Reviewed Preprint
v1 • October 1, 2025
Not revised

Reviewed Preprint
v2 • April 17, 2026
Revised by authors

Neural Representation of Associative Threat Learning in Pulvinar Divisions, Lateral Geniculate Nucleus, and Mediodorsal Thalamus in Humans

Muhammad Badarnee[1], Zhenfu Wen[1], B Isabel Moallem[1], Stephen Maren[2,3], Mohammed R Milad[1]

[1]Department of Psychiatry and Behavioral Sciences, The University of Texas, Health Science Center at Houston, McGovern Medical School, Houston, United States • [2]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, United States • [3]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, United States

For correspondence: mohammed.r.milad@uth.tmc.edu

Competing interests: No competing interests declared

Funding: See page 23

Reviewing editor: Mathieu Wolff, CNRS, University of Bordeaux, France

eLife Assessment

This study provides **valuable** insights into the role of thalamic nuclei in associative threat and extinction learning, underpinned by a large dataset and rigorous, multipronged analyses. The evidence provided is **solid**, supporting the main conclusions. Minor analytical refinements notwithstanding, the manuscript will be of broad interest to researchers in learning and memory, fear, thalamic circuitry, and related mental health conditions.

https://doi.org/10.7554/eLife.108043.2.sa4

Abstract

Understanding the neural mechanisms underlying associative threat learning is essential for advancing behavioral models of threat and adaptation. We investigated distinct activation patterns across thalamic pulvinar divisions, lateral geniculate nucleus (LGN), and mediodorsal thalamus (MD) during the acquisition of associative threat learning in the MRI. The anterior pulvinar and MD exhibited parallel activation patterns, which we interpret as relating to automatic and more deliberative learning processes. Additionally, our findings suggest a hierarchical pulvinar organization during fear conditioning, in which coordinated activation among inferior, lateral, medial, and anterior divisions may support the integration of threat-related information. Pulvinar divisions and the MD showed activation during extinction learning and exhibited patterns consistent with salience processing and safety–threat memory expression during extinction recall and threat renewal. LGN activation patterns during threat learning were consistent with feedforward processing of visual information. This study extends dominant brain models of threat learning and memory, reframing our understanding of distinct thalamic roles in these psychological processes.

Introduction

Over a century ago, Ivan Pavlov provided foundational behavioral evidence for associative learning, demonstrating that pairing a neutral stimulus with an unconditioned stimulus could elicit a conditioned response (Pavlov, 1904). This discovery laid the groundwork for modern learning theories and has been particularly influential in understanding how humans learn to associate neutral stimuli with threats, a process known as associative threat learning. This survival-oriented learning plays a key role in shaping a broad range of behavioral and emotional responses, such as avoidance, decision-making under threats, and fear regulation (Badarnee et al., 2025; Kolling et al., 2014; Korn and Bach, 2019, 2018; Milad et al., 2013). Research on

the neural mechanisms underlying associative threat learning has primarily focused on brain regions involved in emotional regulation, such as the amygdala, hippocampus, and prefrontal cortex (PFC) (Fullana et al., 2015; Krasne et al., 2021; Milad and Quirk, 2012; Vuilleumier et al., 2003). More recently, the thalamus has received growing attention in threat learning (Lithari et al., 2015; Penzo et al., 2015; Ramanathan et al., 2018; Ramanathan and Maren, 2019; Ratigan et al., 2023; Totty et al., 2023), since its role has been reconsidered beyond the traditional view of a mere relay station. Yet, the distinct contribution of this complex structure to associative threat learning remains poorly understood.

The thalamus is a hub structure in the mammalian brain that plays multiple critical roles, ranging from basic sensory processing to higher-order functions and threat learning (Halassa and Sherman, 2019; Hummos et al., 2022; Hwang et al., 2017; Saalmann and Kastner, 2011; Sherman, 2016, 2007). It has long been proposed that threat detection is mediated by the pulvinar and the lateral geniculate nucleus (LGN) (Carr, 2015; Pessoa and Adolphs, 2010; Silverstein and Ingvar, 2015), two major nuclei of the visual thalamus (Arcaro et al., 2015; Casanova and Chalupa, 2023; Takakuwa et al., 2021). The pulvinar projects directly to the amygdala and is believed to subconsciously facilitate rapid survival reactions to threats via a subcortical path ('low road') (Carr, 2015; Pessoa and Adolphs, 2010; Rafal et al., 2015; Wei et al., 2015). The LGN, on the other hand, is part of a slower but more accurate neural pathway, connecting peripheral information to cortical regions for comprehensive processing ('high road'). Additionally, the involvement of the mediodorsal thalamus (MD) in threat learning has also been proposed(Lee and Shin, 2016; Lee et al., 2011). Although this nucleus is not typically classified as part of the visual thalamus, it serves as a high-order region reciprocally connected to the amygdala and PFC while playing a critical role in executive functions (Hwang et al., 2020; Li et al., 2022; Mukherjee et al., 2021; Wolff and Halassa, 2024).

The general implication of these nuclei in threat processing has been demonstrated in primates and rodent models. It has been shown that exposing primates to snake images elicited increased neuronal firing within the pulvinar (Le et al., 2013). Excitation of neurons within the LGN enhanced the acquisition of eyeblink conditioning in rodents (Halverson and Freeman, 2010). Inhibiting this nucleus was associated with impaired conditioned responses (Shi and Davis, 2001; Steinmetz et al., 2013), and activating GABA neurons in the LGN reduced freezing response to an overhead dark shadow that mimics a real-world predator in rats (Salay and Huberman, 2021). Reduced freezing was also observed after MD lesions (Li et al., 2004) and when the connections with the anterior cingulate cortex were ablated (Zheng et al., 2020). Lesions within the MD are also associated with impaired fear extinction (Lee et al., 2011), and injecting gabazine, a modulator of extra-synaptic GABA receptors, into the MD facilitated extinction (Paydar et al., 2014).

Evidence regarding the involvement of the thalamus in the context of threat learning in humans is relatively sparse. Much of our knowledge comes from studies focusing on attention, perception, and decision-making. For example, the LGN has been mainly viewed as a transmitter of peripheral information to other brain regions and found to be associated with selective attention and anticipation of visual stimuli (Mahoney and Schmidt, 2024; O'Connor et al., 2002; Saalmann and Kastner, 2009). The MD contribution to perception, spatial attention, and decision-making has also been reported (Griffiths et al., 2022; Wurtz et al., 2011). Beyond cognitive research, some human studies have specifically investigated the pulvinar's engagement in threat detection. Individuals with increased fiber density in the pulvinar-amygdala pathway showed an enhanced ability to recognize fearful faces (McFadyen et al., 2019). A patient with a complete lesion in the left pulvinar showed slower responses to threatening images when stimuli were presented on the ipsilesional, but not contralesional, field (Ward et al., 2005). In agreement with this, unseen stimuli (fearful vs. happy faces) presented on the blind side of patients with hemianopia moderated their performance when the pulvinar was spared but not when it was lesioned (Bertini et al., 2018).

Specific thalamic contribution to threat processing is still largely unknown, and translational research on this topic is limited. We aim to address these critical gaps in our understanding by investigating the distinct neural representations of associative threat learning in the human pulvinar divisions, LGN, and MD. We analyzed the neuroimaging data of 293-412 controls. All participants underwent a two-day threat learning paradigm (Milad et al., 2009, 2007; Wen et al., 2024, 2022) while in the fMRI scanner. The conditioned stimulus (CS+) (e.g., red light) was associated with an electric shock (unconditioned stimulus US, 62.5% reinforcement rate), while a control light (e.g., blue) was never paired with the US (CS-). This conditioning phase occurred in a computer-displayed visual context A (e.g., an office). The extinction learning included presenting the CS+ and CS- with no US reinforcement in a distinct context B (e.g., bookcase). In extinction recall and threat renewal, the extinguished stimuli were presented with no US reinforcement within contexts B (safe contextual cues) and A (threat contextual cues), respectively. Schematic illustrations of each phase of the paradigm are presented in Figs. 1a, 4a-6a.

We focus on neural activation patterns within pulvinar divisions, LGN, and MD while acquiring the CS-US association. As associative learning is a rapid psychological process (Konrad et al., 2024), we analyzed the first four trials of all threat learning phases. The purpose of this study is to investigate thalamic function during each learning phase separately, focusing on CS+ vs. CS− differences within phases rather than comparing activation across phases. This phase-specific approach allows us to characterize thalamic functional dynamics within each stage of learning and memory, avoiding potential confounds arising from the distinct processes of conditioning, extinction, and recall. We compared brain activation (BOLD) to the CS+ vs. CS- at the block level by averaging the activation across all four trials and at a trial-by-trial level to provide finer activation temporal resolution. This dual approach allows us to capture the general neural responses associated with associative threat learning and provides new insights into trial-level dynamics (Wen et al., 2022). The current dominant neurocognitive thalamic models (Sherman, 2007; Sherman and Guillery, 2006) highlight the thalamus's role in mediating cortical-cortical communications and facilitating high-order functions. Based on this view, we anticipate distinct pulvinar, MD, and LGN roles.

The LGN, as a first-order nucleus, is expected to serve relay functions, transmitting information with minimal integrations. The pulvinar and MD, on the other hand, as higher-order nuclei, are likely involved in more complex processing, potentially integrating threat-related information.

Results

Parallel functional representation consistent with associative threat learning in the anterior pulvinar and MD during conditioning

At the block level, both the anterior pulvinar and MD showed increased activation to CS+ vs. CS− (anterior pulvinar: $t_{(292)} = 4.41$, $p = 0.00001$, $d = 0.25$; MD: $t_{(292)} = 6.41$, $p = 5.83x10^{-10}$, $d = 0.37$; Fig. 1b–c), suggesting a possible involvement of these regions in early associative threat learning. The trial-wise analysis revealed a distinct activation pattern. On the first trial, activation did not differ between CS+ and CS− in either the anterior pulvinar ($p_{FDR} = 0.44$) or the MD ($p_{FDR} = 0.20$). In contrast, on the subsequent trial, both regions showed significantly greater BOLD responses to CS+ than CS− (anterior pulvinar: $p_{FDR} = 0.000001$; MD: $p_{FDR} = 0.000003$; Fig. 1b–c; Detailed statistical parameters are provided in Supplementary Tables 1–2). In our paradigm, the electric shock was paired with CS+ at the end of the CS presentation. The similarity in BOLD responses to CS+ and CS- during the first trial likely reflects an initial equivalence in the emotional valence of the stimuli. The gradual increase in activation for CS+ by the second trial suggests a shift in the emotional valence, consistent with rapid associative learning.

The similarity in trial-level activation patterns in the anterior pulvinar and MD raises the question of whether this apparent similarity truly associated with a parallel functional contribution to associative threat learning. To test this possibility, we first quantified the trial-wise relationships

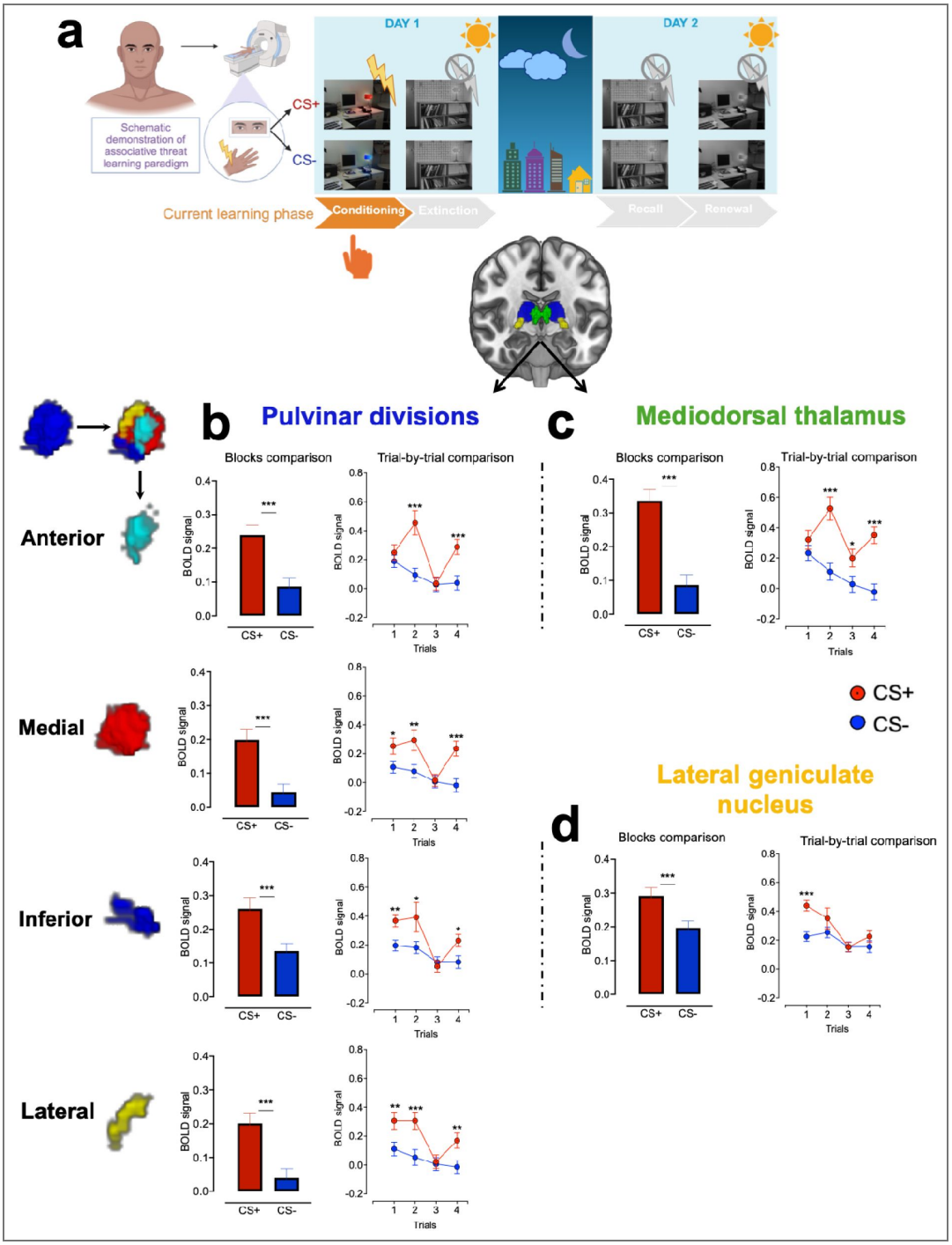


Fig. 1. Neural representation during associative threat learning in pulvinar divisions, MD, and LGN.

(a) Human fear conditioning paradigm during fMRI. (b–d) Means ± SE of activation in response to CS+ vs. CS− at both block-wise and trial-wise levels within pulvinar divisions (b), MD (c), and LGN (d). Block-level comparisons were assessed using paired *t*-tests, while trial-level effects were examined using a 2 × 2 repeated-measures ANOVA, followed by post hoc comparisons between CS+ and CS− across the four trials. Multiple comparisons were controlled using false discovery rate (FDR) correction. Conditioning sample size: $n = 293$. Detailed statistical parameters are provided in Supplementary Tables 1–2. Additionally, cross-validation analyses were conducted to assess the robustness of the CS+ vs. CS− signal in the anterior pulvinar and MD, consistent with our interpretation of learning-related effects (Supplementary Figs. 1–2 and Supplementary File 1). *p < 0.05, **p < 0.01, ***p < 0.001. SE, standard error; MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; CS+, conditioned stimulus predicting shock; CS−, conditioned stimulus predicting no shock. Display items in panel (a) were created using *BioRender*.

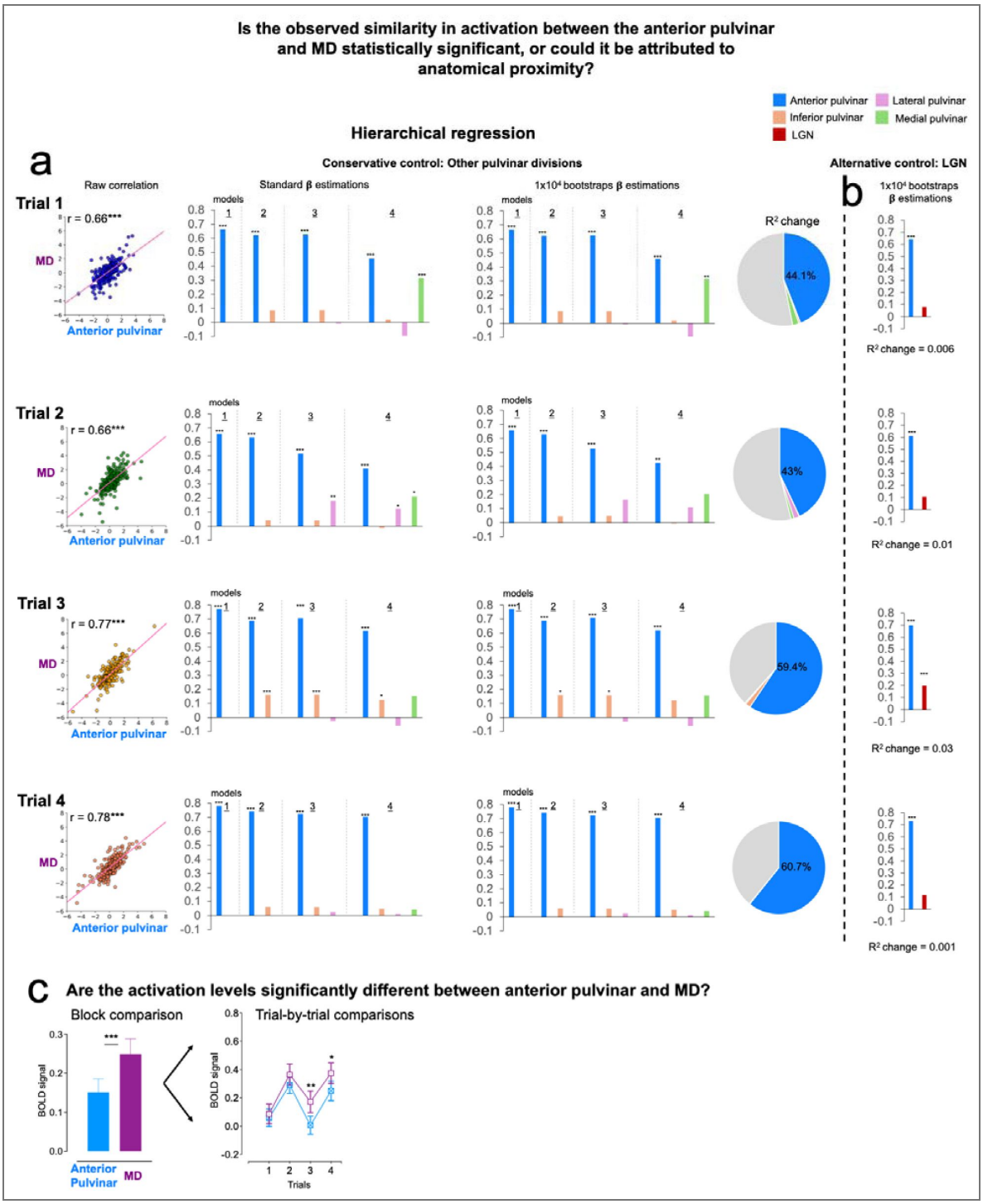


Fig 2. Quantifying the relationships between the anterior pulvinar and MD during conditioning.

a. Hierarchical regression models of trial-wise relationships between the anterior pulvinar and MD activations while controlling for a potential effect of anatomical proximity. The effect of anatomical proximity was controlled by progressively adding other pulvinar divisions as controls. **b.** Results of a hierarchical model that included the LGN as an alternative control, distinct from the pulvinar. **c.** Comparison of activation levels in the anterior pulvinar and MD at both block-wise and trial-wise levels (means ± SE). Sample size in these analyses: *n* = 293. Supplementary analyses testing alternative hypotheses regarding the correlations between the MD and all pulvinar divisions are presented in Supplementary Fig. 3. *p<0.05, **p<0.01, ***p<0.001 MD: Mediodorsal thalamus. LGN: Lateral geniculate nucleus. SE: Standard error.

