-$ with the US while the previous CS$^+$ became not reinforced. While mice progressively switched their avoidance responses towards the reversed CS$^-$ and did not avoid the reversed CS$^+$ during subsequent training sessions, this reversal was also observed in terms of decoding accuracies (Figure 1F).

Together, these results show that dmPFC neuronal populations reflect associative processes during learned fear, and neural representations during threat-predicting cues are sustained even in the absence of sensory inputs.

2.2. *Sustained threat-related representations are not explained by specific defensive behaviours*

We next tested if CS$^+$ decoding accuracy levels reflected certain animals' defensive response at CS$^+$ onset. We first observed that spontaneous freezing episodes outside of CSs presentation could be decoded from CS$^+$ trials, indicating that high decoding accuracy observed at CS$^+$ onset did not merely reflect freezing behaviour (Figure 2A). Indeed, while dmPFC populations carried information as to whether mice were freezing or not during CS$^+$ presentations, the accuracy levels observed were somewhat limited (Figure 2B). Alternatively, to test that high accuracy at CS$^+$ onset was not just related to an action preparation state, after avoidance conditioning learning, mice were subjected to a "Confined task" where the placement of a wall between compartments precluded the avoidance response and non-reinforced CS$^+$ and CS$^-$ were presented, which induced a switch in behaviour to a freezing strategy upon CS$^+$ presentations (Figure 2C, left). This condition also induced sustained CS$^+$ representations, despite avoidance was prevented (Figure 2C, right). Thus, to compare both experimental conditions we trained decoders on the Confined task and tested the performance of these decoders on the Active Avoidance task data (and vice versa). In this condition CS$^+$ evoked neuronal population patterns were still correctly classified despite the change in active or passive defensive strategies, indicating that accuracy observed at CS$^+$ onset in the Active Avoidance task was not merely reflecting the avoidance action preparation (Figure 2D). Accordingly, the direct comparison by decoding avoided from non-avoided CS$^+$ trials did not differ from shuffle accuracies, reflecting that encoded information at CS$^+$ onset was not related to impending avoidance actions (Figure 2E).

We consistently observed sustained CS$^+$ population representations even in the absence of sensory inputs (i.e. in between sound-pips), despite the heterogeneous individual neuronal response dynamics (Figure 1C). Indeed, dmPFC populations during CS$^+$ described strong dynamics led by a few dominant features capturing a large portion of the variance (Figure 2F). In this scenario, performing a cross-temporal decoding of CS$^+$ from baseline activity showed that coding dynamics of CS$^+$ was largely consistent with a stable coding of information at CS onset (Figure 2G).

Altogether, these results suggest that high information observed by neuronal population patterns in the dmPFC during CS$^+$ presentations displayed dynamics of information coding consistent with a stable coding, and such representations convey information about the learned association rather than specific defensive behaviours.

2.3. *dmPFC threat representations depend on amygdala and are necessary to drive defensive actions*

We hypothesised that information related to threat-predicting cues is acquired in remote brain areas and integrated into dmPFC networks. Because amygdala