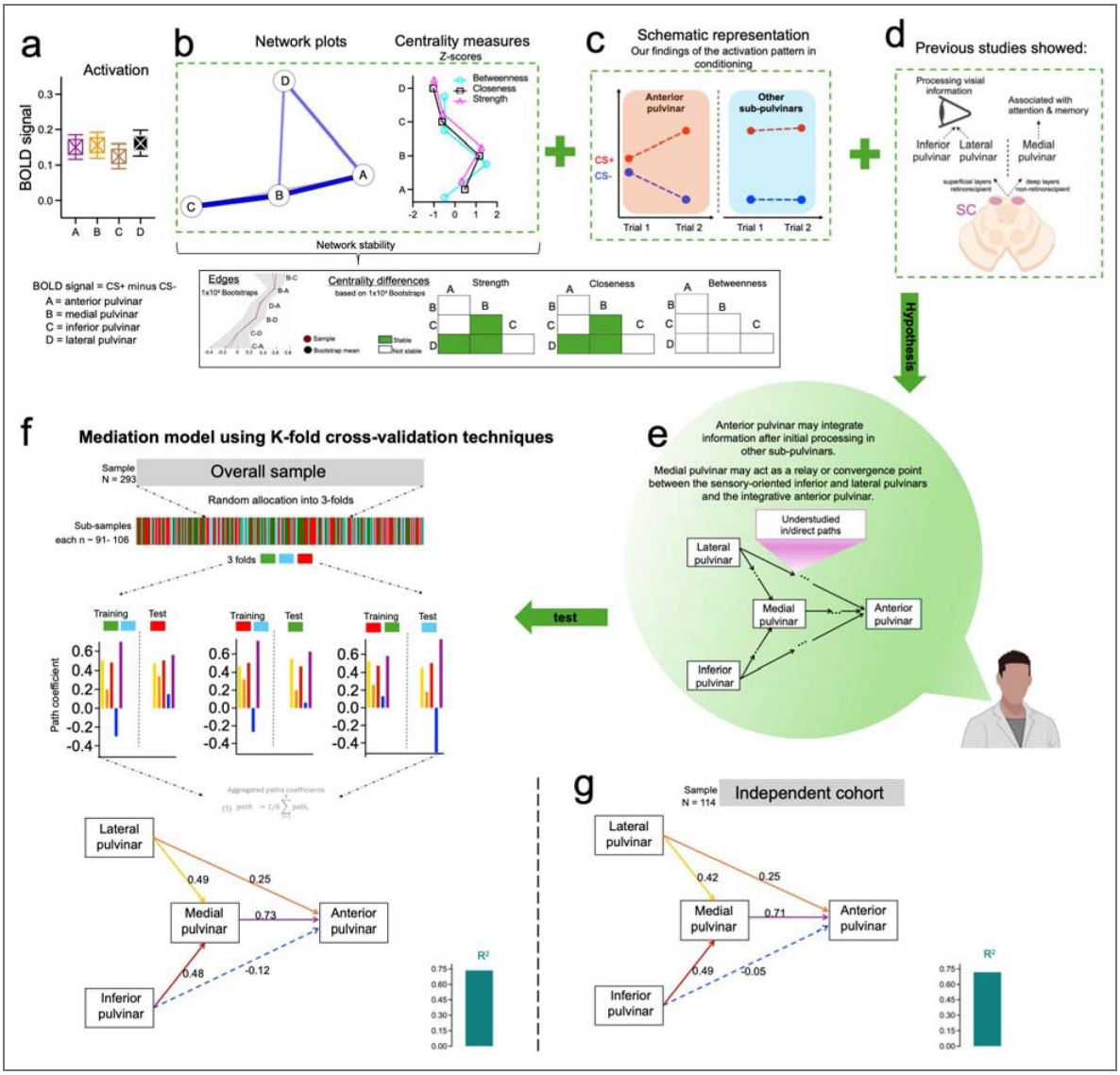


Fig 3. A data-driven approach to understanding the functional relationships between pulvinar divisions during conditioning.

a. Means ± SE of activation differences between pulvinar divisions. **b.** Network analysis reveals that the medial pulvinar serves as a central hub, mediating interactions among pulvinar divisions and exhibiting increased centrality measures. **c.** Schematic visualization of activation patterns we observed in pulvinar divisions. **d.** Previous studies suggest that the inferior and lateral pulvinar are involved in processing basic visual information (Berman and Wurtz, 2010, 2008; Cortes et al., 2024), while the medial pulvinar is associated with higher-level functions, including working memory (Homman-Ludiye and Bourne, 2019). **e.** Based on **b-d**, we hypothesize that the medial pulvinar mediates the relationships with other divisions. **f.** Mediation analysis supports our hypothesis (panel **e**). **g.** Validation of the mediation model on an additional independent sample. Dashed paths in panels **f** and **g** represent statistically unstable paths, while the continuous paths indicate stable paths. *p<0.05, **p<0.01, ***p<0.001 SE: Standard error. SC: Superior colliculus. Display items in panels **d** and **e** were created using *BioRender*. **Note:** The functional relationships among pulvinar divisions during threat learning should be interpreted as computational dependencies derived from statistical associations. These effects may reflect indirect interactions mediated by corticothalamic and thalamocortical pathways (e.g., via visual cortex), rather than direct inter-nuclear connectivity. Elucidating the underlying anatomical mechanisms will require future studies.

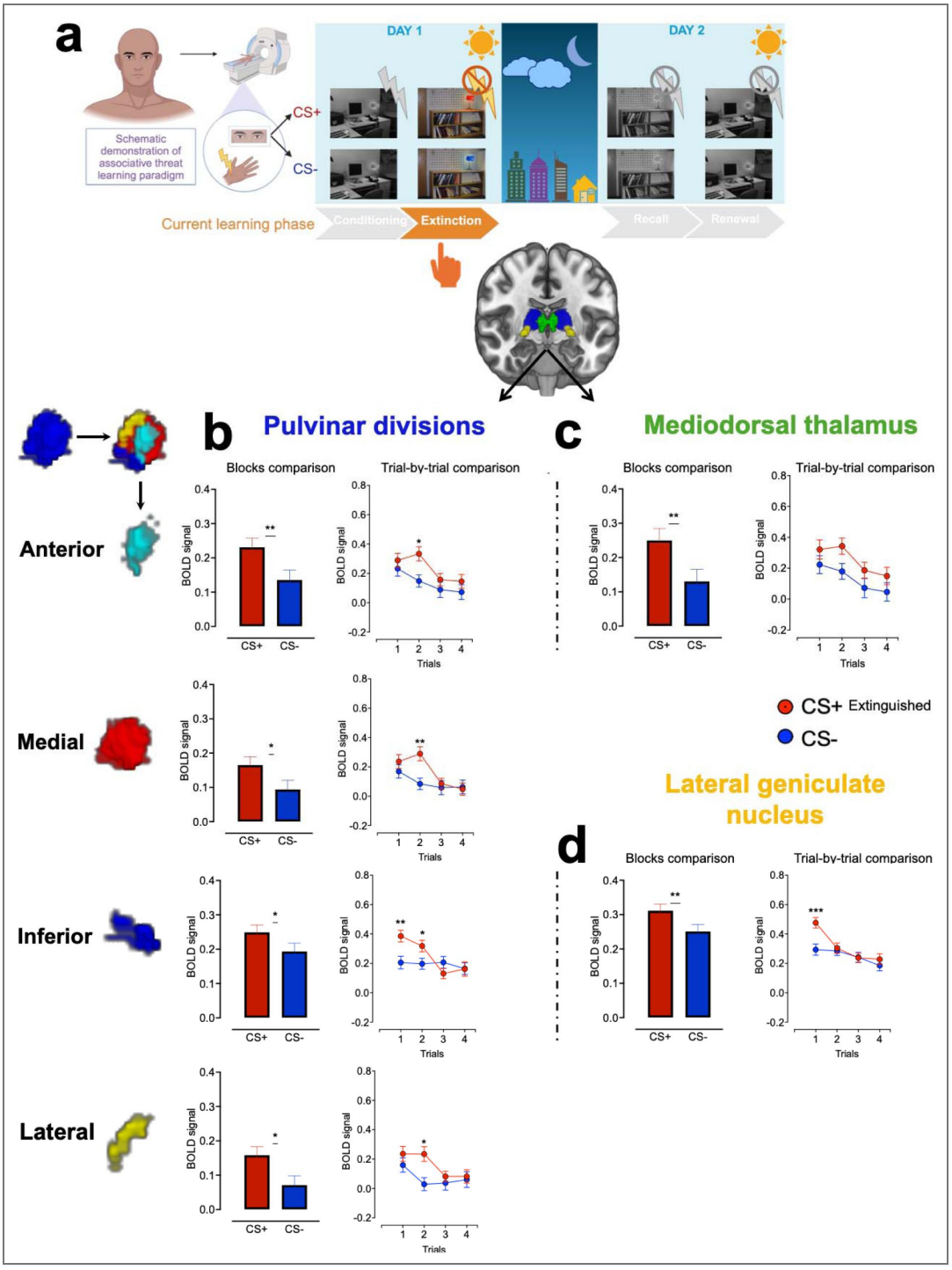


Fig. 4. Neural representation during extinction learning in pulvinar divisions, MD, and LGN.

(a) Human extinction learning paradigm during fMRI. (b–d) Means ± SE of activation in response to CS+ vs. CS− at both block-wise and trial-wise levels within pulvinar divisions (b), MD (c), and LGN (d). Block-level comparisons were assessed using paired *t*-tests, while trial-level effects were examined using a 2 × 2 repeated-measures ANOVA, followed by post hoc comparisons between CS+ and CS− across the four trials. Multiple comparisons were controlled using false discovery rate (FDR) correction. Extinction sample size: $n = 320$. Detailed statistical parameters are provided in Supplementary Tables 4–5. *p < 0.05, **p < 0.01, ***p < 0.001. SE, standard error; MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; Extinguished CS+, conditioned stimulus that no longer predicts shock; CS−, conditioned stimulus predicting no shock. Display items in panel (a) were created using *BioRender*.

using Pearson correlation coefficients between the two regions. The observed correlations support consistent co-activation across all corresponding trials (r *values* ≥ 0.66, $p < 0.001$; Fig. 2a, left). However, this approach could not exclude the possibility that the shared variance is a consequence of shared anatomical proximity, particularly because both regions are part of the same brain structure, i.e., the thalamus. To control for this confound, we conducted a hierarchical regression model in four steps for each trial. We modeled MD activation as the target and progressively added the pulvinar divisions as covariates. Despite minor variations in regression coefficients across control models, the anterior pulvinar consistently showed the largest standardized regression coefficients (all $p < 0.01$; Fig. 2a, middle) and accounted for a substantial proportion of the explained variance in MD activation ($R^2 = 0.43–0.61$). In contrast, other pulvinar divisions contributed minimally to the explained variance (*across all other pulvinar divisions, the maximum observed* ΔR^2 *was 0.02* in trial 1, *0.01* in trial 2, *0.01* in trial 3, and *0.002* in trial 4; Fig. 2a). These results were stable across 10,000 bootstrap resamples with replacement, indicating that the association between anterior pulvinar activation and MD activation is robust to resampling variability (Fig. 2a, right; additionally, see alternative hypothesis tests of the anterior pulvinar–MD relationships in Supplementary Fig. 3).

Applying the same analytical approach while additionally controlling for LGN activation as an anatomical control, beyond the pulvinar itself, yielded comparable results. Anterior pulvinar activation continued to account for the majority of the explained variance in MD activation, whereas the inclusion of LGN activation produced only an incremental additional increase in explained variance ($\Delta R^2 = 0.006$ in trial 1, *0.01* in trial 2, *0.03* in trial 3, and *0.01* in trial 4; Fig. 2b). Together, these analyses support preferential trial-wise functional co-activations between the anterior pulvinar and MD, independent of shared anatomical proximity.

Although co-activation does not necessarily imply similar activation magnitude, we further tested the differences in activation levels in the anterior pulvinar and MD. We used a t-test to capture the differences in the overall activation at the block level and repeated measures analysis of variance (RM-ANOVA) to capture the activation differences at the trial level. The results showed that the overall MD activation in response to CS+ was higher than the anterior pulvinar response ($t_{(292)} = 3.49$, $p = 0.0005$, $d = 0.20$, Fig. 2c). –The trial-wise analysis revealed that the overall differences between the two regions is driven particularly from the activation in trials 3 and 4 ($F_{(1, 292)} = 5.84$, $p = 0.01$; *Trial 1:* $t_{(292)} = 0.49$, $p_{FDR} = 0.82$, $d = 0.02$; *Trial 2:* $t_{(292)} = 0.03$, $p_{FDR} = 0.97$, $d = 0.002$, *Trial 3:* $t_{(292)} = 3.40$, $p_{FDR} = 0.003$, $d = 0.199$; *Trial 4:* $t_{(292)} = 2.63$, $p_{FDR} = 0.01$, $d = 0.15$; Fig. 2c). The increased activation within the MD might be associated with different levels of threat processing in the two regions.

A data-driven approach reveals hierarchical functional processing of threat in pulvinar divisions

The anatomical and functional interconnections between pulvinar divisions remain insufficiently characterized. To address this gap and understand the pulvinar's role in early associative threat learning, we integrated findings from complementary analyses into a novel computational model designed to investigate the functional relationships between pulvinar divisions.

We first defined the activation within each pulvinar division as the differences in BOLD signal at the block level in response to CS+ compared to CS-. We then performed RM-ANOVA to test whether the pulvinar divisions were engaged with different activation levels. We found no activation differences, suggesting similar processing levels of CS+ information across all pulvinar divisions ($F_{(2.29, 670.33)} = 0.96$, $p = 0.39$; Fig. 3a). Using network analysis, we explored the underlying functional dynamics between the divisions. We employed the EBICglasso method to estimate a sparse Gaussian graphical model (Foygel and Drton, 2010). To assess the stability of the network and provide robust estimates of edges and centrality measures, we conducted 10,000 bootstrap resamples. This approach allowed us to compute 95% confidence intervals (CI), ensuring the reliability of our findings. The combined use of EBICglasso and bootstrapping is essential for accurately capturing the dynamics of pulvinar division interactions. The resulting network included four nodes and five edges. The network plot pointed to stable direct edges connecting the

lateral pulvinar to the anterior pulvinar and indirect edges connecting both the lateral and inferior pulvinar to the anterior through the medial pulvinar (Fig. 3b). Additionally, the medial pulvinar exhibited the highest centrality values, a measure that captures the node importance within a network, indicating a potential hub role of this division in associative threat learning (Fig. 3b).

These underlying dynamics and the activation timing in the trial-wise analyses (overview in Fig. 3c and details in Fig. 1b) led us to propose a possible functional model of the relationships between pulvinar divisions. Specifically, the increased activation induced by CS+ during trial 1 in the medial (p_{FDR} = *0.026*), inferior (p_{FDR} = *0.008*), and lateral pulvinar (p_{FDR} = *0.01*) suggests that these divisions process CS+ information before the anterior pulvinar, which exhibited a delayed BOLD response starting in trial 2 only (p_{FDR} = *0.44* in trial 1, while p_{FDR} = *0.000001* in trial 2). The medial pulvinar mediates connections within the pulvinar network and is associated with elevated centrality values. Together, these findings highlight a possible hub role of the medial pulvinar, integrating CS+ information at a higher level, compared to sensory-driven processing in the inferior and lateral pulvinar. This possible hierarchical organization is supported by previous evidence showing that the inferior and lateral pulvinar are associated with processing basic sensory information (Berman and Wurtz, 2010, 2008; Cortes et al., 2024), while the medial pulvinar is implicated in higher-order processing, including attention and working memory (Homman-Ludiye and Bourne, 2019) (Fig. 3d).

This data-driven approach provides a possible new perspective on the pulvinar divisions' functional specialization during associative threat learning. We hypothesize that, during this process, the activation in the medial pulvinar mediates the functional relationships between the inferior and lateral divisions with the anterior pulvinar (Fig. 3e). To test this hypothesis, we conducted a mediation model analysis and evaluated the model robustness using the k-fold cross-validation method. The sample (*N* = 293) was randomly divided into three groups of 91, 96, and 106 subjects. Each sub-sample was used to test the model while the remaining sub-samples were used to assess the stability of the parameter estimates. This resulted in six iterations of the mediation model. The 95% CI for each path coefficient were estimated using 10,000 bootstrapping replications. This choice was made to achieve greater precision and stability in the CI estimates. Finally, applying the mediation analysis to another independent cohort (*N* = 114) confirmed the model's applicability. The results in Figs. 3f-3g support the mediating role of the medial pulvinar, demonstrating significant indirect computational paths from the lateral and inferior divisions to the anterior pulvinar (*both p-values < 0.05*). Detailed statistical parameters are provided in Supplementary Tables 3.

Pulvinar divisions and MD seem to support extinction learning while preserving threat memory across contextual cues

After our comprehensive examination of the thalamic contribution to threat conditioning, we moved forward to examine the neural contributions of the same regions to extinction learning. During this phase, at the block level, we found higher activation in response to CS+E than CS- across all pulvinar divisions ($t_{(319)}$ = *2.02–2.85, p = 0.005–0.0.044, d = 0.11–0.15*) and MD ($t_{(319)}$ = *3.04, p = 0.003, d = 0.17)*. The trial-wise analysis indicated that these differences were primarily driven by the first two trials in the pulvinar; across pulvinar divisions, all second-trial comparisons survived FDR correction (p_{FDR} < *0.05*, Fig. 4b), whereas the MD showed no significant effects at the trial-wise level (*all* p_{FDR} > *0.05*, Fig. 4c). Thus, these nuclei—particularly the pulvinar—showed possible engagement in rapid extinction learning, whereas cumulative block-level differences in the MD suggest a role in updating higher-order representations of the overall task structure. Detailed statistical parameters are provided in Supplementary Tables 4-5.

During extinction recall, the anterior pulvinar and MD exhibited similar functional patterns at both block and trial levels, with increased activation to CS+E compared to CS- during the first two trials (Anterior pulvinar: at the block level $t_{(411)}$ = *2.49, p = 0.013, d = 0.12,* p_{FDR} = *0.026* in trial 1 and p_{FDR} = *0.026* in trial 2; while p_{FDR}> *0.05* in trial 3 and 4. MD: at the block level $t_{(411)}$ = *4.3, p* =

0.00002, d = 0.21, p_{FDR} *= 0.000002* in trial 1 and p_{FDR} *= 0.008* in trial 2 while p_{FDR} *> 0.05* in trial 3 and 4; Fig. 5b-c). In contrast, the medial and inferior pulvinar maintained the BOLD signal to CS+E in the second trial (both p_{FDR} in trial 2 *< 0.01* while p_{FDR} in other trials *> 0.05*). We observed no activation differences in the lateral pulvinar (*p-values* at the block and trial-wise level *> 0.05*). See Fig. 5b-c, and Detailed statistical parameters in Supplementary Tables 6-7. These findings highlight the involvement of most pulvinar divisions and MD in sustaining threat memory under safe contextual cues and suggests association with retrieval suppression of the extinguished threat.

Changing the contextual cues to the threat background in which the original associative learning occurred was associated with increased activation to CS+E on the first trial in the anterior pulvinar (p_{FDR} *= 0.0001*), lateral pulvinar (p_{FDR} *= 0.008*), and MD (p_{FDR} *= 0.0003*). No significant activation was observed in the medial or inferior pulvinar at the trial level (all p_{FDR} *> 0.05* across trials; Fig. 6b–c; Detailed statistical parameters are provided in Supplementary Tables 8-9.

LGN activation patterns during threat learning are consistent with feedforward processing

At the block level, activation was greater for CS+ than CS− across conditioning ($t_{(292)}$ *= 3.35, p = 0.0009, d = 0.19*), extinction ($t_{(319)}$ *= 2.66, p = 0.008, d = 0.14*), recall ($t_{(411)}$ *= 2.61, p = 0.009, d = 0.12*), and renewal ($t_{(317)}$ *= 2.41, p = 0.016, d = 0.13*). These differences were primarily driven by the first trial (p_{FDR} *< 0.001* in trial 1 across phases, except recall: p_{FDR} *= 0.056*), during which CS+ elicited increased BOLD responses that declined over subsequent trials. See Figs. 1d, 4d–6d; Detailed statistical parameters are provided in Supplementary Tables 2, 5, 7, and 9. Collectively, these findings suggest that the LGN contributes to multiple phases of threat learning, with early trial responses consistent with feedforward visual processing.

Thalamic connectivity underlies threat learning and memory

We tested the relationships between each thalamic nucleus and core brain regions within the 'fear circuit' (Herry et al., 2010; Maren and Quirk, 2004; Milad et al., 2014; Milad and Quirk, 2012; Tovote et al., 2015) by seeding the nuclei to target the amygdala, hippocampus, ventromedial prefrontal cortex (vmPFC), subgenual anterior cingulate cortex (sgACC), and dorsal anterior cingulate cortex (dACC). During conditioning, the anterior pulvinar showed significant CS+ > CS− connectivity with the amygdala ($t_{(292)}$ *= 2.46,* p_{FDR} *= 0.02, d = 0.14*), vmPFC ($t_{(292)}$ *= 2.60,* p_{FDR} *= 0.02, d = 0.15*), and hippocampus ($t_{(292)}$ *= 2.42,* p_{FDR} *= 0.02, d = 0.14*), but not with the dACC ($t_{(292)}$ *= -0.95,* p_{FDR} *= 0.42, d = -0.05*) or sgACC ($t_{(292)}$ *= 0.79,* p_{FDR} *= 0.42, d = 0.04*), Fig. 7a.

Additionally, to assess the robustness of the connectivity differences between CS+ and CS−, we performed nonparametric bootstrapping (10,000 resamples) and estimated the bias-corrected and accelerated (BCa) CI for the mean paired differences. The BCa intervals excluded zero for the amygdala (BCa 95% CI [0.002, 0.015]), vmPFC (BCa 95% CI [0.002, 0.017]), and hippocampus (BCa 95% CI [0.0016, 0.016]), but included zero for the dACC (BCa 95% CI [-0.01, 0.003]) and sgACC (BCa 95% CI [-0.005, 0.011]). The connectivity with the amygdala likely supports encoding the emotional valence of the newly learned CS-US associations, while the engagement of the hippocampus suggests prioritizing the contextual CS+ information. The connectivity with the vmPFC may underlie a process of top-down control release to enhance fear expression or encode threat information for tracking and supporting decision-making in future encounters.

During extinction learning, the LGN showed increased CS+E > CS− connectivity with the sgACC ($t_{(319)}$ *= 2.60,* p_{FDR} *= 0.04, d = 0.14*). No significant differences were observed with other cortical regions, including the vmPFC ($t_{(319)}$ *= 1.13,* p_{FDR} *= 0.32, d = 0.06*) and dACC ($t_{(319)}$ *= 0.41,* p_{FDR} *= 0.68, d = 0.02*), or with limbic regions, including the amygdala ($t_{(319)}$ *= 1.19,* p_{FDR} *= 0.32, d = 0.06*) and hippocampus ($t_{(319)}$ *= 1.34,* p_{FDR} *= 0.32, d = 0.07*), Fig. 7b. Additionally, to assess the robustness of the connectivity differences between CS+E and CS−, we performed nonparametric bootstrapping analysis (10,000 resamples). The BCa interval for the mean differences excluded zero for the LGN–

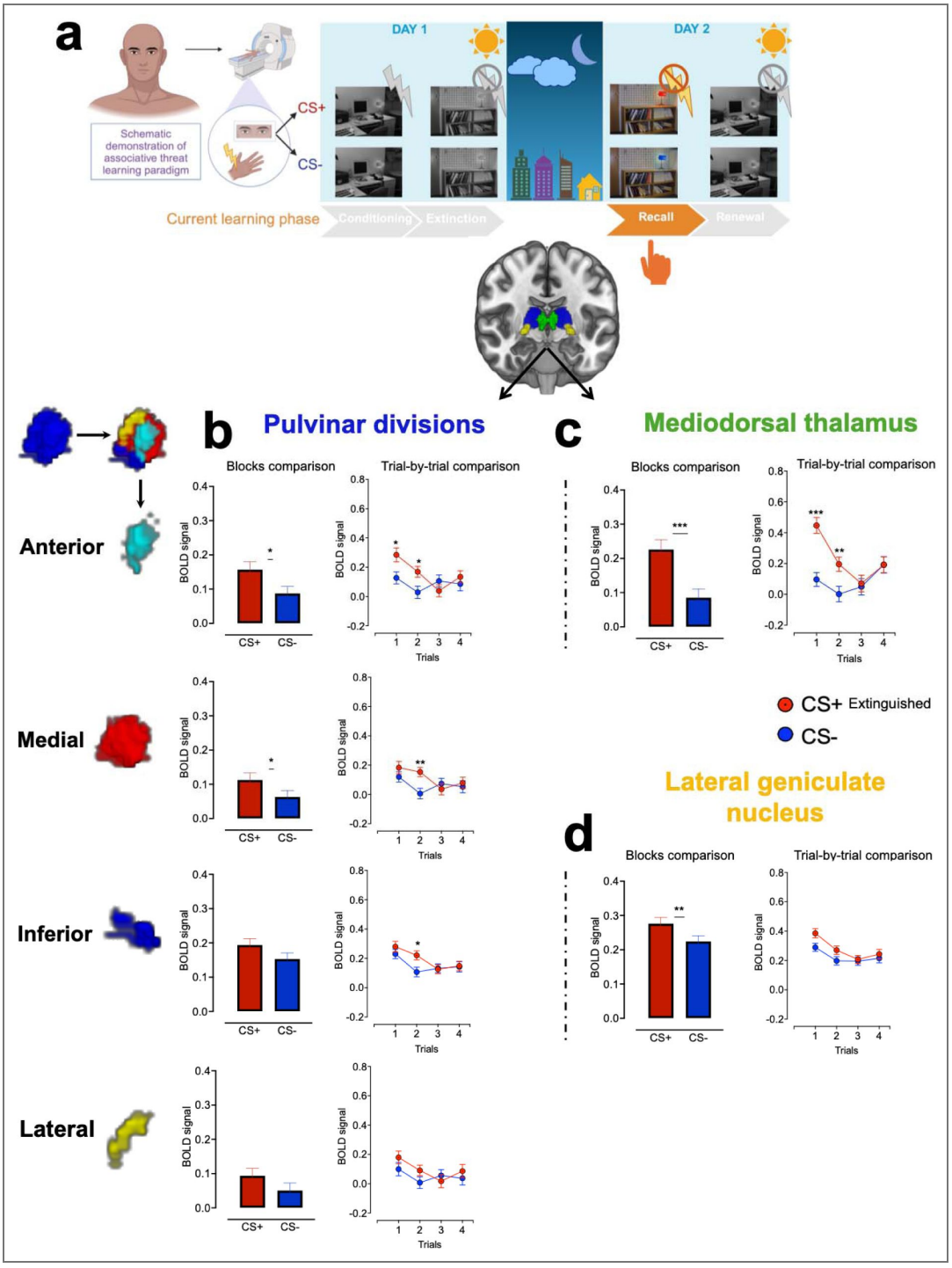


Fig. 5. Neural representation during extinction recall in pulvinar divisions, MD, and LGN.

(a) Human extinction recall paradigm in the fMRI conducted within safe contextual cues. (b–d) Means ± SE of activation in response to CS+ vs. CS− at both block-wise and trial-wise levels within pulvinar divisions (b), MD (c), and LGN (d). Block-level comparisons were assessed using paired *t*-tests, while trial-level effects were examined using a 2 × 2 repeated-measures ANOVA, followed by post hoc comparisons between CS+ and CS− across the four trials. Multiple comparisons were controlled using false discovery rate (FDR) correction. Extinction recall sample size: $n = 412$. Detailed statistical parameters are provided in Supplementary Tables 6–7. *p < 0.05, **p < 0.01, ***p < 0.001. SE, standard error; MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; Extinguished CS+, conditioned stimulus that no longer predicts shock; CS−, conditioned stimulus predicting no shock. Display items in panel (a) were created using *BioRender*.

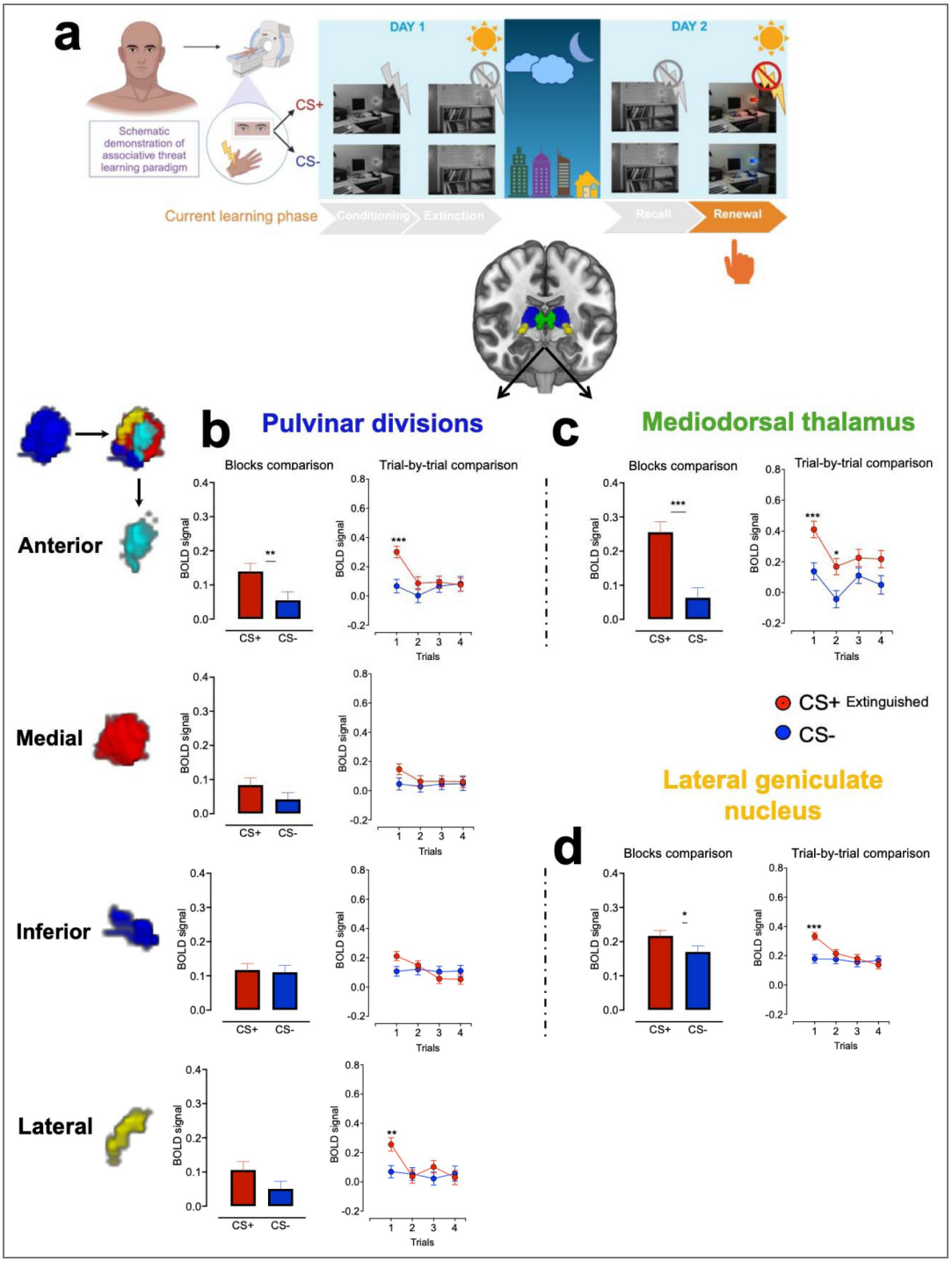


Fig. 6. Neural representation during threat renewal in pulvinar divisions, MD, and LGN.

(a) Human threat renewal paradigm in the fMRI conducted within threat contextual cues in the original context where fear conditioning occurred. (b–d) Means ± SE of activation in response to CS+ vs. CS− at both block-wise and trial-wise levels within pulvinar divisions (b), MD (c), and LGN (d). Block-level comparisons were assessed using paired *t*-tests, while trial-level effects were examined using a 2 × 2 repeated-measures ANOVA, followed by post hoc comparisons between CS+ and CS− across the four trials. Multiple comparisons were controlled using false discovery rate (FDR) correction. Threat renewal sample size: n = 318. Detailed statistical parameters are provided in Supplementary Tables 8–9. *p < 0.05, **p < 0.01, ***p < 0.001. SE, standard error; MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; Extinguished CS+, conditioned stimulus that no longer predicts shock; CS−, conditioned stimulus predicting no shock. Display items in panel (a) were created using *BioRender*.

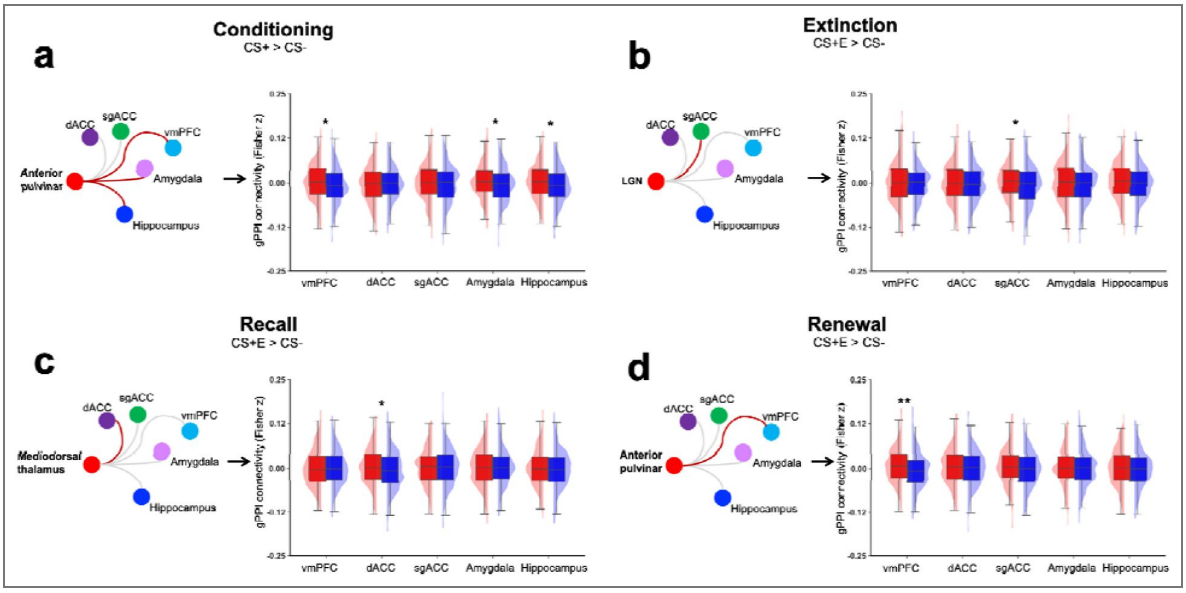


Fig. 7. Task-related thalamic connectivity with threat-related brain regions.