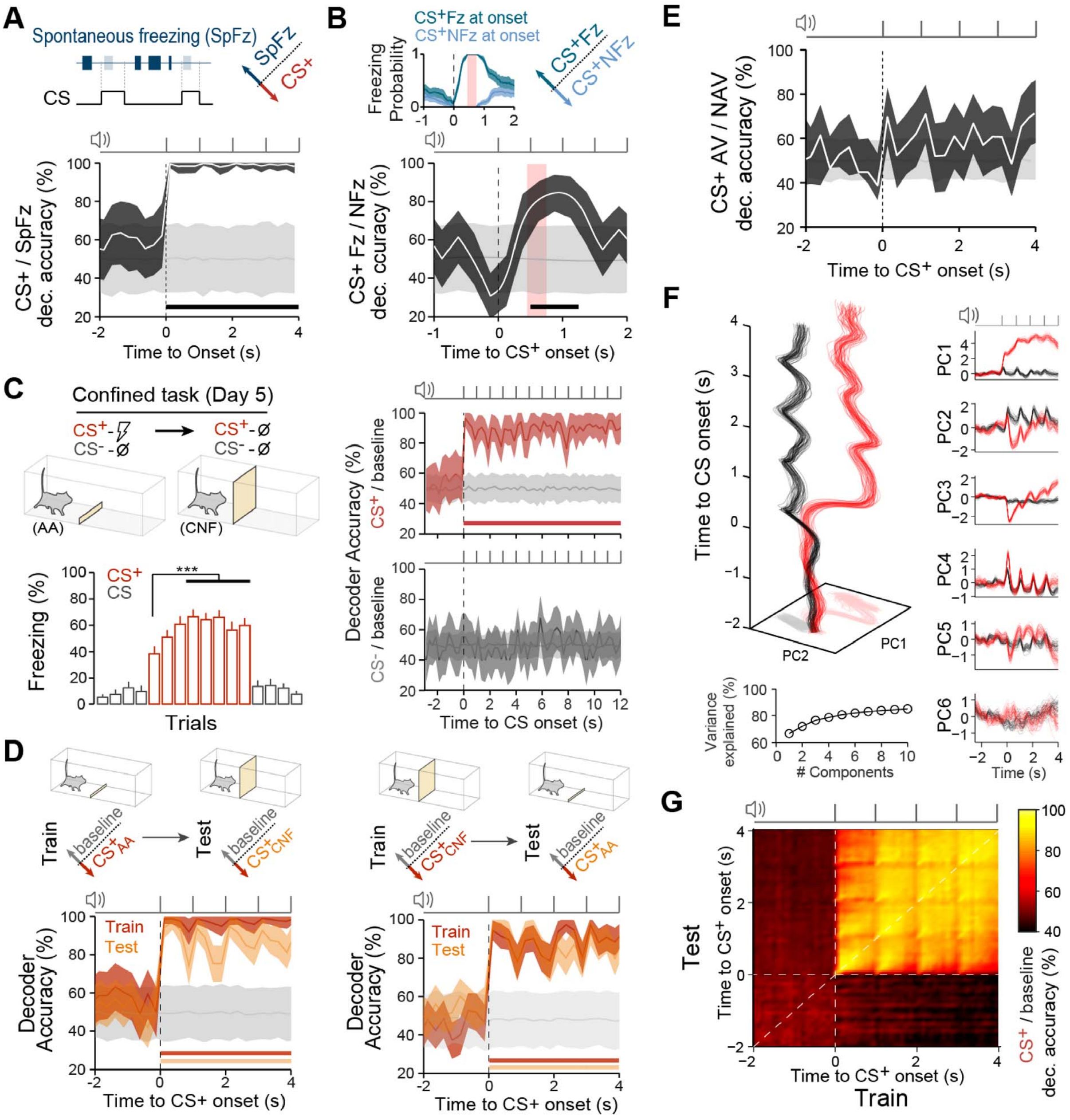

Figure 2. Associated threat representations by dmPFC neuronal population are not specific for defensive behaviors. (A) Decoding Spontaneous freezing events (outside CS) and CS$^+$ evoked activity. (B) Decoding CS$^+$ trials when mice exhibit freezing (CS$^+$ Fz) or non-freezing (CS$^+$ NFz) at CS onset. Inset shows freezing probabilities for both trial types (error bars display 95% CI), light red area indicate the period during which animals showed pure freezing or non-freezing bouts (n = 13 sessions). (C) After training, mice were confined to one of the compartments where unreinforced conditioned stimuli were presented (confined task). Freezing to CS$^+$ progressively increases during the first three CS$^+$ trials (5th trial) and reaches stable levels from trial 8 to 12 ($^{***}P < 0.001$, one-way RM ANOVA; n = 21 mice). Decoding accuracy of CS$^+$ and CS$^-$ from baseline population activity during the confined task (right).

Figure 2. (**cont.**) (D) Trained decoders to classify CS^+ trials from baseline in the active avoidance task (CS^+_{AA}) displayed high accuracy when decoding CS^+ in the confined task (CS^+_{CNF}) (left), and vice versa (right). (E) Decoding accuracy between CS^+ avoided and CS^+ non-avoided trials at CS-onset was not different from shuffle values. (F) Population trajectories at CS onset defined via PCA on trial-averaged activity (50% trials for each session and trial condition; 100 repetitions). Inset showing variance explained and dynamics of first 6 principal components. (G) Cross-temporal decoding of CS^+ from Baseline activity (activity patterns in 0.1 s bins), consistent with stable dynamics of information coding. Accuracy data are mean ±1 s.d. Significant decoding accuracy periods over shuffle accuracies were represented by thick lines ($P < 0.05$, permutation test).

subnuclei are involved in active avoidance behaviour [17, 18], we next studied how amygdala CSs processing impacts its representation in dmPFC networks. Consistent with previous studies [18], we observed that amygdala inactivation using muscimol impaired avoidance behaviour (Figure 3A). Importantly, dmPFC decoding analyses during amygdala inactivation revealed that threat-predicting cues were still encoded during individual pips, but that the maintenance of information between pip presentations was nearly absent (Figure 3B). Altogether, these results indicate that dmPFC populations at CS onset represent threat information in a sustained manner and through associative processes, in that these representations are amygdala-dependent.