(a–d) Thalamic seed-to-ROI connectivity with the vmPFC, sgACC, dACC, amygdala, and hippocampus during conditioning (*n = 293*), extinction (*n = 320*), recall (*n = 412*), and renewal (*n = 318*). Red lines indicate significant positive connectivity for CS+ vs. CS− following FDR correction across regions ($p_{FDR} < 0.05$); gray lines indicate non-significant effects. For each panel, boxplots and kernel density estimates show the distribution of connectivity values for CS+ and CS−. All thalamic regions of interest—including pulvinar divisions (anterior, inferior, lateral, medial), LGN, and MD—were analyzed using identical statistical procedures. To maintain figure clarity and focus on effects relevant to the main conclusions, we display only seeds showing at least one significant CS+ vs. CS− differences after FDR correction. Non-significant seeds did not show systematic condition-related differences. *p < 0.05, **p < 0.01. MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; vmPFC, ventromedial prefrontal cortex; sgACC, subgenual anterior cingulate cortex; dACC, dorsal anterior cingulate cortex; CS+, conditioned stimulus predicting shock; CS−, conditioned stimulus predicting no shock. gPPI; generalized psychophysiological interaction; connectivity values are Fisher z–transformed correlation coefficients.

sgACC connectivity (BCa 95% CI [0.002, 0.017]), but included zero for all other regions: vmPFC (BCa 95% CI [−0.003, 0.011]), dACC (BCa 95% CI [−0.005, 0.009]), amygdala (BCa 95% CI [−0.002, 0.012]), and hippocampus (BCa 95% CI [−0.002, 0.012]).

Interestingly, in extinction recall and threat renewal, we found that the learners, i.e., the anterior pulvinar and MD seems to support either safe or threatening memory, depending on contextual cues. Specifically, during extinction recall, which occurs in the safe context, we observed increased connectivity to CS+E > CS- between the MD and dACC ($t_{(411)} = 2.60$, $p_{FDR} = 0.04$, $d = 0.12$) but not with other regions: amygdala ($t_{(411)} = 0.038$, $p_{FDR} = 0.97$, $d = 0.002$), hippocampus ($t_{(411)} = 1.2$, $p_{FDR} = 0.57$, $d = 0.059$), sgACC ($t_{(411)} = 0.18$, $p_{FDR} = 0.97$, $d = 0.009$), and vmPFC ($t_{(411)} = -0.49$, $p_{FDR} = 0.97$, $d = -0.02$), Fig. 7c.

These findings were robust to nonparametric bootstrapping (10,000 resamples): the BCa 95% confidence interval for the mean connectivity difference excluded zero for the MD–dACC (BCa 95% CI [0.002, 0.015]), but included zero for the amygdala (BCa 95% CI [−0.006, 0.006]), hippocampus (BCa 95% CI [−0.002, 0.010]), sgACC (BCa 95% CI [−0.006, 0.007]), and vmPFC (BCa 95% CI [−0.008, 0.005]).

During threat renewal, we observed increased CS+E > CS− connectivity between the anterior pulvinar and the vmPFC ($t_{(317)} = 3.47$, $p_{FDR} = 0.002$, $d = 0.19$). In contrast, no significant CS+E > CS− connectivity differences were observed between the anterior pulvinar and other regions, including the amygdala ($t_{(317)} = 0.59$, $p_{FDR} = 0.55$, $d = 0.03$), hippocampus ($t_{(317)} = 1.09$, $p_{FDR} = 0.44$, $d = 0.06$), sgACC ($t_{(317)} = 1.90$, $p_{FDR} = 0.14$, $d = 0.10$), and dACC ($t_{(317)} = 0.92$, $p_{FDR} = 0.44$, $d = 0.05$), Fig. 7d.

Nonparametric bootstrapping (10,000 resamples) supported the robustness of these results. The BCa 95% confidence interval for the mean difference excluded zero only for the anterior pulvinar –vmPFC connectivity (BCa 95% CI [0.005, 0.02]), whereas intervals for the amygdala (BCa 95% CI [-0.005, 0.01]), hippocampus (BCa 95% CI [-0.003, 0.012]), sgACC (BCa 95% CI [-0.0002, 0.015]), and dACC (BCa 95% CI [-0.003, 0.01]) all included zero.

Discussion

We examined the neural representation of associative threat learning within the pulvinar divisions, LGN, and MD, providing new insights into thalamic involvement in this adaptive behavior in humans. We identified distinct thalamic activation profiles during threat learning and memory. The anterior pulvinar and MD exhibited parallel activation patterns consistent with associative learning, raising a possibility that these nuclei are involved in automatic survival responses and deliberate threat processing. We propose a novel hierarchical computational model for processing threat information in the pulvinar divisions. The medial pulvinar seems to mediate basic sensory information from the inferior and lateral divisions to the anterior pulvinar for higher-order integrative learning. Both pulvinar and MD were involved in extinction and showed activation consistent with the salience processing of threat-related memories during extinction recall and threat renewal. The LGN primarily seems to represent feedforward processing, anticipating upcoming visual stimuli throughout threat learning. We integrated these insights into schematic models underlying the emotional and behavioral expression of threat learning and memory, providing a possible neural framework for studying related human behaviors (Fig. 8a-b).

The anterior pulvinar and MD co-activation patterns may demonstrate a parallel contribution to threat learning. The similar activation in response to CS+ and CS- at the first trial, followed by a heightened activation specific to CS+ in the next trial, is likely consistent with the acquisition of the CS-US association. The responses' similarity to both types of CS during the first trial suggests a possible equivalent initial emotional valence. The gradual increase in the activation, specifically to CS+ by the end of trial 1, suggests a shift in the emotional valence of CS+, likely indicating rapid associative learning. This suggested parallel specialized role highlights the anterior pulvinar and MD as central thalamic hubs for integrating CS-US information, apparently specifying the two-system model proposed by LeDoux and Pine (LeDoux and Pine, 2016). Briefly, fear processing,

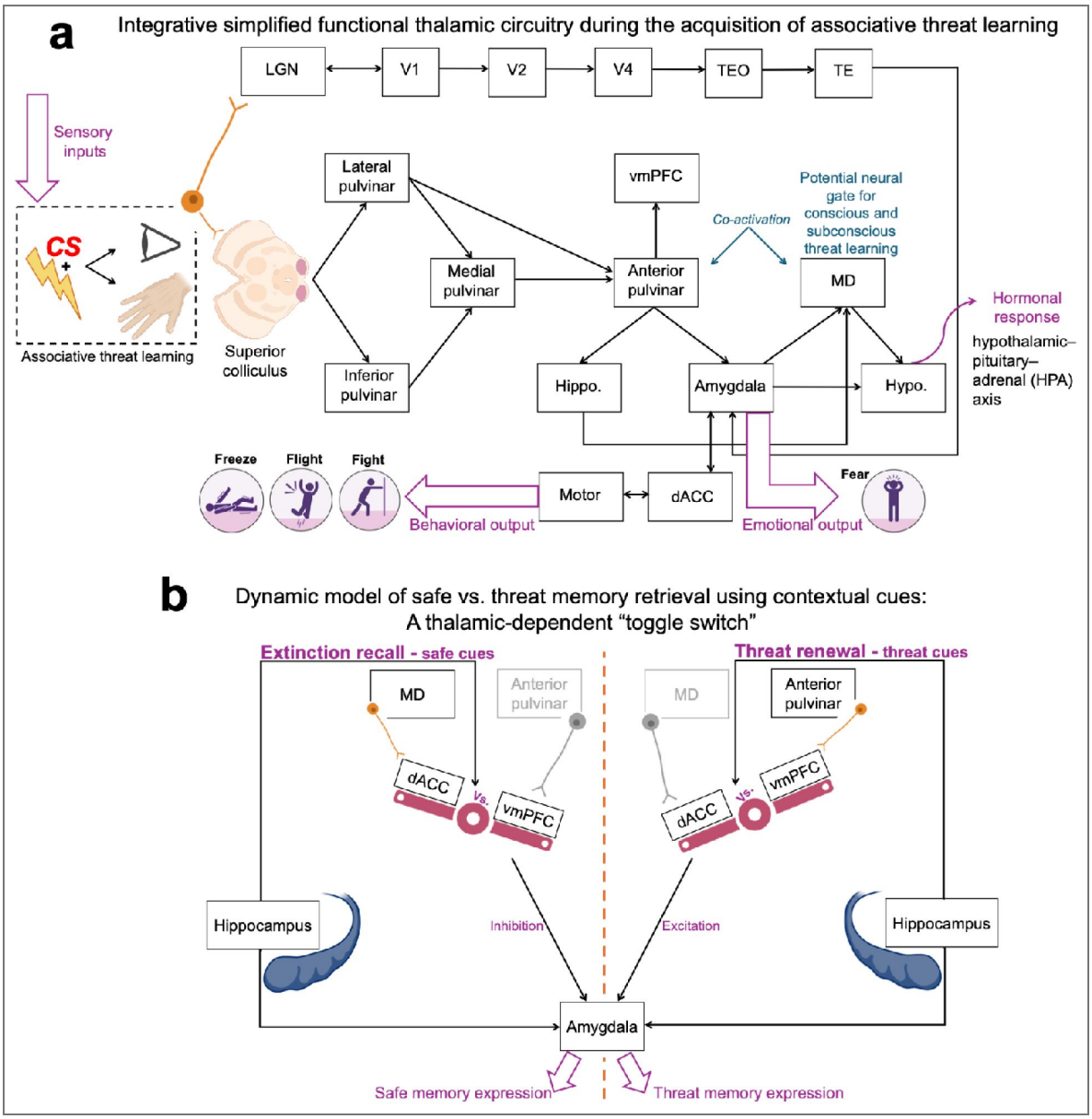


Fig 8. Neurobehavioral models of thalamic involvement in associative threat learning and memory.

a. Schematic illustration of thalamic circuitry during the acquisition of associative threat learning, highlighting interactions between pulvinar divisions, MD, and LGN with key brain regions involved in fear expression. **b.** A thalamic-dependent "toggle switch" regulates the retrieval of safety vs. threat-related memory. The MD-dACC connectivity modulates the interaction between dACC and vmPFC, promoting vmPFC dominance during extinction recall. In contrast, the anterior pulvinar-vmPFC connectivity promotes dACC dominance, enhancing the expression of threat memory during threat renewal. MD: Mediodorsal thalamus. LGN: Lateral geniculate nucleus. vmPFC = Ventromedial prefrontal cortex. dACC = Dorsal anterior cingulate cortex. V1, V2, and V4: Primary, secondary, and fourth visual areas. TEO, TE: Temporal cortex regions. Hypo.: Hypothalamus. Hippo.: Hippocampus. Created using *BioRender*. **Note (panel a):** Known pulvinar–cortical connections, as well as sensory input pathways (e.g., visual inputs via the retina/LGN and nociceptive inputs via the spinothalamic tract), are not explicitly shown. These connections are well established anatomically but were omitted due to their heterogeneity and incomplete characterization at the level of pulvinar subnuclei. Their absence should not be interpreted as a lack of anatomical or functional relevance.

according to this model, involves two distinct pathways: a subcortical route for rapid threat detection and species-specific defense reactions and a cortical route for deliberate threat evaluation and conscious fear experience. Learning within the anterior pulvinar appears to support rapid automatic processes of the subcortical pathway. This aligns with previous reports that demonstrated the role of the pulvinar-amygdala pathway in encoding negative emotions in humans (Bridge et al., 2015; Koller et al., 2019; Kragel et al., 2021; McFadyen et al., 2019; Rafal et al., 2015). Indeed, the increased connectivity that we observed with the amygdala, hippocampus, and vmPFC during threat conditioning aligns with this role. Specifically, in this context, the amygdala likely contributes to tagging a negative emotional valence to CS+ and triggering fight, flight, or freeze reactions (Adolphs et al., 1994; Costa et al., 2022; Fanselow and LeDoux, 1999; Phelps and LeDoux, 2005; Wen et al., 2024, 2022). The connectivity with the hippocampus likely facilitates contextual encoding of the environmental characteristics of the aversive event (Maren, 2001; Maren et al., 2013). In turn, the connectivity with vmPFC appears to support salience encoding for tracking future encounters with the learned threat (Battaglia et al., 2020) and facilitating decision-making and emotion regulation (Milad and Quirk, 2012, 2002; Nejati et al., 2021).

On the other hand, MD activation aligns more closely with the cortical pathway, facilitating conscious and deliberate threat encoding. This is supported by our findings, which showed increased MD activation in response to CS+, compared to the anterior pulvinar during threat learning, probably indicating a different level of threat processing. Additionally, this interpretation is supported by substantial evidence pointing to well-established anatomical pathways connecting the MD with the PFC and demonstrating its role in decision-making and learning (Behrens et al., 2003; Hwang et al., 2020; Li et al., 2022; Wolff and Halassa, 2024). This is along with reports that underscored the contribution of the MD-cortical loops to supporting a conscious experience in humans (Griffiths et al., 2022; Whyte et al., 2024). Together, these findings align with the framework of the two-system model of threat processing, highlighting the anterior pulvinar and MD possible role in the acquisition of the CS-US association and suggesting that these regions might be involved in a parallel conscious vs. unconscious learning proposed by LeDoux and colleagues (LeDoux and Pine, 2016), Fig. 8a.

This suggested rapid initial acquisition of the predictive value of the CS+ is thought to be pronounced during the first two trials. The attenuated CS+ vs. CS− differentiation on the third trial specifically in the pulvinar may reflect a decreased requirement for differential thalamic engagement once the initial association has been acquired, or an initial survival fear reaction is expressed. Notably, because the MD sustained the BOLD response to the CS+ in the third trial which may indicate involvement of this nucleus in the consolidation or stabilization of the learned association. This aligns with the well-established MD-PFC circuit involved in cognitive processes (Wolff and Halassa, 2024). Additionally, in a previous study using a similar paradigm, we observed sustained CS+ vs. CS− differentiation on the third trial in the nucleus reuniens, as well (Tuna et al., 2025). These findings suggest that trial-dependent learning dynamics may vary across thalamic nuclei rather than reflecting a uniform thalamic learning signal. Together, while our paradigm does not inherently distinguish between different stages of learning, such as early acquisition and stabilization, our findings are consistent with stronger associative learning–related engagement during the first two trials, with a reduced differential response by the third trial that may reflect the involvement of different neural processes.

We suggest that different pulvinar divisions seems to hierarchically contribute to the integration of CS-US association in the anterior pulvinar, likely through bottom-up processing. Although the anatomical projections within pulvinar divisions are not well characterized, our functional data-driven approach suggests a possible critical role for the medial pulvinar as a hub region facilitating communication among pulvinar divisions. Specifically, our trial-wise analysis suggested that the medial, inferior, and lateral pulvinar process threat-related information earlier than the anterior pulvinar. The anterior pulvinar exhibited activation in response to CS+ starting at trial 2, whereas activation in the other pulvinar divisions occurred as early as trial 1. Network modeling further identified the medial pulvinar as a potential central hub, likely facilitating

communication with other divisions. Using a robust mediation model, we demonstrated that the medial pulvinar appears to mediate the relay of sensory CS+ information from the inferior and lateral pulvinar to the anterior division probably for higher-order integrative learning. This hierarchical model aligns with previous evidence suggesting that the inferior and lateral pulvinar are associated with processing basic sensory information (Berman and Wurtz, 2010, 2008; Bridge et al., 2015; Cortes et al., 2024). While the medial pulvinar, which receives projections from deep layers of the superior colliculus, supports more advanced functions such as attention and working memory (Bridge et al., 2015; Homman-Ludiye and Bourne, 2019). These findings seem to be indicative of novel insights into the functional specialization of pulvinar divisions during threat learning, suggesting feedforward functions in the inferior and lateral pulvinar and high-order integration in the anterior pulvinar.

Yet, these intrapulvinar relationships should be understood as a functional and computational model, reflecting statistical dependencies among pulvinar divisions during threat learning, rather than as evidence of direct monosynaptic anatomical connections. Because detailed inter-nuclear anatomical connectivity within the pulvinar remains incompletely characterized, our analysis does not presuppose strong direct excitatory projections between subnuclei. Instead, our findings are intended to highlight candidate functional relationships within the pulvinar during conditioning, rather than to provide a definitive anatomical map.

The pulvinar divisions and the MD showed patterns consistent with engagement during extinction learning. In the pulvinar, responses to the CS+E remained relatively stable across the first two trials and subsequently attenuated; a pattern that may reflect rapid updating of threat value. In contrast, MD activation patterns were more consistent with representing the broader task structure across blocks. Despite the involvement in extinction, these nuclei seem to remain sensitive to the extinguished threat during safe contextual cues in extinction recall and exhibited increased activation to CS+E when the stimuli were presented within threat contextual cues during renewal. This pattern suggests engagement in retrieval suppression during recall, along with threat salience processing during both recall and renewal. The MD and pulvinar is suggestive to continue to monitor and evaluate the extinguished threat across contexts rather than relying on static safety memories. These nuclei appear to engage in dynamic evaluation, learning, and decision-making during future encounters with the extinguished CS+.

The classical Pavlovian model proposes that extinction learning forms a new safety memory, competing with the original threat memory acquired during conditioning (Bouton, 2002; Milad and Quirk, 2002). Contextual changes often favor either the safe or threat memory. Maren and colleagues (Maren et al., 2013; Maren and Quirk, 2004) described the hippocampal-prefrontal-amygdala model for contextual memory. The hippocampus projects to the basolateral amygdala, vmPFC, and dACC. Although the vmPFC supports safe memory recall by projecting to the intercalated cells (which inhibit the central nucleus of the amygdala), the dACC enhances threat memory renewal by projecting to the basolateral amygdala (which activates the central nucleus of this structure). Our findings suggest that the thalamic connectivity may influence the balance between recalling safe vs. threat memory by modulating the competitive interaction between the dACC and vmPFC. The increased MD-dACC connectivity during recall may reflect a thalamic-driven reconfiguration of prefrontal circuits, enabling the vmPFC to become more functionally dominant and promote safe memory expression. On the other hand, the anterior pulvinar-vmPFC connectivity during renewal may destabilize the influence of the vmPFC, allowing the dACC to regain dominance and facilitate threat memory retrieval. The thalamic connectivity, thus, appears to orchestrate a context-dependent functional balance between the dACC and vmPFC. This balance may serve as a neural “toggle switch” between safe vs. threat memory expression. We integrated this suggested flexible responding mechanism to changing environmental contexts in Maren’s circuit model of emotional memory, Fig. 8b.

The LGN contribution seems to be consistent across most phases of threat learning, as evidenced by our trial-wise results, which pointed to increased activation in response to CS+ that diminished immediately after the first trial. This distinct pattern might indicate readiness for imminent visual stimuli, regardless of the actual emotional valence, as conditioning, extinction, and renewal

elicited similar BOLD patterns. This aligns with previous studies that highlighted the LGN's engagement in selective attention and anticipation of visual stimuli (Mahoney and Schmidt, 2024; O'Connor et al., 2002; Saalmann and Kastner, 2009). The LGN, as a first-order nucleus (Cortes et al., 2024), primarily supports a feedforward function, while the pulvinar, as part of the visual thalamus, is more closely associated with a broader functional processing (Cortes et al., 2024) including emotional valences of stimuli. This distinction likely underscores their joint but specialized contributions to adaptive threat responses.

Limitations and Future Directions

Although distinct thalamic roles in threat learning have been proposed, fMRI data do not fully capture the complexity of this structure. Pulvinar divisions, MD, and LGN each contain diverse neuron subtypes and finer anatomical divisions that may serve distinct functions. Importantly, the absence of CS+ vs. CS− differences in BOLD activation should not be interpreted as a lack of stimulus-specific processing, as such distinctions may occur without changes in overall activation detectable by fMRI. Future advancements, including higher-resolution human brain atlases, may improve anatomical and functional precision. Additionally, since different sensory modalities preferentially engage distinct thalamic nuclei, the specific thalamic roles we investigated may not be consistent across different experimental designs, particularly in studies using auditory threat learning. The observed relationships among pulvinar divisions during conditioning are purely functional and do not distinguish direct inter-nuclear interactions from indirect coupling mediated by corticothalamic and thalamocortical pathways, including visual cortical regions. Thus, the pulvinar model may reflect indirect cortical loops, or currently undocumented inter-nuclear interactions, or a combination of both. Finally, given the indirect nature of fMRI data and the absence of direct brain signal manipulations, our findings should not be interpreted as evidence of causality. Further research is needed to examine causal mechanisms underlying dynamic neural representations of threat learning within the thalamus.

This study's insights raise critical future questions at the intersection between neuroscience, psychology, and mental health, particularly regarding the neural mechanisms underlying associative threat learning. First, the crucial role of the anterior pulvinar and MD during the acquisition of the CS-US association sets the stage for developing more precise brain interventions, focusing on refining maladaptive fear reactions. Future studies could explore the optimal parameters for applying non-invasive techniques to inhibit these thalamic regions during or immediately after fear conditioning. This approach is particularly promising for clinical populations and individuals exposed to high-risk environments, such as emergency medical staff, firefighters, and paramedics.

Second, the distinct contributions of the MD-dACC and anterior pulvinar-vmPFC circuits to safe vs. threatening memory suggest potential intervention targets for prioritizing one memory over the other. Activating the MD-dACC may facilitate recalling safe memories while engaging the anterior pulvinar-vmPFC circuits might reinforce fear memory pathways. Experimental designs in controlled laboratory settings could target these circuits to identify specific parameters for enhancing safe memories. This avenue holds therapeutic potential, particularly in conditions such as aviophobia; activating a safe memory circuit before a flight might suppress fear relapse during the actual flight.

Third, the anterior pulvinar-MD relationships raise questions related to the wide range of human behaviors, focusing on understanding neural gateways between conscious and unconscious learning. Studying the dynamics of information flow across this promising pathway could deepen our understanding of how sensory and cognitive information transits between conscious and unconscious states and brings about behaviors. This research line may pave the way for innovative learning methods and reveal neurocognitive mechanisms underlying the acquisition of new information in humans.

Together, our findings and emerging research directions underscore the thalamic nuclei's vital role as a hub for fear acquisition and memory processing, offering possible avenues for theoretical advancements and clinical applications. As the primary neural gateway to the human brain, the

thalamus is the first station for all sensory information except olfactory inputs. Modulating its functions is potentially an impactful strategy to induce widespread functional changes across the brain and influence different mental and behavioral expressions.

Materials and Methods

Participants

We analyzed fMRI data of 293 participants during fear conditioning, 320 during extinction, 412 during extinction recall, and 312 during threat renewal. Participants included individuals of both sexes, aged 18–70 years (mean ± SD = 32.17 ± 13.1 years). All participants were proficient in English, right-handed, and had normal or corrected-to-normal vision. The exclusion criteria included a history of seizures or significant head trauma, current substance abuse or dependence, metal implants, pregnancy, breastfeeding, or positive urine toxicology screen for drugs of abuse. We followed the latest version of Helsinki's declaration, and all procedures were approved by the Partners HealthCare Institute Review Board of the Massachusetts General Hospital, Harvard Medical School. All subjects provided written informed consent before taking part in the study. Results from this dataset have been published elsewhere with a different focus (Marin et al., 2020; Wen et al., 2024, 2022). The current results are novel and have not been previously published.

Experimental procedure

Threat learning

Participants underwent two-day sessions of a validated threat learning paradigm in the MRI scanner (Figs. 1a, 4a-6a); the experimental contexts consisted of visual scenes on a computer display. On day 1, participants underwent a Pavlovian conditioning phase in which a neutral stimulus in context A (e.g., red light in an office) was paired with a 500ms electric shock (US) with a partial reinforcement rate of 62.5%. Another neutral stimulus in context A (e.g., blue light in the same office) was also presented but never paired with the shock (CS-). On the same day, we also conducted extinction learning in context B (e.g., red and blue lights in a casebook background). The stimuli were repeatedly presented with the removal of the expected reinforcement (i.e., no shock).

On day 2, the participants underwent two phases of a memory test. The first is extinction recall, including presenting the extinguished CS+ and CS- within safe contextual cues (context B used during extinction learning). The second is threat renewal, in which the stimuli were presented within threat contextual cues (context A used in conditioning). Both extinction recall and threat renewal included no US reinforcement. Across phases of threat learning, the duration of each trial was 6s, and the inter-trial intervals with a fixation screen ranged between 12 to 18 s (15 s on average). Finally, to control for potential confounds related to the experimental design, we pseudorandomized and counterbalanced the order of CS+ and CS− across phases and between subjects.

Shock level

The shock intensity used in the fear learning paradigm was determined during a pre-experiment calibration. Electrodes were attached to the participant's right hand, and stimulation began at a low level (0.1 mA), gradually increasing in small increments. After each increment, participants verbally rated their discomfort. The procedure continued until the participant identified a level they described as "highly annoying but not painful." This individualized intensity was then used for that participant throughout the experiment. For safety and ethical reasons, the maximum intensity was capped at 20 mA, and no participant received a shock above this limit.

Instructions to the participants

Each visual stimulus in our paradigm was first shown to participants for 6 seconds. This initial presentation served as habituation, allowing us to isolate the responses to genuinely new stimuli. Before the experiment began, participants were informed that they would see pictures illuminated with different colored lights, such as red or blue. During the experiment, some pictures might be paired with an electric shock, while others might not. Participants were instructed to pay attention to whether a specific color or pattern was associated with the shock. These instructions were adopted from previous studies in which our group developed this paradigm and found them highly effective for human learning. We therefore used the same approach in the current experiment. These instructions were provided throughout all phases of threat learning, and participants were informed that any shocks delivered would be at the same intensity determined on Day 1.

MRI data acquisition and preprocessing

Two MRI settings were used to acquire the neuroimaging data. The first is a Trio 3T whole-body MRI scanner (Siemens Medical Systems, Iselin, New Jersey) using an 8-channel head coil. The functional data in this setting were acquired using a T2* weighted echo-planar pulse sequence with these parameters: TR = 3.0s, TE = 30 ms, slice number = 45, voxel size = 3 × 3 × 3 mm^3. The second setting was also in the same scanner using a 32-channel head coil. The functional images were obtained using a T2* weighted echo-planar pulse sequence using TR = 2.56s TE = 30 ms, slice number = 48, voxel size = 3 × 3 × 3 mm^3. The anatomical brain images were collected using a T1-weighted MP-RAGE pulse sequence, parcellated into 1 × 1 × 1 mm^3 voxels. Elastic bands were affixed to the head coil device to reduce head motions.

Using the default pipeline in fMRIPrep, version 20.0.2 (Esteban et al., 2020, 2019), we preprocessed the data and applied correction of slice timing, realignment of the functional images, and coregistration. In addition, spatial normalization was performed with data normalized to Montreal Neurological Institute (MNI) space and resampled to a 2 × 2 × 2 mm^3 voxel grid, followed by spatial smoothing with a 6-mm full-width at half-maximum Gaussian kernel.

Activation analyses

We applied the least-squares-based generalized linear model (GLM) for each participant to estimate the BOLD response to CS+ and CS- using SPM 12. We estimated the beta values for each voxel during each learning phase in the paradigm. Overall, the model included 32 regressors for the CS+ and CS-, a regressor for the context, and a regressor for shock in conditioning but not in other phases. The GLM also included six head movement parameters (x, y, z directions, and rotations). This first-level analysis resulted in contrast maps that we used to estimate the variability of these maps across all subjects at the group-level analysis. We then used the contrast maps from the group-level analysis to extract the averaged values across the voxels within predefined masks of the pulvinar divisions, MD, and LGN. These outputs were used to compare the BOLD response during the first four CS+ and four CS- trials. We averaged the activation for each CS type across trials to obtain the block-level activation.

Statistical tests

All statistical analyses were conducted using JASP versions 0.18.3 and 0.19.3 (JASP Team, 2024).

Activation differences across threat-learning phases

For each threat-learning phase, block-level activation was defined as the average activation across trials. Differences in activation between CS+ and CS− conditions within each thalamic nucleus were assessed using paired-samples t-tests.

To examine trial-wise activation dynamics, we conducted 2 × 2 RM-ANOVAs with condition (CS+, CS−) and Trial as within-subject factors. Assumptions of sphericity were evaluated using Mauchly's test, and Greenhouse–Geisser corrections were applied when violations were detected. Significant

effects were followed by post hoc comparisons at the trial level, with false discovery rate (FDR) correction applied to control for multiple comparisons.

Network analyses of within-pulvinar relationships during conditioning

Network analyses examined functional relationships between pulvinar divisions. Nodes corresponded to block-level CS+ minus CS− activation estimates for each pulvinar division, yielding four nodes (one per division). Networks were estimated using a Gaussian graphical model with EBICglasso (LASSO regularization) based on Pearson correlation matrices, with the EBIC tuning parameter set to $\gamma = 0.5$. Edge weights represent partial correlations.

Three centrality measures were computed on the estimated weighted partial-correlation network: node strength, defined as the sum of absolute edge weights connected to a node; closeness, defined as the inverse of the average shortest path length from a node to all other nodes; and betweenness, defined as the proportion of shortest paths between all node pairs that pass through a given node. Shortest paths were computed using inverse edge weights. Centrality indices were normalized. Network accuracy and centrality stability were assessed using nonparametric bootstrapping (10,000 iterations) to estimate CI for edge weights and centrality measures. All network analyses were conducted using the default settings unless otherwise specified, following Epskamp et al., (2018)

Mediation analyses of within-pulvinar relationships during conditioning

Mediation models were estimated using the lavaan package (Rosseel, 2012) with maximum likelihood estimation. All variables were z-standardized prior to analysis. Block-level activation estimates from the inferior and lateral pulvinar were entered as predictors, activation in the medial pulvinar as the mediator, and activation in the anterior pulvinar as the outcome.

To assess the robustness and generalizability of the mediation effects, we conducted K-fold cross-validation ($K = 3$). The full sample ($N = 293$) was randomly partitioned into three non-overlapping sub-samples ($n = 91$, 96, and 106). In each iteration, the mediation model was estimated in one sub-sample, and consistency of indirect effects was evaluated across the remaining sub-samples, yielding six cross-validation iterations.

Indirect effects were evaluated using BCa bootstrap CI based on 10,000 resamples, as recommended by Biesanz et al., (2010). An indirect effect was considered significant when the 95% confidence interval did not include zero and $p < 0.05$.

Connectivity analyses

Functional connectivity was computed using the CONN toolbox (version 22.a) for MATLAB (Nieto-Castanon and Whitfield-Gabrieli, 2022; Whitfield-Gabrieli and Nieto-Castanon, 2012). Anatomical images were segmented into gray matter, white matter, and CSF, and functional images were preprocessed following CONN's default pipeline, including realignment, slice-timing correction, and coregistration. Functional images were then denoised using a component-based noise correction approach, along with regression of motion parameters and outlier scans, followed by band-pass filtering of the BOLD time series between 0.008 and 0.09 Hz.

Following standard denoising procedures—including regression of motion- and physiology-related confounds and temporal filtering—condition-dependent connectivity effects were inferred from subject-level generalized psychophysiological interaction (gPPI) contrast estimates rather than from raw time-series correlations. This GLM-based framework reduces the likelihood that observed PPI effects reflect differences in temporal autocorrelation or spectral properties across regions rather than genuine task-dependent interactions.

For ROI-to-ROI analyses, the pulvinar divisions, MD, and LGN were defined as seed regions, and connectivity was assessed with target regions including the dACC, sgACC, vmPFC, amygdala, and hippocampus. At the first level, functional connectivity between ROIs was computed as Pearson correlation coefficients of the denoised BOLD time series, which were then Fisher z-transformed to normalize the distribution. To estimate task-related connectivity, the first eight trials of each condition were analyzed as a block.

At the second level, these Fisher z-transformed connectivity values were entered into group-level GLMs to evaluate condition-dependent differences. Paired t-tests were used to assess differences between conditions, with false discovery rate (FDR, $p < 0.05$) correction for multiple comparisons. To provide robust estimates of mean differences, 10,000 bootstrap resamples were used to calculate the BCa CI.

Masks

We defined the pulvinar divisions, MD, and LGN nuclei using predefined masks based on the neuroanatomical guidelines of the Automated Anatomical Labelling Atlas(Rolls et al., 2020). We also applied the same atlas guidelines to determine the masks for the two other anatomical regions; the amygdala and hippocampus. The masks for the functional regions were created using Neurosynth (Yarkoni et al., 2011) and the keyword 'conditioning.' For each region, we created 8mm spheres around the following identified peak coordinates: vmPFC (MNIxyz = −2, 46, −10), sgACC (MNIxyz = 0, 26, −12), and dACC (MNIxyz = 0, 14, 28).

Data availability

The data supporting the findings of this study are part of a larger, actively used database and cannot be publicly released at this time due to ongoing projects. The data underlying the analyses reported in this manuscript are available from the corresponding author upon reasonable request and subject to standard data-use agreements. All scripts used for preprocessing and analysis, including workflows implemented in SPM12, and CONN, are available upon request to facilitate replication and reuse.

Acknowledgements

This work was supported by the National Institute of Mental Health grants R01MH123736, R01MH125198, R33MH111907, R01MH097880, and R01MH097964 to M.R.M.