To test the causal role of such threat representation in dmPFC networks, we used an optogenetic inactivation approach on dmPFC excitatory neurons (Figure 3C, left). Whereas photo-inhibition of dmPFC excitatory neurons during CS^- presentations had no effect compared to GFP controls or non-stimulated trials, it strongly impaired avoidance behaviour during CS^+ trials (Figure 3C, right). In addition, avoidance impairment was not explained by a switch in defensive behaviour, as freezing levels during CS^+ presentations were unaltered (Figure 3D). To further evaluate whether representation of threat-predicting cues was necessary to drive the selection of active avoidance behaviour, we perform temporally specific optogenetic inhibition of the dmPFC specifically around CS onset (Figure 3E). Importantly, we observed that inhibiting dmPFC activity at CS onset delayed the initiation of avoidance behaviour (Figure 3E, left). Conversely, dmPFC inhibition after CS onset impaired active avoidance behaviour (Figure 3E, right). Altogether, these circuit manipulations indicate that generation of CS-induced threat information was amygdala-dependent, but importantly this information was integrated within the dmPFC network to generate sustained threat representations. In contrast, avoidance action initiation critically relied on dmPFC activity after CS onset, indicating that dmPFC links threat information with actions to ultimately drive active defensive responses.

3. Discussion

Using a combination of behavioural, pharmacological and optogenetic approaches along with single-unit recordings and neuronal decoding techniques, we showed that threat representations in dmPFC populations mainly represent learned associations rather than specific defensive behaviours. Moreover, these representations exhibiting a stable coding dynamics are necessary to bridge threat-predicting stimuli to defensive actions. Our data also indicate that the dmPFC threat representations depend on amygdala functional integrity. Moreover, the dmPFC optogenetic inhibition during CS presentation selectively impaired active avoidance behaviour without impacting animals' threat assessment or altering the selection of defensive behaviour. Finally, our data indicate that whereas dmPFC inhibition restricted to CS onset delayed avoidance behaviour, the same manipulation performed after CS onset reduced avoidance probability.

We observed that discrimination of threatening and non-threatening CSs is not impaired during inhibition of the dmPFC, and that the decoding accuracy of CS^+ pips is preserved in the dmPFC upon amygdala inactivation. These data first suggest that threat-related information is acquired in remote brain areas and subsequently represented in dmPFC networks. This idea is consistent with previous results indicating that the formation of fear-related CS-US associations occurs in subcortical networks [19–22] and is integrated by the mPFC to drive defensive-related behaviours [23–25]. In addition, previous

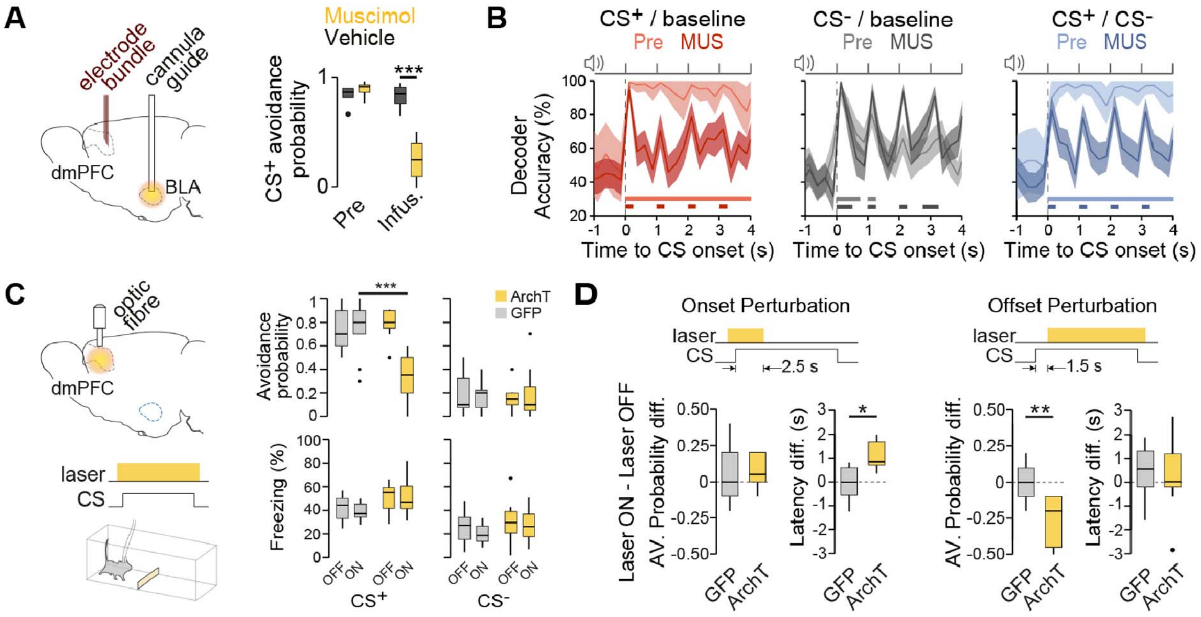

Figure 3. Aversive association representations requires the functional integrity of amygdala and are critical to drive defensive actions. (A) Muscimol inactivation of the amygdala impaired avoidance behaviour. $^{***}P < 0.0001$, two-way repeated-measures ANOVA; muscimol, $n = 6$ mice; vehicle, $n = 5$ mice; dots indicate outliers. Pre, pre-infusion session; Infus., infusion session. BLA, basolateral amygdala. (B) Impact of amygdala inactivation on dmPFC CS decoding accuracies comparing pre-infusion (light colours) with muscimol infusion sessions (Mus, dark colours). (C) The dmPFC was infected with either ArchT or GFP under the CaMKII promoter, and continuous light was delivered during CS presentations for half of the trials (ON/OFF laser conditions). CS^+ avoidance was significantly reduced upon dmPFC inhibition ($^{***}P < 0.0001$, repeated measures mixed-effects model; GFP $n = 9$ mice, ArchT $n = 12$ mice). No significant effect was observed on CS-evoked freezing during laser OFF and ON conditions ($P > 0.05$, repeated-measures mixed-effects model). (D) Photostimulation restricted at CS onset (Onset Perturbation; GFP $n = 8$ mice, ArchT $n = 5$ mice) did not changed avoidance probability (unpaired t-test, $P = 0.76$) but significantly delayed avoidance responses (unpaired t-test, $^{*}P = 0.01$). Photostimulation restricted to after CS onset periods (Offset Perturbation; GFP $n = 12$ mice, ArchT $n = 13$ mice) reduced avoidance probability (unpaired t-test, $^{**}P = 0.001$) whereas avoidance was not delayed (unpaired t-test, $P = 0.86$). Thick lines indicate significant decoding accuracies ($P < 0.05$, permutation test). Accuracy data are mean ±1 s.d.

studies demonstrate that amygdala inactivations [8] or lesions [26] alter dmPFC single-cell responses to cue-predicting threats. Our data indicate rather than CS^+ sensory information encoding in the dmPFC is largely independent of the amygdala, but that amygdala processing of threatening cues is critical for the maintenance of sustained threat representations in the dmPFC. Second, the fact that dmPFC inhibition reduces freezing upon threat presentations during tone fear conditioning [27, 28] but remains unaffected during avoidance tasks despite the reduction in avoidance behaviour [18], highlights the flexibility of the dmPFC in the control of passive or active defensive responses depending on contextual inputs. Third, although a current view postulates that the basolateral amygdala (BLA) mediate avoidance behaviour [17, 29, 30], our data demonstrate that the amygdala allows sustained threat representations in the dmPFC, that ultimately links threats with the initiation of active avoidance responses.

Despite the heterogeneous dynamics in individual cells' activity, we observed a relatively stable dynamics of information coding by dmPFC neuronal populations during an ongoing associated threat, consistent to what is observed during the delay period of working memory tasks [31]. The resemblance in how mnemonic and threat information is encoded by neuronal populations suggests that stable coding is a fundamental computational principle in the prefrontal cortex for connecting associated stimuli to goal-directed actions.

We observed that sustained threat representations are largely not reflecting sensory inputs or specific defensive behaviours. What these threat representations actually represent? Overall, our data is consistent with the general view that the dmPFC guide actions specifically based on internal outcome expectations [32]. However, a precise demonstration if our observed threat representations convey information about the expected aversive outcomes require complex behavioural paradigms. Indeed, a recent study in mice performing an approach-avoidance task suggests that neuronal population representations are liked with the value of expected aversive outcome [33].

Finally, our results highlight a dynamic process of encoding sensory-threat and goal-directed avoidance states by the concerted activity of populations of cells in the dmPFC. Such changes in prefrontal representations have been described during contextual changes in appetitive-based sensory-decision representations in non-human primates [34]. In our study, we observed that dmPFC neuronal population dynamics bridge a transition from representing an aversive state to representing an active defensive response. This sustained, multi-representational, transfer of environmental information by the dmPFC is required for accurate threat encoding and the computations of action-outcome, which ultimately control the selection of active defensive responses strategies.

4. Methods

All procedures were performed in accordance with standard ethical guidelines (European Communities Directive 86/60-EEC) and were approved by the committee on Animal Health and Care of Institut National de la Santé et de la Recherche Médicale and French Ministry of Agriculture and Forestry (agreement #A3312001). Male C57BL6/J mice (Janvier) aged 8–9 weeks were individually housed under a 12 h light–dark cycle, and provided with food and water ad libitum. Active Avoidance task was performed in a shuttlebox comprised of a plexiglass box (40 × 10 × 30 cm) with a floor grid connected to a shocker, where a small plastic hurdle (1 cm height) divided the arena into two equal compartments while infrared beams detection automatically monitored the mice shuttling between compartments (Imetronic). Shuttlebox was enclosed inside an acoustic foam isolated box where two speakers mounted on top of each compartment delivered the auditory conditioned stimulus (CS) consisting in either 7 kHz or white-noise 50 ms pips at 1 Hz (maximum 13 pips, i.e. CS maximum duration of 12 s), 2 ms rise and fall, 80 dB sound pressure level. Scrambled foot shocks of 50 ms (~7 Hz) at an intensity of 0.7 mA and maximum duration 5 s applied through the grid floor (Imetronic) served as an unconditioned stimulus (US). LEDs mounted on top of each compartment provided house light. During the Confined task, an additional opaque white wall was placed in the middle of the maze preventing shuttling between compartments.