Additional files

Supplementary Table 1

Supplementary Table 2

Supplementary Table 3

Supplementary Table 4

Supplementary Table 5

Supplementary Table 6

Supplementary Table 7

Supplementary Table 8

Supplementary Table 9

Supplementary Figure 1

Supplementary Figure 2

Supplementary Figure 3

Supplementary File 1

Additional information

Funding

Funder	Grant reference number	Author
HHS | NIH | National Institute of Mental Health (NIMH)	R01MH123736	Mohammed R Milad
HHS | NIH | National Institute of Mental Health (NIMH)	R01MH125198	Mohammed R Milad
HHS | NIH | National Institute of Mental Health (NIMH)	R33MH111907	Mohammed R Milad
HHS | NIH | National Institute of Mental Health (NIMH)	R01MH097880	Mohammed R Milad
HHS | NIH | National Institute of Mental Health (NIMH)	R01MH097964	Mohammed R Milad

Author ORCID iDs

Stephen Maren: https://orcid.org/0000-0002-9342-7411
Mohammed R Milad: https://orcid.org/0000-0002-8620-5132

References

Adolphs R, Tranel D, Damasio H, Damasio A (1994) Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**:669-672 https://doi.org/10.1038/372669a0 | PubMed

Arcaro MJ, Pinsk MA, Kastner S (2015) The Anatomical and Functional Organization of the Human Visual Pulvinar. *Journal of Neuroscience* **35**:9848-9871 https://doi.org/10.1523/JNEUROSCI.1575-14.2015 | PubMed

Badarnee M, Wen Z, Hammoud MZ, Glimcher P, Cain CK, Milad MR (2025) Intersect between brain mechanisms of conditioned threat, active avoidance, and reward. *Communications Psychology* **3**:32 https://doi.org/10.1038/S44271-025-00197-7 | PubMed

Battaglia S, Garofalo S, di Pellegrino G, Starita F. (2020) Revaluing the Role of vmPFC in the Acquisition of Pavlovian Threat Conditioning in Humans. *Journal of Neuroscience* **40**:8491-8500 https://doi.org/10.1523/JNEUROSCI.0304-20.2020 | PubMed

Behrens TEJ, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CAM, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, *et al.* (2003) Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience* **6**:750-757 https://doi.org/10.1038/nn1075 | PubMed

Berman RA, Wurtz RH (2010) Functional Identification of a Pulvinar Path from Superior Colliculus to Cortical Area MT. *Journal of Neuroscience* **30**:6342-6354 https://doi.org/10.1523/JNEUROSCI.6176-09.2010 | PubMed

Berman RA, Wurtz RH (2008) Exploring the pulvinar path to visual cortex. *Progress in brain research* **171**:467 https://doi.org/10.1016/S0079-6123(08)00668-7 | PubMed

Bertini C, Pietrelli M, Braghittoni D, Làdavas E (2018) Pulvinar lesions disrupt fear-related implicit visual processing in hemianopic patients. *Frontiers in Psychology* **9**:412050 https://doi.org/10.3389/fpsyg.2018.02329 | PubMed

Biesanz JC, Falk CF, Savalei V (2010) Assessing Mediational Models: Testing and Interval Estimation for Indirect Effects. *Multivariate Behav Res* **45**:661-701 https://doi.org/10.1080/00273171.2010.498292

Bouton ME (2002) Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry* **52**:976-986 https://doi.org/10.1016/S0006-3223(02)01546-9 | PubMed

Bridge H, Leopold DA, Bourne JA (2015) Adaptive pulvinar circuitry supports visual cognition. *Trends in cognitive sciences* **20**:146 https://doi.org/10.1016/J.TICS.2015.10.003 | PubMed

Carr JA (2015) I'll take the low road: The evolutionary underpinnings of visually triggered fear. *Frontiers in Neuroscience* **9**:165437 https://doi.org/10.3389/fnins.2015.00414 | PubMed

Casanova C, Chalupa LM (2023) The dorsal lateral geniculate nucleus and the pulvinar as essential partners for visual cortical functions. *Frontiers in Neuroscience* **17**:1258393 https://doi.org/10.3389/fnins.2023.1258393 | PubMed

Cortes N, Ladret HJ, Abbas-Farishta R, Casanova C (2024) The pulvinar as a hub of visual processing and cortical integration. *Trends in Neurosciences* **47**:120-134 https://doi.org/10.1016/J.TINS.2023.11.008 | PubMed

Costa M, Lozano-Soldevilla D, Gil-Nagel A, Toledano R, Oehrn CR, Kunz L, Yebra M, Mendez-Bertolo C, Stieglitz L, Sarnthein J, *et al.* (2022) Aversive memory formation in humans involves an amygdala-hippocampus phase code. *Nature Communications* **13**:1-16 https://doi.org/10.1038/s41467-022-33828-2 | PubMed

Epskamp S, Borsboom D, Fried EI (2018) Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods* **50**:195-212 https://doi.org/10.3758/s13428-017-0862-1 | PubMed

Esteban O, Ciric R, Finc K, Blair RW, Markiewicz CJ, Moodie CA, Kent JD, Goncalves M, DuPre E, Gomez DEP, *et al.* (2020) Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature protocols* **15**:2186-2202 https://doi.org/10.1038/S41596-020-0327-3 | PubMed

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, *et al.* (2019) fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods* **16**:111-116 https://doi.org/10.1038/S41592-018-0235-4 | PubMed

Fanselow MS, LeDoux JE (1999) Why we think plasticity underlying pavlovian fear conditioning occurs in the basolateral amygdala. *Neuron* **23**:229-232 https://doi.org/10.1016/S0896-6273(00)80775-8 | PubMed

Foygel R, Drton M (2010) Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems* **23** https://doi.org/10.48550/arxiv.1011.6640

Fullana MA, Harrison BJ, 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**:500-508 https://doi.org/10.1038/mp.2015.88 | PubMed

Griffiths BJ, Zaehle T, Repplinger S, Schmitt FC, Voges J, Hanslmayr S, Staudigl T (2022) Rhythmic interactions between the mediodorsal thalamus and prefrontal cortex precede human visual perception. *Nature Communications* **13**:1-11 https://doi.org/10.1038/s41467-022-31407-z | PubMed

Halassa MM, Sherman SM (2019) Thalamo-cortical circuit motifs: a general framework. *Neuron* **103**:762 https://doi.org/10.1016/J.NEURON.2019.06.005 | PubMed

Halverson HE, Freeman JH (2010) Ventral lateral geniculate input to the medial pons is necessary for visual eyeblink conditioning in rats. *Learning & memory* **17**:80-85 https://doi.org/10.1101/LM.1572710 | PubMed

Herry C, Ferraguti F, Singewald N, Letzkus JJ, Ehrlich I, Lüthi A (2010) Neuronal circuits of fear extinction. *European Journal of Neuroscience* **31**:599-612 https://doi.org/10.1111/J.1460-9568.2010.07101.X | PubMed

Homman-Ludiye J, Bourne JA (2019) The medial pulvinar: function, origin and association with neurodevelopmental disorders. *Journal of anatomy* **235**:507-520 https://doi.org/10.1111/JOA.12932 | PubMed

Hummos A, Wang BA, Drammis S, Halassa MM, Pleger B (2022) Thalamic regulation of frontal interactions in human cognitive flexibility. *PLOS Computational Biology* **18**:e1010500 https://doi.org/10.1371/JOURNAL.PCBI.1010500 | PubMed

Hwang K, Bertolero MA, Liu WB, D'Esposito M (2017) The Human Thalamus Is an Integrative Hub for Functional Brain Networks. *Journal of Neuroscience* **37**:5594-5607 https://doi.org/10.1523/JNEUROSCI.0067-17.2017 | PubMed

Hwang K, Bruss J, Tranel D, Boes AD (2020) Network Localization of Executive Function Deficits in Patients with Focal Thalamic Lesions. *Journal of cognitive neuroscience* **32**:2303 https://doi.org/10.1162/JOCN_A_01628 | PubMed

Pavlov I (1904) Nobel Prize Lecture: Physiology of Digestion. https://www.nobelprize.org/prizes/medicine/1904/pavlov/lecture/

JASP Team (2024) JASP. version: v. 0.18.3 and 0.19.3

Koller K, Rafal RD, Platt A, Mitchell ND (2019) Orienting toward threat: Contributions of a subcortical pathway transmitting retinal afferents to the amygdala via the superior colliculus and pulvinar. *Neuropsychologia* **128**:78-86 https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2018.01.027 | PubMed

Kolling N, Wittmann M, Rushworth MFS (2014) Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron* **81**:1190-1202 https://doi.org/10.1016/j.neuron.2014.01.033 | PubMed

Konrad C, Neuhoff L, Adolph D, Goerigk S, Herbert JS, Jagusch-Poirier J, Weigelt S, Seehagen S, Schneider S (2024) Associative learning via eyeblink conditioning differs by age from infancy to adulthood. *Communications Psychology* **2**:1-13 https://doi.org/10.1038/s44271-024-00176-4 | PubMed

Korn CW, Bach DR (2019) Minimizing threat via heuristic and optimal policies recruits hippocampus and medial prefrontal cortex. *Nature Human Behaviour* **3**:733-745 https://doi.org/10.1038/s41562-019-0603-9 | PubMed

Korn CW, Bach DR (2018) Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications* **9**:1-15 https://doi.org/10.1038/s41467-017-02750-3 | PubMed

Kragel PA, Čeko M, Theriault J, Chen D, Satpute AB, Wald LW, Lindquist MA, Feldman Barrett L, Wager TD (2021) A human colliculus-pulvinar-amygdala pathway encodes negative emotion. *Neuron* **109**:2404-2412.e5. https://doi.org/10.1016/j.neuron.2021.06.001 | PubMed

Krasne FB, Zinn R, Vissel B, Fanselow MS, Franklin Krasne CB, Howland-Rose V, Battersby D, King D (2021) Extinction and discrimination in a Bayesian model of context fear conditioning (BaconX). *Hippocampus* **31**:790-814 https://doi.org/10.1002/HIPO.23298 | PubMed

Van Le Q, Isbell LA, Matsumoto J, Nguyen M, Hori E, Maior RS, Tomaz C, Tran AH, Ono T, Nishijo H. (2013) Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proceedings of the National Academy of Sciences of the United States of America* **110**:19000-19005 https://doi.org/10.1073/pnas.1312648110 | PubMed

LeDoux JE, Pine DS (2016) Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework. *Am J Psychiatry* **173**:1083-1093 https://doi.org/10.1176/APPI.AJP.2016.16030353

Lee S, Shin HS (2016) The role of mediodorsal thalamic nucleus in fear extinction. *Journal of Analytical Science and Technology* **7**:1-5 https://doi.org/10.1186/s40543-016-0093-6

Lee S, Ahmed T, Lee S, Kim H, Choi S, Kim DS, Kim SJ, Cho J, Shin HS (2011) Bidirectional modulation of fear extinction by mediodorsal thalamic firing in mice. *Nature Neuroscience* **15**:308-314 https://doi.org/10.1038/nn.2999 | PubMed

Li K, Fan L, Cui Y, Wei X, He Y, Yang J, Lu Y, Li W, Shi W, Cao L, *et al.* (2022) The human mediodorsal thalamus: Organization, connectivity, and function. *NeuroImage* **249**:118876 https://doi.org/10.1016/J.NEUROIMAGE.2022.118876 | PubMed

Li XB, Inoue T, Nakagawa S, Koyama T (2004) Effect of mediodorsal thalamic nucleus lesion on contextual fear conditioning in rats. *Brain Research* **1008**:261-272 https://doi.org/10.1016/J.BRAINRES.2004.02.038 | PubMed

Lithari C, Moratti S, Weisz N (2015) Thalamocortical interactions underlying visual fear conditioning in humans. *Human Brain Mapping* **36**:4592 https://doi.org/10.1002/HBM.22940 | PubMed

Mahoney HL, Schmidt TM (2024) The cognitive impact of light: illuminating ipRGC circuit mechanisms. *Nature Reviews Neuroscience* **25**:159-175 https://doi.org/10.1038/s41583-023-00788-5 | PubMed

Maren S (2001) Nuerobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience* **24**:897-931 PubMed | https://doi.org/10.1146/annurev.neuro.24.1.897

Maren S, Phan KL, Liberzon I (2013) The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience* **14**:417-428 https://doi.org/10.1038/nrn3492 | PubMed

Maren S, Quirk GJ (2004) Neuronal signalling of fear memory. *Nature Reviews Neuroscience* **5**:844-852 https://doi.org/10.1038/nrn1535 | PubMed

Marin MF, Hammoud MZ, Klumpp H, Simon NM, Milad MR (2020) Multimodal Categorical and Dimensional Approaches to Understanding Threat Conditioning and Its Extinction in Individuals With Anxiety Disorders. *JAMA Psychiatry* **77**:618-627 https://doi.org/10.1001/JAMAPSYCHIATRY.2019.4833 | PubMed

McFadyen J, Mattingley JB, Garrido MI (2019) An afferent white matter pathway from the pulvinar to the amygdala facilitates fear recognition. *eLife* **8** https://doi.org/10.7554/ELIFE.40766 | PubMed

Milad MR, Furtak SC, Greenberg JL, Keshaviah A, Im JJ, Falkenstein MJ, Jenike M, Rauch SL, Wilhelm S (2013) Deficits in Conditioned Fear Extinction in Obsessive-Compulsive Disorder and Neurobiological Changes in the Fear Circuit. *JAMA Psychiatry* **70**:608-618 https://doi.org/10.1001/JAMAPSYCHIATRY.2013.914 | PubMed

Milad MR, Pitman RK, Ellis CB, Gold AL, Shin LM, Lasko NB, Zeidan MA, Handwerger K, Orr SP, Rauch SL (2009) Neurobiological Basis of Failure to Recall Extinction Memory in Posttraumatic Stress Disorder. *Biological Psychiatry* **66**:1075-1082 https://doi.org/10.1016/J.BIOPSYCH.2009.06.026 | PubMed

Milad MR, Quirk GJ (2012) Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. *Annual review of psychology* **63**:129 https://doi.org/10.1146/ANNUREV.PSYCH.121208.131631 | PubMed

Milad MR, Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* **420**:70-74 https://doi.org/10.1038/nature01138 | PubMed

Milad MR, Rosenbaum BL, Simon NM (2014) Neuroscience of fear extinction: Implications for assessment and treatment of fear-based and anxiety related disorders. *Behaviour Research and Therapy* **62**:17-23 https://doi.org/10.1016/J.BRAT.2014.08.006 | PubMed

Milad MR, Wright CI, Orr SP, Pitman RK, Quirk GJ, Rauch SL (2007) Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological psychiatry* **62**:446-454 https://doi.org/10.1016/J.BIOPSYCH.2006.10.011 | PubMed

Mukherjee A, Lam NH, Wimmer RD, Halassa MM (2021) Thalamic circuits for independent control of prefrontal signal and noise. *Nature* **600**:100-104 https://doi.org/10.1038/s41586-021-04056-3 | PubMed

Nejati V, Majdi R, Salehinejad MA, Nitsche MA (2021) The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions. *Scientific Reports* **11**:1-12 https://doi.org/10.1038/s41598-021-81454-7 | PubMed

Nieto-Castanon A, Whitfield-Gabrieli S (2022) CONN functional connectivity toolbox: RRID SCR_009550, release 22. https://doi.org/10.56441/HILBERTPRESS.2246.5840

O'Connor DH, Fukui MM, Pinsk MA, Kastner S (2002) Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience* **5**:1203-1209 https://doi.org/10.1038/nn957 | PubMed

Paydar A, Lee B, Gangadharan G, Lee S, Hwang EM, Shin HS (2014) Extrasynaptic GABAA receptors in mediodorsal thalamic nucleus modulate fear extinction learning. *Molecular Brain* **7**:39 https://doi.org/10.1186/1756-6606-7-39 | PubMed

Penzo MA, Robert V, Tucciarone J, De Bundel D, Wang M, Van Aelst L, Darvas M, Parada LF, Palmiter RD, He M, *et al.* (2015) The paraventricular thalamus controls a central amygdala fear circuit. *Nature* **519**:455-459 https://doi.org/10.1038/NATURE13978 | PubMed

Pessoa L, Adolphs R (2010) Emotion processing and the amygdala: from a "low road" to "many roads" of evaluating biological significance. *Nature Reviews Neuroscience* **11**:773-782 https://doi.org/10.1038/nrn2920 | PubMed

Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* **48**:175-187 https://doi.org/10.1016/j.neuron.2005.09.025 | PubMed

Rafal RD, Koller K, Bultitude JH, Mullins P, Ward R, Mitchell AS, Bell AH (2015) Connectivity between the superior colliculus and the amygdala in humans and macaque monkeys: virtual dissection with probabilistic DTI tractography. *Journal of neurophysiology* **114**:1947-1962 https://doi.org/10.1152/JN.01016.2014 | PubMed

Ramanathan KR, Jin J, Giustino TF, Payne MR, Maren S (2018) Prefrontal projections to the thalamic nucleus reuniens mediate fear extinction. *Nature Communications* **9**:1-12 https://doi.org/10.1038/s41467-018-06970-z | PubMed

Ramanathan KR, Maren S (2019) Nucleus reuniens mediates the extinction of contextual fear conditioning. *Behavioural Brain Research* **374**:112114 https://doi.org/10.1016/J.BBR.2019.112114 | PubMed

Ratigan HC, Krishnan S, Smith S, Sheffield MEJ (2023) A thalamic-hippocampal CA1 signal for contextual fear memory suppression, extinction, and discrimination. *Nature Communications* **14**:1-17 https://doi.org/10.1038/s41467-023-42429-6 | PubMed

Rolls ET, Huang CC, Lin CP, Feng J, Joliot M (2020) Automated anatomical labelling atlas 3. *NeuroImage* **206** https://doi.org/10.1016/J.NEUROIMAGE.2019.116189 | PubMed

Rosseel Y (2012) lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* **48**:1-36 https://doi.org/10.18637/JSS.V048.I02

Saalmann YB, Kastner S (2011) Cognitive and Perceptual Functions of the Visual Thalamus. *Neuron* **71**:209 https://doi.org/10.1016/J.NEURON.2011.06.027 | PubMed

Saalmann YB, Kastner S (2009) Gain control in the visual thalamus during perception and cognition. *Current opinion in neurobiology* **19**:408 https://doi.org/10.1016/J.CONB.2009.05.007 | PubMed

Salay LD, Huberman AD (2021) Divergent outputs of the ventral lateral geniculate nucleus mediate visually evoked defensive behaviors. *Cell Reports* **37** https://doi.org/10.1016/j.celrep.2021.109792 | PubMed

Sherman SM (2016) Thalamus plays a central role in ongoing cortical functioning. *Nature Neuroscience* **19**:533-541 https://doi.org/10.1038/nn.4269 | PubMed

Sherman SM (2007) The thalamus is more than just a relay. *Current opinion in neurobiology* **17**:417-422 https://doi.org/10.1016/J.CONB.2007.07.003 | PubMed

Sherman SM, Guillery RW. (2006) *Exploring the thalamus and its role in cortical function* **484**

Shi C, Davis M (2001) Visual pathways involved in fear conditioning measured with fear-potentiated startle: behavioral and anatomic studies. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **21**:9844-9855 https://doi.org/10.1523/JNEUROSCI.21-24-09844.2001 | PubMed

Silverstein DN, Ingvar M (2015) A multi-pathway hypothesis for human visual fear signaling. *Frontiers in Systems Neuroscience* **9**:137726 https://doi.org/10.3389/fnsys.2015.00101 | PubMed

Steinmetz AB, Buss EW, Freeman JH (2013) Inactivation of the ventral lateral geniculate and nucleus of the optic tract impairs retention of visual eyeblink conditioning. *Behavioral Neuroscience* **127**:690-693 https://doi.org/10.1037/A0033729 | PubMed

Takakuwa N, Isa K, Onoe H, Takahashi J, Isa T (2021) Contribution of the Pulvinar and Lateral Geniculate Nucleus to the Control of Visually Guided Saccades in Blindsight Monkeys. *Journal of Neuroscience* **41**:1755-1768 https://doi.org/10.1523/JNEUROSCI.2293-20.2020 | PubMed

Totty MS, Tuna T, Ramanathan KR, Jin J, Peters SE, Maren S (2023) Thalamic nucleus reuniens coordinates prefrontal-hippocampal synchrony to suppress extinguished fear. *Nature Communications* **14**:1-12 https://doi.org/10.1038/s41467-023-42315-1 | PubMed

Tovote P, Fadok JP, Lüthi A (2015) Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience* **16**:317-331 https://doi.org/10.1038/nrn3945 | PubMed