4.1. *Behavioural paradigm*

4.1.1. *Active avoidance task*

Mice (n = 34, implanted) were first habituated to the context and tones (Day 0). During the habituation, alternating 20 CS^+ and 20 CS^- were presented without any US. Training days (Day > 0) consisted of randomly intermingled 30 CS^+ and 30 CS^-, where CS^+ trials were paired with the US if mice did not shuttle compartment before 7 s after CS^+ onset. CSs started independently from the animal location in the shuttle box after an inter-trial interval of 25–40 s, and shuttling between compartments stopped any ongoing CS or US. The first CS^+ trial of the first training session was always paired with the US disregarding the behaviour of the animal. All sessions started with 2 min acclimation periods before the first CS presentation. In a subset of mice (n = 9, non-implanted), the tones used for CS^- and CS^+ were switched.

4.1.2. *Confined task*

On Day 5, a fraction of the mice (n = 21, implanted) were trained in the Confined task. The session started as an active avoidance training session of intermingled 15 CS^+ and 15 CS^- trials. Next, a plastic wall was introduced at the hurdle position, preventing mice from shuttling between compartments. In this condition, a sequence of 4 CS^- followed by non-reinforced 8 CS^+, followed by 4 CS^- was presented. CSs were unavoidable/unescapable, therefore lasted for 12 s (inter-trial interval 25–40 s). Two minutes after the last trial, the plastic wall was removed and the active avoidance training session was resumed with intermingled 15 CS^+ and 15 CS^- trials. Mice were never disconnected during the entire session.

4.1.3. *Extinction training*

On Day 5 and 6, mice (n = 14, implanted) were submitted into an extinction training procedure consisting of non-reinforced 50 CS^+ and 50 CS^- trial presentations. The first 5 CS^+ of the first extinction session were still conditionally paired with the US to induce high aversion towards CS^+ over a longer timescale.

4.1.4. *Reversal training*

After 2 days of extinction training, mice (n = 14, implanted) were submitted to 3 consecutive reversal training sessions consisting of sessions analogous to the active avoidance training sessions but where tones associated with the CSs during the active avoidance training sessions were exchanged, using the same structure as the aforementioned Active Avoidance task. Therefore, CS^- was paired with the US (rCS^+) while CS^+ was not reinforced (rCS^-). Only mice displaying rCS^+ avoidance probability greater than 0.3 after 3 days of reversal training were included in the analyses (11/14 mice).

4.2. *Electrode implantation and electrophysiological recordings*

Mice (10 weeks old) were anaesthetised with isoflurane (induction 3%, maintenance 1.5%) in O_2. Body temperature was maintained at 37 °C with a temperature controller system (FHC), and eyes were hydrated with Lacrigel (Europhta Laboratories). Mice were placed in a stereotaxic frame (Kopf Instruments), and 3 stainless steel screws were attached to the skull. Following craniotomy, mice were bilaterally implanted in the dmPFC with electrode array targeting the following coordinates relative to bregma: +2.0–2.1 mm AP; ±0.55–0.7 mm ML; and 1.20–1.30 mm DV from dura with an angle of 14°. Each electrode bundle consisted of 16 individually insulated nichrome wires (13 mm diameter, impedance 60–100 KU; Kanthal) fixed to an electrode guide. Each electrode bundle was attached to one 18-pin connector (Omnetics). Connectors were referenced and grounded via a silver wires (127 μm diameter, A-M Systems) placed above the cerebellum. All implants were secured using Super-Bond cement (Sun Medical). During surgery, long- and short-lasting analgesic agents were injected (Metacam, Boehringer; Lurocaïne, Vetoquinol). After surgery, mice were allowed to recover for at least 10 days. Electrodes were connected to a headstage (Plexon) containing sixteen unity-gain operational amplifiers. Each headstage was connected to a 16-channel PBX preamplifier where the signal was replicated and bandpass-filtered at 300 Hz and 8 kHz and at 0.5 Hz and 200 Hz for local field potential recordings. Spiking activity was digitised at 40 kHz and isolated by time-amplitude window discrimination and template matching using an Omniplex system (Plexon). Single-unit spike sorting was performed using Off-Line Spike Sorter (OFSS, Plexon), where Pairwise-P (multivariate ANOVA) statistics was used to assess unit isolation quality ($P < 0.05$). At the end of the experiment, electrolytic lesions were administered before transcardial perfusion to verify electrode tip location using standard histological techniques.

4.3. *Muscimol inactivations*

Mice were bilaterally implanted with stainless steel cannula guide (26 gauge; Plastics One) aimed at the PFC (n = 12 mice) or BLA (n = 22 mice). To target the dmPFC, guides were implanted following coordinates relative to bregma: +2.1 mm AP; ±0.55 mm ML; and 0.7 mm DV; with a 14° angle. To target the BLA, guides were implanted following coordinates relative to bregma: −1.4 mm AP; ±3.7 mm ML; 3.5 mm DV from dura; with a 4° angle. Cannula guides were secured using Super-Bond cement (Sun Medical). Dummy cannulas were used to fill the guide and removed only during the injection period. For a subset of mice implanted with cannula guides in the BLA (n = 11 mice), were also implanted with recording electrodes in the dmPFC as described in the “Electrode implantation and electrophysiological recordings” methodological section. On the injection day, muscimol (Sigma; 0.25 μg/μL mM in PBS 0.1 M, pH = 7.2–7.3; based on [19]) or vehicle (PBS0.1 M, pH = 7.2–7.3) was infused at a rate of 0.1 μL/min (total volume of 0.15 μL for dmPFC, 0.2 μL for BLA) with 0.5 mm protruding injector cannulas. After infusion, injecting cannulas were left in place for 5 min to allow drug diffusion. Muscimol was infused 40 min before the behavioural test.

4.4. *Virus injections and optogenetics*

For optical silencing of dmPFC CaMKIIa-expressing neurons, 0.15–0.2 μL of either GFP (AAV5-CaMKIIa-GFP, titer 1.5 × 10^{12}, UNC Vector Core Facility) or ArchT (AAV5-CaMKIIa-ArchT-GFP, titer 5.3 × 10^{12}, Addgene) were bilaterally injected into the dmPFC of 8/9 weeks old wild-type mice from glass pipettes (tip diameter 20–30 μm) at the following coordinates relative to bregma: +2.1 mm AP; ±0.55–0.65 mm ML; 1.15–1.3 mm DV from dura; with a 14° angle. At 2–3 weeks after the injection, mice were implanted bilaterally with custom-built optic fibers (diameter: 200 mm; numerical aperture: 0.39; Thorlabs) above the dmPFC at the following coordinates relative to bregma: +2.1 mm AP; ±0.65 mm ML; −0.65 mm DV from dura; with a 14° angle. For optical silencing BLA to dmPFC projecting cells, we injected 0.15–0.2 μL retro-ArchT (ssAAV-retro/2-mCaMKIIalpha-eArchT3.0, titer 8.8 × 10^{12}, Zurich Viral Vector Facility) into the dmPFC using the same coordinates and procedure. At 4 weeks after injection, mice were bilaterally implanted with commercial optic fibers (diameter: 400 mm; numerical aperture: 0.66; Doric) targeting the upper part of the BLA, guides were implanted following coordinates relative to bregma: −1.4 mm AP; ±3.7 mm ML; 3.5 mm DV from dura; with a 4° angle. Only mice with correct placement of optic fibers or virus expression restricted to dmPFC were included in the analyses.

Three different protocols for optogenetic stimulation at a continuous green light (532 nm, ~6–8 mW at fiber tip) were used: (i) from 1 s before CS onset until 1 s after CS offset, for global CS perturbation (n = 16 mice); (ii) from 1 s before CS onset until 2.5 s after CS onset, for the onset perturbation experiments (n = 13 mice) and (iii) from 1.5 s after CS onset until 1 s after CS offset, for the offset perturbation experiments (n = 25 mice). The optogenetics protocol structure for different conditions was the first half of the trials within a session were non-stimulated, followed by the second half of the trials being stimulated, providing an internal control of light effect for each condition (except retro-ArchT experiments, where also all trials were stimulated on Day 4).

Data analysis

To get an estimate of the neuronal responses for different trial types or conditions, we included recording sessions from Day 3 and 4 (stabilized performance). Due to the low number and behavioural variability observed in CS^- avoided trials, we consider in the analysis CS^- as the CS^- non-avoided trials. CS^+ avoided trials considered for CS onset aligned related analyses were those having avoidance shuttle latency above 4 s.