Tuna T, Totty MS, Badarnee M, Mourão FAG, Peters S, Milad MR, Maren S (2025) Associative coding of conditioned fear in the thalamic nucleus reuniens in rodents and humans. *Communications Biology* **8**:1-14 https://doi.org/10.1038/s42003-025-08580-0 | PubMed

Vuilleumier P, Armony JL, Driver J, Dolan RJ (2003) Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience* **6**:624-631 https://doi.org/10.1038/nn1057 | PubMed

Ward R, Danziger S, Bamford S (2005) Response to visual threat following damage to the pulvinar. *Current biology : CB* **15**:571-573 https://doi.org/10.1016/J.CUB.2005.01.056 | PubMed

Wei P, Liu N, Zhang Z, Liu X, Tang Y, He X, Wu B, Zhou Z, Liu Y, Li J, *et al.* (2015) Processing of visually evoked innate fear by a non-canonical thalamic pathway. *Nature Communications* **6**:1-13 https://doi.org/10.1038/ncomms7756 | PubMed

Wen Z, Pace-Schott EF, Lazar SW, Rosén J, Åhs F, Phelps EA, LeDoux JE, Milad MR (2024) Distributed neural representations of conditioned threat in the human brain. *Nature Communications* **15**:1-14 https://doi.org/10.1038/s41467-024-46508-0 | PubMed

Wen Z, Raio CM, Pace-Schott EF, Lazar SW, LeDoux JE, Phelps EA, Milad MR (2022) Temporally and anatomically specific contributions of the human amygdala to threat and safety learning. *Proceedings of the National Academy of Sciences of the United States of America* **119**:e2204066119 https://doi.org/10.1073/pnas.2204066119 | PubMed

Whitfield-Gabrieli S, Nieto-Castanon A (2012) Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain connectivity* **2**:125-141 https://doi.org/10.1089/BRAIN.2012.0073 | PubMed

Whyte CJ, Redinbaugh MJ, Shine JM, Saalmann YB (2024) Thalamic contributions to the state and contents of consciousness. *Neuron* **112**:1611-1625 https://doi.org/10.1016/j.neuron.2024.04.019 | PubMed

Wolff M, Halassa MM (2024) The mediodorsal thalamus in executive control. *Neuron* **112**:893-908 https://doi.org/10.1016/j.neuron.2024.01.002 | PubMed

Wurtz RH, McAlonan K, Cavanaugh J, Berman RA (2011) Thalamic pathways for active vision. *Trends in Cognitive Sciences* **15**:177-184 https://doi.org/10.1016/J.TICS.2011.02.004 | PubMed

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8**:665-670 https://doi.org/10.1038/nmeth.1635 | PubMed

Zheng C, Huang Y, Bo B, Wei L, Liang Z, Wang Z (2020) Projection from the Anterior Cingulate Cortex to the Lateral Part of Mediodorsal Thalamus Modulates Vicarious Freezing Behavior. *Neuroscience Bulletin* **36**:217-229 https://doi.org/10.1007/s12264-019-00427-z | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

Badarnee and colleagues analyse fMRI data collected during an associative threat-learning task. They find evidence for parallel processes mediated by the mediodorsal, LGn and pulvinar nuclei of the thalamus. The evidence for these conclusions is promising, but limited by a lack of clarity regarding the preprocessing and statistical methods.

Strengths:

The approach is inventive and novel, providing information about thalamocortical interactions that are scant in the current literature.

Weaknesses:

(1) There are not sufficient details present to allow for the direct interrogation of the methods used in the study.

(2) The figures do not contain sufficiently granular details, making it challenging to determine whether the observed effects were robust to individual differences.

Comments on revisions:

I continue to recommend the plotting of individual data points. While there may be individual variance, it is important to quantify this in publication so that future studies can appreciate the uncertainty surrounding test statistics.

https://doi.org/10.7554/eLife.108043.2.sa3

Reviewer #2 (Public review):

Summary:

The authors quantify human fMRI BOLD responses in pulvinar and mediodorsal thalamic nuclei during a fear conditioning and extinction task across two days, in a large sample size (hundreds of participants). They show that the BOLD responses in these areas differentiate the conditioned (CS+) and safety (CS-) stimulus. Additionally this changes with repeated trials which could be a neural correlate of fear learning. They show that the anterior pulvinar is most correlated with the MD, and that this is not due to anatomical proximity. They perform graph analysis on the pulvinar sub nuclei which suggests that the medial pulvinar is a hub between the sensory (lateral/inferior) and associative (anterior) pulvinar. They show different patterns of thalamic activity across conditioning, extinction, recall, and renewal.

Strengths:

The data has a large sample size (n=293 in some measures, n=412 in others). This is a validated human fear conditioning/extinction task that Dr Milad's group has been working with for several years. Few labs have investigated the thalamus activity during fear conditioning and extinction, particularly with a large sample size. There is an independent replication of the pulvinar network structure (Fig. 3), which suggests that the processing in the more sensory-related inferior and lateral pulvinar is relayed to the anterior pulvinar (and possibly thereby to more action-related prefrontal areas) via an intermediate step in the medial pulvinar - potentially a novel discovery but that needs more validation.

Weaknesses:

(1) The authors cannot make causal claims about their results based on correlational neuroimaging evidence. Causal claims should be pared back. E.g. Sentence 1 in results "The anterior pulvinar and MD contribute to early associative threat learning, as evidenced by increased functional activation in response to CS+ compared to CS- at the block level (Fig. 1b-c)." needs to be reworded to something like 'the anterior pulvinar and MD have increased functional activation... This suggests that these areas may contribute to early associate threat learning"

(2) Fig .1 The fact that the difference in BOLD activity between CS+ and CS- goes away on the third trial is not addressed. This is a very large effect in the data.

(3) Fig. 3 Could the observed network structure be due to anatomical proximity? Perhaps the authors should do an analogous analysis to what they did in Fig. 2 for this intra-pulvinar analysis. This analysis doesn't take into account the indirect connections through corticothalamic and thalamocortical connections with visual cortex and the pulvinar. There is an implicit assumption that there are interconnections between the pulvinar sub nuclei, but there are few strong excitatory projections between these sub nuclei to my knowledge. If visual areas are included in the graph, it would make things more complex, but would probably dramatically change the story. In this way, the message is somewhat constructed or arbitrary.

(4) In the results section describing Fig. 4-7, there are no statistics supporting the claims made.

There needs to be a set of graphs comparing the results across the study sessions and days, with statistical comparisons between the different experiments to confirm differences.

(5) FIg. 7 does not include the major corticothalamic and thalamocortical projections from early, mid-level, and higher visual cortex to the different pulvinar nuclei. I doubt that there are strong direct projections between the pulvinar nuclei, rather the functional connections are probably mediated through interconnections with cortical visual areas.

(6) Stylistic: There are a lot of hypotheses and interpretations presented in this primary literature paper which may be better suited for a review or perspective piece.

(7) In the discussion there is an assumption that the fMRI BOLD responses to CS+ and CS- need to be different to indicate that an area is processing these distinctly, but the BOLD signal can only detect large scale changes in overall activity. It's easy to imagine that an area could be involved in processing these two stimuli distinctly without showing an overall difference in the gross amount of activity.

(8) There is strong evidence that the BOLD responses to the threat-related and safety-related stimuli are different, modest evidence for their claims of learning/plasticity in these pathways, and circumstantial evidence supporting their hypothesized graph network models. Overall most of the claims made in the discussion are better considered possible interpretations rather than proven findings - this is not a criticism, as these experiments and subject matter are extremely complex.

(9) This study continues to validate the power and utility of this in human fear conditioning/extinction paradigm, and extends this paradigm to investigating fear learning beyond the traditional limbic system pathways. It's possible that their models for the pulvinar nuclei interconnections could guide future neuromodulation or DBS studies that could provide more causal evidence for their hypotheses.

Comments on revisions:

The reviewers addressed my major concerns appropriately in the modified manuscript. As long as the MRI analysis concerns of Reviewer 3 are satisfied (MRI analysis is not my expertise), I am satisfied with the modified manuscript.

https://doi.org/10.7554/eLife.108043.2.sa2

Reviewer #3 (Public review):

Summary:

The present work was aimed at investigating the specific contributions of thalamic nuclei to associative threat learning and extinction. Using fMRI, it examined activation patterns across pulvinar divisions, the lateral geniculate nucleus (LGN), and the mediodorsal thalamus (MD) during threat acquisition, extinction, and recall. It goals was to uncover whether distinct thalamic systems support different modes of learning-automatic survival mechanisms versus more deliberate processes-and to propose a hierarchical pulvinar model of fear conditioning. The manuscript also tried to refine current neuroanatomical models of threat learning and memory, highlighting the role of thalamic nuclei in it.

Strengths:

(1) Valuable theoretical elaboration and modeling regarding the differential role of pulvinar subdivisions on feedforward (inferior, lateral) and higher-order integration (anterior), and their functional interplay with other relevant subcortical and cortical structures in associative threat and extinction learning.

(2) Large sample sizes and multipronged analytical approaches were used for hypothesis testing.

(3) Exhaustive literature review in the field of associative threat, as well as regarding the role of thalamic nuclei and other brain structures in it.

Weaknesses:

(1) The manuscript has improved methodologically and analytically after the review. Several weaknesses remain, in my opinion, but still findings are valuable and the evidence can be considered as convincing.

a) fMRI data have low resolution (3 cubic mm), which certainly limits the examination of small nuclei such as the ones investigated here, and especially the examination of the LGN and inferior pulvinar.

b) fMRI was normalized to standard space. Analyzing the data in individual-subject space would have given you the options of avoiding altering every participant's brain and of using more precise atlases than the normalized AAL for ROI selection.

c) Motion during scanning was poorly controlled. Including the motion parameters as covariates of no interest in the GLM/analysis does not fully guarantee that motion is not influencing the results, and that motion is not differentially influencing some experimental conditions more than others.

https://doi.org/10.7554/eLife.108043.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***Public review:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *Badarnee and colleagues analyse fMRI data collected during an associative threatlearning task. They find evidence for parallel processes mediated by the mediodorsal, LGn, and pulvinar nuclei of the thalamus. The evidence for these conclusions is promising, but limited by a lack of clarity regarding the preprocessing and statistical methods.*
>
> *Strengths:*
>
> *The approach is inventive and novel, providing information about thalamocortical interactions that are scant in the current literature.*
>
> *Weaknesses:*
>
> *(1) There are not sufficient details present to allow for the direct interrogation of the methods used in the study.*

We thank the reviewer for this comment. We have added more detailed information about the methods to clarify our procedure. In addition to the original description of our threat learning paradigm in humans, we included the following to page 39-40:

"Experimental procedure

Threat learning: Please see the original description in the manuscript.

Shock level: The shock intensity used in the fear learning paradigm was determined during a preexperiment calibration. Electrodes were attached to the participant's right hand, and stimulation began at a low level (0.1 mA), gradually increasing in small increments. After each increment, participants verbally rated their discomfort. The procedure continued until the participant identified a level they described as "highly annoying but not painful." This individualized intensity was then used for that participant throughout the experiment. For safety and ethical reasons, the maximum intensity was capped at 20 mA, and no participant received a shock above this limit.

Instructions to the participants: Each visual stimulus in our paradigm was first shown to participants for 6 seconds. This initial presentation served as habituation, allowing us to isolate the responses to genuinely new stimuli. Before the experiment began, participants were informed that they would see pictures illuminated with different colored lights, such as red or blue. During the experiment, some pictures might be paired with an electric shock, while others might not. Participants were instructed to pay attention to whether a specific color or pattern was associated with the shock. These instructions were adopted from previous studies in which our group developed this paradigm and found them highly effective for human learning. We therefore used the same approach in the current experiment. These instructions were provided throughout all phases of threat learning, and participants were informed that any shocks delivered would be at the same intensity determined on Day 1."

> *(2) The figures do not contain sufficiently granular details, making it challenging to determine whether the observed effects were robust to individual differences.*

We thank the reviewer for this suggestion. We agree that visualizations exposing the full data distribution can be highly informative, and we therefore present distribution-based plots for several analyses (e.g., connectivity results in Figure 7). However, for the activation analyses, our primary goal was to highlight trial-to-trial changes and overall patterns across thalamic nuclei, rather than the distribution of individual data points per se. For this purpose, bar plots with standard errors provide a clearer representation of the directional effects and facilitate comparison across trials and conditions.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *The authors quantify human fMRI BOLD responses in pulvinar and mediodorsal thalamic nuclei during a fear conditioning and extinction task across two days, in a large sample size (hundreds of participants). They show that the BOLD responses in these areas differentiate the conditioned (CS+) and safety (CS-) stimuli. Additionally, this changes with repeated trials, which could be a neural correlate of fear learning. They show that the anterior pulvinar is most correlated with the MD, and that this is not due to anatomical proximity. They perform graph analysis on the pulvinar subnuclei, which suggests that the medial pulvinar is a hub between the sensory (lateral/inferior) and associative (anterior) pulvinar. They show different patterns of thalamic activity across conditioning, extinction, recall, and renewal.*
>
> *Strengths:*
>
> *The data has a large sample size (n=293 in some measures, n=412 in others). This is a validated human fear conditioning/extinction task that Dr Milad's group has been working with for several years. Few labs have investigated the thalamus activity during fear conditioning and extinction, particularly with a large sample size. There is an independent replication of the pulvinar network structure (Figure 3), which suggests that the processing in the more sensory-related inferior and lateral pulvinar is relayed to the anterior pulvinar (and possibly thereby to more action-related prefrontal areas) via an intermediate step in the medial pulvinar - potentially a novel discovery, but that needs more validation.*
>
> *Weaknesses:*
>
> *(1) The authors cannot make causal claims about their results based on correlational neuroimaging evidence. Causal claims should be pared back. E.g., sentence 1 in the Results section: "The anterior pulvinar and MD contribute to early associative threat learning, as evidenced by increased functional activation in response to CS+ compared to CS- at the block level (Fig. 1b-c)." needs to be reworded to something like "The anterior pulvinar and MD have increased functional activation... This suggests that these areas may contribute to early associate threat learning."*

We acknowledge the limitations of fMRI studies and agree with the reviewer that causal claims cannot be made based on correlational neuroimaging evidence. Accordingly, we revised the text to reduce causal interpretations. Specifically, we reworded the sentence identified by the reviewer in the Results section and systematically updated language throughout the manuscript.

Page 9: “At the block level, both the anterior pulvinar and MD showed increased activation to CS+ vs. CS− (anterior pulvinar: $t_{(292)} = 4.41$, $p = 0.00001$, $d = 0.25$; *MD*: $t_{(292)} = 6.41$, $p = 5.83x10^{-}$

^{10}, *d = 0.37*; Fig. 1b–c), suggesting a possible involvement of these regions in early associative threat learning."

Throughout the manuscript, we replaced terms such as "reflects" with "likely reflects" and "indicating" with "consistent with," and introduced explicitly correlational phrasing where appropriate (e.g., "apparently," "closely align," and "seems to"). All revisions are highlighted in green in the revised manuscript.

> *(2) Figure 1: The fact that the difference in BOLD activity between CS+ and CS- goes away on the third trial is not addressed. This is a very large effect in the data.*

We thank the reviewer for highlighting this important pattern in Trial 3. The CS+ vs. CS− contrast in the third trial in the mediodorsal thalamus remained statistically significant after FDR correction and was correctly reported in the Supplementary Tables. However, we acknowledge that the statistical marker was inadvertently omitted from Figure 1. We have now corrected the figure to include the appropriate significance annotation.

In addition, we now explicitly describe the attenuation of the CS+ vs. CS− difference by the third trial in the mediodorsal thalamus but not in the pulvinar (page 32):

"This suggested rapid initial acquisition of the predictive value of the CS+ is thought to be pronounced during the first two trials. The attenuated CS+ vs. CS− differentiation on the third trial specifically in the pulvinar may reflect a decreased requirement for differential thalamic engagement once the initial association has been acquired, or an initial survival fear reaction is expressed. Notably, because the MD sustained the BOLD response to the CS+ in the third trial which may indicate involvement of this nucleus in the consolidation or stabilization of the learned association. This aligns with the wellestablished MD-PFC circuit involved in cognitive processes (Wolff and Halassa, 2024). Additionally, in a previous study using a similar paradigm, we observed sustained CS+ vs. CS− differentiation on the third trial in the nucleus reuniens, as well (Tuna et al., 2025). These findings suggest that trialdependent learning dynamics may vary across thalamic nuclei rather than reflecting a uniform thalamic learning signal. Together, while our paradigm does not inherently distinguish between different stages of learning, such as early acquisition and stabilization, our findings are consistent with stronger associative learning–related engagement during the first two trials, with a reduced differential response by the third trial that may reflect the involvement of different neural processes".

> *(3) Figure 3: Could the observed network structure be due to anatomical proximity? Perhaps the authors should do an analogous analysis to what they did in Figure 2 for this intra-pulvinar analysis. This analysis doesn't take into account the indirect connections through corticothalamic and thalamocortical connections with the visual cortex and the pulvinar. There is an implicit assumption that there are interconnections between the pulvinar subnuclei, but there are few strong excitatory projections between these subnuclei to my knowledge. If visual areas are included in the graph, it would make things more complex, but would probably dramatically change the story. In this way, the message is somewhat constructed or arbitrary.*

We thank the reviewer for this insightful comment. We agree that the network analysis in Figure 3 does not provide a direct anatomical account of pulvinar connectivity and cannot distinguish between direct inter-nuclear interactions and indirect coupling mediated via corticothalamic and thalamocortical pathways, including visual cortex.

Our intention with this analysis was to characterize functional statistical dependencies among pulvinar divisions during conditioning, rather than to infer monosynaptic anatomical connectivity. Accordingly, the observed network structure should not be interpreted as evidence for direct excitatory projections between pulvinar subnuclei.

We agree that including visual cortical regions in the network would substantially increase model complexity and could alter the inferred network structure. However, doing so would require a trial-wise, multiregional modeling framework that goes beyond the scope of the present intra-pulvinar analysis.