To address how informative the firing rates of the dmPFC cells ensembles were about stimulus identity, threat cues, avoidance action and behavioural states, we used linear-kernel Support Vector Machine (SVM) classifiers. Individual trial pseudo-population firing rate vectors were constructed for each 250 ms time bin using units and trials from different animals, therefore removing decoding contributions due to noise correlations. In every case, the number of trials was equalised in both classes by randomly subsampling the overrepresented class, and randomly sampling the minimum amount of available trials across animals. For each time bin, an independent set of SVM classifiers were used to perform our decoding accuracies estimations and statistical tests. The decoding accuracy of a classifier (5-fold cross validation) was defined as the proportion of correctly classified trials in the cross-validation procedure. To get an estimate of the accuracy variability, we performed a bootstrap 200 times on randomly selecting trials for pseudo-population trial construction. To evaluate the statistical significance of decoding accuracy, pseudo-population surrogate trial construction was obtained by randomly shuffling the label of the classes over 1000 bootstrap runs. The shuffled labels decoding was used to get a null distribution of the decoding accuracies that would occur by chance. For each time bin, we performed a Permutation test by computing P-values as the proportion of shuffle repetitions that exceed the real mean decoding accuracies [15], and significant mean decoding accuracies were defined as those time bins where $P < 0.05$. Since decoding accuracy levels depend on the number of included units, in order to compare the stimulus identity and decoding accuracy across days/conditions, we randomly sampled the same number of units (250 units) for each day/condition and bootstrap run. For spontaneous freezing decoding analyses we selected spontaneous freezing and non-freezing episodes longer than 4 s taking place outside of the CS. For freezing versus non-freezing behaviour decoding at CS^+ onset, we considered CS^+ AV and CS^+ NAV trials of at least 2.5 s, for which first freezing intervals started within 0.5 s after CS^+ onset (CS^+ FzOns) and trials for which freezing started at least 0.75 s after CS onset (CS^+ NFzOns), ($n = 13$ out of 68 sessions, total 269 units). During the time window from 0.45 s till 0.75 s, mice exhibited freezing for CS^+ FzOns trials and non-freezing for CS^+ NFzOns trials with probability 1. Cross-decoding CS^+ versus baseline in Active Avoidance and Confined tasks, was performed by doing the described decoding procedure at CS onset aligned condition to train the SVM models in one task, and testing the decoding performance in the other one. Principal component analysis (PCA) was performed on trial-averaged data by randomly selecting 50% of trials for each recording session and neuron, and repeated 100 times. All the analyses were performed using built-in implementation together with in-house codes in Matlab (The MathWorks, Inc., Natick, MA, USA).

Data availability

Data supporting the findings of this study are available from the corresponding author on request.

Code availability

Custom-written codes used to analyse data from this study are available from the corresponding author on request.

Declaration of interests

The authors do not work for, advise, own shares in, or receive funds from any organization that could benefit from this article, and have declared no affiliations other than their research organizations.

Acknowledgements

We thank N. Winke and other members of the Herry laboratory for helpful discussions. In addition, we thank S. Laumond, J. Tessaire and the technical staff of the housing and experimental animal facility of the Neurocentre Magendie. Microscopy was performed in the Bordeaux Imaging Center of the CNRS-INSERM and Bordeaux University, a member of France BioImaging. This work was supported by grants from the French National Research Agency (ANR-DOPAFEAR, ANR-PRELONGIN, ANR-10-EQPX-08 OPTOPATH), the Conseil Regional d'Aquitaine, and the Fondation pour la Recherche Médicale (FRM-Equipes FRM 2017).

References

[1] R. A. Rescorla, "Behavioral studies of Pavlovian conditioning", *Annu. Rev. Neurosci.* **11** (1988), p. 329-352.

[2] T. C. M. Bienvenu, C. Dejean, D. Jercog, B. Aouizerate, M. Lemoine, C. Herry, "The advent of fear conditioning as an animal model of post-traumatic stress disorder: Learning from the past to shape the future of PTSD research", *Neuron* **109** (2021), p. 2380-2397.

[3] C. Grillon, "Associative learning deficits increase symptoms of anxiety in humans", *Biol. Psychiatry* **51** (2002), p. 851-858.

[4] C. Herry, D. Jercog, "Decoding defensive systems", *Curr. Opin. Neurobiol.* **76** (2022), article no. 102600.

[5] R. R. Rozeske, C. Herry, "Neuronal coding mechanisms mediating fear behavior", *Curr. Opin. Neurobiol.* **52** (2018), p. 60-64.

[6] A. Burgos-Robles, I. Vidal-Gonzalez, G. J. Quirk, "Sustained conditioned responses in prelimbic prefrontal neurons are correlated with fear expression and extinction failure", *J. Neurosci.* **29** (2009), p. 8474-8482.

[7] M. R. Milad, G. J. Quirk, "Neurons in medial prefrontal cortex signal memory for fear extinction", *Nature* **420** (2002), p. 70-74.

[8] F. Sotres-Bayon, D. Sierra-Mercado, E. Pardilla-Delgado, G. J. Quirk, "Gating of fear in prelimbic cortex by hippocampal and amygdala inputs", *Neuron* **76** (2012), p. 804-812.

[9] U. Livneh, R. Paz, "Amygdala-prefrontal synchronization underlies resistance to extinction of aversive memories", *Neuron* **75** (2012), p. 133-142.

[10] T. Seidenbecher, T. R. Laxmi, O. Stork, H. C. Pape, "Amygdalar and hippocampal theta rhythm synchronization during fear memory retrieval", *Science* **301** (2003), p. 846-850.

[11] E. Likhtik, J. M. Stujenske, M. A. Topiwala, A. Z. Harris, J. A. Gordon, "Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety", *Nat. Neurosci.* **17** (2014), p. 106-113.

[12] N. Karalis, C. Dejean, F. Chaudun, S. Khoder, R. R. Rozeske, H. Wurtz, S. Bagur, K. Benchenane, A. Sirota, J. Courtin, C. Herry, "4-Hz oscillations synchronize prefrontal-amygdala

circuits during fear behavior", *Nat. Neurosci.* **19** (2016), p. 605-612.

[13] S. Saxena, J. P. Cunningham, "Towards the neural population doctrine", *Curr. Opin Neurobiol.* **55** (2019), p. 103-111.

[14] R. Quian Quiroga, S. Panzeri, "Extracting information from neuronal populations: information theory and decoding approaches", *Nat. Rev. Neurosci.* **10** (2009), p. 173-185.

[15] D. Jercog, N. Winke, K. Sung, M. M. Fernandez, C. Francioni, D. Rajot, J. Courtin, F. Chaudun, P. E. Jercog, S. Valerio, C. Herry, "Dynamical prefrontal population coding during defensive behaviours", *Nature* **595** (2021), p. 690-694.

[16] E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, "Dynamic population coding of category information in inferior temporal and prefrontal cortex", *J. Neurophysiol.* **100** (2008), p. 1407-1419.

[17] J. S. Choi, C. K. Cain, J. E. LeDoux, "The role of amygdala nuclei in the expression of auditory signaled two-way active avoidance in rats", *Learn. Mem.* **17** (2010), p. 139-147.

[18] C. Bravo-Rivera, C. Roman-Ortiz, E. Brignoni-Perez, F. Sotres-Bayon, G. J. Quirk, "Neural structures mediating expression and extinction of platform-mediated avoidance", *J. Neurosci.* **34** (2014), p. 9736-9742.

[19] G. J. Quirk, C. Repa, J. E. LeDoux, "Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat", *Neuron* **15** (1995), p. 1029-1039.

[20] S. Duvarci, D. Pare, "Amygdala microcircuits controlling learned fear", *Neuron* **82** (2014), p. 966-980.

[21] H. C. Pape, D. Pare, "Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear", *Physiol. Rev.* **90** (2010), p. 419-463.

[22] K. A. Goosens, J. A. Hobin, S. Maren, "Auditory-evoked spike firing in the lateral amygdala and Pavlovian fear conditioning: mnemonic code or fear bias?", *Neuron* **40** (2003), p. 1013-1022.

[23] O. Klavir, M. Prigge, A. Sarel, R. Paz, O. Yizhar, "Manipulating fear associations via optogenetic modulation of amygdala inputs to prefrontal cortex", *Nat. Neurosci.* **20** (2017), p. 836-844.

[24] A. Burgos-Robles, E. Y. Kimchi, E. M. Izadmehr, M. J. Porzenheim, W. A. Ramos-Guasp, E. H. Nieh, A. C. Felix-Ortiz, P. Namburi, C. A. Leppla, K. N. Presbrey, K. K. Anandalingam, P. A. Pagan-Rivera, M. Anahtar, A. Beyeler, K. M. Tye, "Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment", *Nat. Neurosci.* **20** (2017), p. 824-835.

[25] V. Senn, S. B. Wolff, C. Herry, F. Grenier, I. Ehrlich, J. Gründemann, J. P. Fadok, C. Müller, J. J. Letzkus, A. Lüthi, "Long-range connectivity defines behavioral specificity of amygdala neurons", *Neuron* **81** (2014), p. 428-437.

[26] A. Poremba, M. Gabriel, "Amygdalar lesions block discriminative avoidance learning and cingulothalamic training-induced neuronal plasticity in rabbits", *J. Neurosci.* **17** (1997), p. 5237-5244.

[27] J. Courtin, F. Chaudun, R. R. Rozeske, N. Karalis, C. Gonzalez-Campo, H. Wurtz, A. Abdi, J. Baufreton, T. C. Bienvenu, C. Herry, "Prefrontal parvalbumin interneurons shape neuronal activity to drive fear expression", *Nature* **505** (2014), p. 92-96.

[28] D. Sierra-Mercado, N. Padilla-Coreano, G. J. Quirk, "Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear", *Neuropsychopharmacology* **36** (2011), p. 529-538.

[29] F. Ramirez, J. M. Moscarello, J. E. LeDoux, R. M. Sears, "Active avoidance requires a serial basal amygdala to nucleus accumbens shell circuit", *J. Neurosci.* **35** (2015), p. 3470-3477.

[30] J. E. LeDoux, J. Moscarello, R. Sears, V. Campese, "The birth, death and resurrection of avoidance: a reconceptualization of a troubled paradigm", *Mol. Psychiatry* **22** (2017), p. 24-36.

[31] E. Spaak, K. Watanabe, S. Funahashi, M. G. Stokes, "Stable and dynamic coding for working memory in primate prefrontal cortex", *J. Neurosci.* **37** (2017), no. 27, p. 6503-6516.

[32] D. R. Euston, A. J. Gruber, B. L. McNaughton, "The role of medial prefrontal cortex in memory and decision making", *Neuron* **76** (2012), no. 6, p. 1057-1070.

[33] N. Winke, C. Herry, D. Jercog, "The geometry of appetitive-aversive value representations in medial prefrontal networks", bioRxiv 2023.03.03.530871, https://doi.org/10.1101/2023.03.03.530871.

[34] V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex", *Nature* **503** (2013), p. 78-84.