In response to this comment, we have now explicitly clarified the assumptions, interpretational limits, and alternative explanations of the network model in the Discussion (page 33):

"Yet, these intrapulvinar relationships should be understood as a functional and computational model, reflecting statistical dependencies among pulvinar divisions during threat learning, rather than as evidence of direct monosynaptic anatomical connections. Because detailed inter-nuclear anatomical connectivity within the pulvinar remains incompletely characterized, our analysis does not presuppose strong direct excitatory projections between subnuclei. Instead, our findings are intended to highlight candidate functional relationships within the pulvinar during conditioning with different level of data processing, rather than to provide a definitive anatomical map."

We also included the following in the Limitations and Future Directions section (page 36):

"The observed relationships among pulvinar divisions during conditioning are purely functional and do not distinguish direct inter-nuclear interactions from indirect coupling mediated by corticothalamic and thalamocortical pathways, including visual cortical regions. Thus, the pulvinar model may reflect indirect cortical loops, weak or currently undocumented inter-nuclear interactions, or a combination of both."

Finally, we added this note to the legend of Fig. 3:

"Note: The functional relationships among pulvinar divisions during threat learning should be interpreted as computational dependencies derived from statistical associations. These effects may reflect indirect interactions mediated by corticothalamic and thalamocortical pathways (e.g., via visual cortex), rather than direct inter-nuclear connectivity. Elucidating the underlying anatomical mechanisms will require future studies."

> *(3) In the results section describing Figures 4-7, there are no statistics supporting the claims made. There needs to be a set of graphs comparing the results across the study sessions and days, with statistical comparisons between the different experiments to confirm differences.*

We thank the reviewer for this suggestion. In this study, each phase (conditioning, extinction, recall, and renewal) was analyzed separately to characterize thalamic function within that specific phase. Our primary conclusions focus on differences between CS+ and CS− within each phase, rather than comparisons across phases or sessions. Direct statistical comparisons across phases were therefore not performed, as they fall outside the scope of our main hypotheses.

We have clarified this in the revised manuscript to make the rationale for our analytic approach explicit. Added to page 8:

"The purpose of this study is to investigate thalamic function during each learning phase separately, focusing on CS+ vs. CS− differences within phases rather than comparing activation across phases. This phase-specific approach allows us to characterize thalamic functional dynamics within each stage of learning and memory, avoiding potential confounds arising from the distinct processes of conditioning, extinction, and recall."

> *(4) Figure 7 does not include the major corticothalamic and thalamocortical projections from early, mid-level, and higher visual cortex to the different pulvinar nuclei. I doubt*

> *that there are strong direct projections between the pulvinar nuclei; rather, the functional connections are probably mediated through interconnections with cortical visual areas.*

We thank the reviewer for this point. Reciprocal connections between the visual cortex and the pulvinar are established, but the precise projections to specific pulvinar divisions remain unknown. We have added a note to the Figure 8a caption to clarify this (Figure 7a in the original version).

"Note (panel a): Known pulvinar–cortical connections, as well as sensory input pathways (e.g., visual inputs via the retina/LGN and nociceptive inputs via the spinothalamic tract), are not explicitly shown. These connections are well established anatomically but were omitted due to their heterogeneity and incomplete characterization at the level of pulvinar subnuclei. Their absence should not be interpreted as a lack of anatomical or functional relevance."

> *(5) Stylistic: There are a lot of hypotheses and interpretations presented in this primary literature paper, which may be better suited for a review or perspective piece.*

We thank the reviewer for this comment. We aimed to integrate our empirical findings within a broader conceptual framework to provide a complementary narrative, rather than presenting isolated observations without connecting them to theoretical context. This approach is intended to strengthen the interpretive value of the study while remaining grounded in primary data.

> *(6) In the discussion, there is an assumption that the fMRI BOLD responses to CS+ and CS- need to be different to indicate that an area is processing these distinctly, but the BOLD signal can only detect large-scale changes in overall activity. It's easy to imagine that an area could be involved in processing these two stimuli distinctly without showing an overall difference in the gross amount of activity.*

We thank the reviewer for raising this important point. We fully agree that the fMRI BOLD signal reflects large-scale changes in population activity and may fail to capture more subtle or distributed neural representations. Accordingly, the absence of a CS+ vs. CS− BOLD difference should not be interpreted as evidence that a region is not involved in discriminating these stimuli. Rather, our inferences are limited to differences in aggregate activation at the spatial and temporal resolution of fMRI.

To partially address this limitation, we analyzed anatomically defined thalamic subregions; however, we acknowledge that finer-scale subdivisions and cell-type– specific processing likely exist that are not currently resolvable in human fMRI. Such distinctions may be better investigated using invasive recordings or circuit-level approaches in rodents or non-human primates. This limitation has now been explicitly acknowledged in the Limitations section of the manuscript (page 36):

"Pulvinar divisions, MD, and LGN each contain diverse neuron subtypes and finer anatomical subdivisions that may serve distinct functions. Importantly, the absence of CS+ vs. CS− differences in BOLD activity should not be interpreted as a lack of stimulus-specific processing, as such distinctions may occur without changes in overall activation detectable by fMRI…"

> *(7) There is strong evidence that the BOLD responses to the threat-related and safetyrelated stimuli are different, modest evidence for their claims of learning/plasticity in these pathways, and circumstantial evidence supporting their hypothesized graph network models. Overall, most of the claims made in the discussion are better considered possible interpretations rather than proven findings - this is not a criticism, as these experiments and subject matter are extremely complex.*

We thank the reviewer for this constructive suggestion. In response, we have revised the discussion to present our interpretations as possible or plausible explanations, rather than definitive conclusions, to better reflect the strength of the current evidence. The changes are marked in green throughout the Discussion section.

> *This study continues to validate the power and utility of this in human fear conditioning/extinction paradigm, and extends this paradigm to investigating fear learning beyond the traditional limbic system pathways. It's possible that their models for the pulvinar nuclei interconnections could guide future neuromodulation or DBS studies that could provide more causal evidence for their hypotheses.*
>
> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *The present work was aimed at investigating the specific contributions of thalamic nuclei to associative threat learning and extinction. Using fMRI, they examined activation patterns across pulvinar divisions, the lateral geniculate nucleus (LGN), and the mediodorsal thalamus (MD) during threat acquisition, extinction, and recall. Their goal was to uncover whether distinct thalamic systems support different modes of learningautomatic survival mechanisms versus more deliberate processes - and to propose a hierarchical pulvinar model of fear conditioning. They also try to refine current neuroanatomical models of threat learning and memory, highlighting the role of thalamic nuclei in it.*
>
> *Strengths:*
>
> *(1) Valuable theoretical elaboration and modeling regarding the differential role of pulvinar subdivisions on feedforward (inferior, lateral) and higher-order integration (anterior), and their functional interplay with other relevant subcortical and cortical structures in associative threat and extinction learning.*
>
> *(2) Large sample sizes and multipronged analytical approaches were used for hypothesis testing.*
>
> *(3) Exhaustive literature review in the field of associative threat, as well as regarding the role of thalamic nuclei and other brain structures in it.*
>
> *Weaknesses:*
>
> *(1) Several weaknesses should be pointed out regarding how fMRI data were collected, as well as decisions regarding how the fMRI data were preprocessed and analyzed:*
>
> *(a) fMRI data have low resolution (3 cubic mm), which certainly limits the examination of small nuclei such as the ones investigated here, and especially the examination of the LGN and inferior pulvinar.*

We thank the reviewer for raising this point. While the spatial resolution of fMRI (3 mm isotropic) does limit voxel-wise examination of very small nuclei, our analyses were not performed at the single-voxel level. Instead, signals were extracted using anatomically defined masks for each thalamic nucleus, which is a standard and widely used approach for studying small subcortical structures with fMRI. This strategy increases signal-to-noise ratio and mitigates partial-volume effects by aggregating activity across voxels belonging to the same anatomical region.

> *(b) fMRI was normalized to standard space. Analyzing the data in individual-subject space would have given you the options of avoiding altering every participant's brain*

> *and of using a probabilistic thalamic atlas that better adapts to each subject's brain and thalamic nuclei (see, for instance, Iglesias et al., 2018). This would have been ideal and would have given the authors more precision, especially considering the low resolution of the fMRI data and the size of the thalamic nuclei of interest.*

We thank the reviewer for pointing out the availability of specialized thalamic atlases. In our study we used the Automated Anatomical Labelling Atlas 3 (AAL3 atlas), which includes thalamic subdivisions (including pulvinar and other nuclei) among its 150+ whole-brain regions and is widely used for ROI extraction in normalized fMRI analyses. This choice allowed us to define consistent ROIs across the entire brain such as the amygdala and hippocampus within the same parcellation framework and to extract functional signals at the resolution of our preprocessed fMRI data.

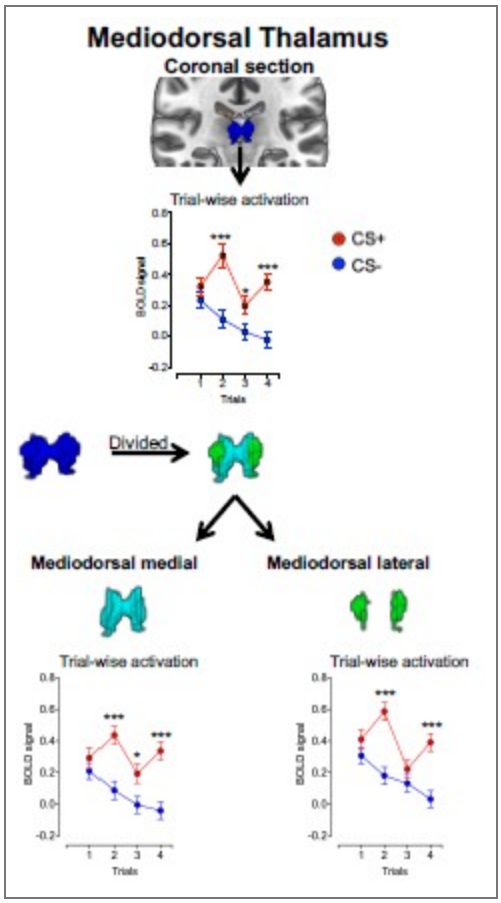
While histology-informed probabilistic atlases offer finer microanatomical segmentation of the thalamus, they are implemented primarily for structural segmentation pipelines (e.g., FreeSurfer) and do not change the fact that AAL3's thalamic subdivisions are established and anatomically reasonable ROIs for functional studies at standard fMRI resolutions. AAL3 thus provides a practical and valid choice for our whole-brain activation and connectivity analyses.

> *(c) On top of the two previous points, the authors decided to smooth the data to 6mm, which means that every single voxel within these small nuclei was blurred/mixed with the 2 immediately contiguous voxels (if they followed the standard SPM12 normalization resampling default which resamples, or upsamples the data in this case, to 2 x 2 x 2mm). Given the strong changes in structural connectivity and function that can occur, especially in the thalamus, on voxels of this size, this and the previous 2 decisions do not favor anatomical precision.*

We thank the reviewer for raising this concern regarding anatomical precision. The data were resampled to 2 × 2 × 2 mm resolution in SPM12, and a 6 mm FWHM Gaussian smoothing kernel was applied. Gaussian smoothing does not uniformly mix immediately adjacent voxels; rather, it applies distance-weighted averaging with a standard deviation of approximately 2.55 mm (FWHM = 2.355σ). At 2 mm resolution, this corresponds to ~1.3 voxels, meaning that signal contribution decreases smoothly with spatial distance rather than reflecting simple voxel averaging. Moreover, all statistical analyses were conducted at the ROI level using anatomically defined masks, rather than voxel-wise inference within nuclei.

To empirically assess whether smoothing may have introduced boundary-driven spillover effects, we divided the mediodorsal (MD) thalamus into medial and lateral divisions and examined the CS effect separately in each. The CS effect did not differ between subdivisions (MD subdivision X CS interaction: $F_{(1, 292)} = 0.50$, $p = 0.48$).

Additionally, across trials, the CS+ vs. CS− effect was observed in both subdivisions and showed comparable magnitudes (see Author response image 1). The effect sizes were also comparable across MD divisions as presented in Author response table 1).



Author response image 1. **Mean activation in MD subdivisions during threat learning**

	MD		MDm		MDl	
CS+ *vs.* CS-	Cohen's d	95% CI [L, U]	Cohen's d	95% CI [L, U]	Cohen's d	95% CI [L, U]
Trial 1	0.07	[-0.04, 0.18]	0.06	[-0.04, 0.18]	0.08	[-0.02, 0.20]
Trial 2	0.28	[0.16, 0.40]	0.26	[0.14, 0.38]	0.32	[0.20, 0.44]
Trial 3	0.13	[0.01, 0.24]	0.14	[0.03, 0.26]	0.07	[-0.04, 0.18]
Trial 4	0.29	[0.18, 0.41]	0.29	[0.17, 0.41]	0.28	[0.16, 0.39]

Note. MD, Mediodorsal thalamus; MDm, Mediodorsal medial thalamus; MDl, Mediodorsal lateral thalamus; L, Lower 95% CI; U, Upper 95% CI; CI, Confidence Intervals.

Author response table 1. **Point estimates and 95% confidence intervals of effect sizes (Cohen's d) for CS+ vs.** CS− contrasts in MD, MDm, and MDl During Early Threat Learning

If smoothing had artificially driven the MD effect via boundary spillover, one would expect consistent asymmetry or substantially larger effects in one subdivision relative to the other. Instead, the CS effect was distributed across both medial and lateral MD, supporting the interpretation that the observed activation reflects intrinsic MD signal rather than smoothing-related contamination.

> *(d) Motion during scanning was poorly controlled in the preprocessing. Including the motion parameters as covariates of no interest in the GLM does not fully guarantee that motion is not influencing the results, and that motion is not differentially influencing some experimental conditions more than others.*

Our analyses are within-subject, so each participant serves as their own control, minimizing the impact of motion differences across conditions. Functional data were preprocessed with fMRIPrep 20.0.2, which estimates motion parameters. The motion estimations are included in the GLM to account for residual motion-related variance in SPM12. The connectivity analyses were conducted in CONN, which also includes these motion parameters as regressors and applies additional denoising steps to further reduce motion-related effects. Together, these procedures make it highly unlikely that motion systematically influenced the observed condition differences.

> *(2) It is not clearly indicated in the manuscript how many subjects and how many trials went into each of the analyses. It would be important to indicate this in the text and/or the figures.*

We thank the reviewer for this important comment. We have now explicitly reported the number of participants and trials contributing to each analysis throughout the manuscript, including the main text, figure captions, and supplementary materials.

Specifically, under Materials and Methods (page 38), we now clarify the sample sizes for each learning phase:

"We analyzed fMRI data from 293 participants during fear conditioning, 320 during extinction, 412 during extinction recall, and 312 during threat renewal."

In addition, all figure captions now report the corresponding sample sizes and trial numbers. For example, the caption to Figure 1 (pages 7–8) states:

"…Block-level comparisons were assessed using paired t-tests, while trial-level effects were examined using a 2 × 2 repeated-measures ANOVA, followed by post hoc comparisons between CS+ and CS− across four trials. Multiple comparisons were controlled using false discovery rate (FDR) correction. Conditioning sample size: n = 293. Detailed statistical parameters are provided in Supplementary Tables 1–2."

> *(3) It is not clear either, why, given the large sample size, some of the results were not conducted using reproducibility strategies such as dividing the sample into 2 or 3 groups or using further cross-validation strategies.*

Cross-validation strategies were applied to the mediation analyses, which are regressionbased and can be sensitive to extreme values or overfitting, ensuring that observed effects generalize beyond the sample. In contrast, the repeated-measures ANOVA tests within-subject condition differences, and is inherently robust to between-subject variability. For these inferential tests, cross-validation or sample-splitting is not typically applied.

However, following the reviewer's recommendation, we conducted a cross-validation analysis focusing on the anterior pulvinar and the mediodorsal thalamus, the primary regions of interest in this study. The full sample ($N = 293$) was randomly divided into three subsamples ($n_1 = 106$, $n_2 = 91$, $n_3 = 96$). For each iteration, we conducted a repeatedmeasures ANOVA (RM-ANOVA) within one subsample and then examined the stability of the CS+ vs. CS− difference in the remaining two subsamples combined. The CS+ vs. CS− difference was statistically significant in most folds for both the mediodorsal thalamus and the anterior pulvinar. Importantly, effect sizes were comparable across folds within each nucleus, indicating stable estimates of the CS effect.

Finally, we observed a comparable pattern of CS+ vs. CS− differences at the trial level in both the mediodorsal thalamus and the anterior pulvinar. Critically, the effect sizes of these differences were stable across most cross-validation folds

> *(4) Limited testing of alternative hypotheses. The results clearly seem to be a selection of the findings supporting the hypotheses that the authors sought to confirm. (just one example: in the analysis reported in Figures 1-2; are there other correlations between the activation of the anterior pulvinar and MD with other pulvinar nuclei? only the MDanterior Puv is reported).*

We thank the reviewer for raising this important point. We would like to clarify that the analyses were not limited to a single, selectively reported association. The relationship between the MD and the anterior pulvinar was evaluated while explicitly accounting for other pulvinar subdivisions, as well as for thalamic input outside the pulvinar.

Specifically, potential contributions from other pulvinar nuclei were controlled by including them in the regression model (Fig. 2 in the manuscript), and the LGN was included as an additional control region. These analyses therefore test whether the MD–anterior pulvinar association is specific, rather than reflecting a more general thalamic or pulvinar-wide effect. With respect to hypothesis testing, the study was explicitly hypothesis-driven, grounded in functional evidence motivating a specific prediction about MD–anterior pulvinar interactions.

Still, in response to the reviewer's suggestion, we further examined pairwise relationships among thalamic subregions. Specifically, we assessed the association between the MD and each pulvinar subdivision using partial correlations, controlling for the remaining pulvinar subdivisions in each analysis. For example, the partial correlation between the MD and the lateral pulvinar was computed while controlling for the activation of the anterior, inferior, and medial pulvinar subdivisions.

The partial correlation between the MD and the anterior pulvinar was consistent across all four trials of threat learning, whereas the other pulvinar subdivisions did not exhibit a consistent pattern. To evaluate the robustness of these effects, we applied a bootstrap procedure (10,000 resamples) to estimate 95% confidence intervals for each partial correlation. As presented in Figure 4b, only the anterior pulvinar–MD association remained robust, with confidence intervals that did not include zero. In contrast, the confidence intervals for most other pulvinar subdivisions included zero, indicating non-robust associations.

> *(5) The manuscript does not contain a limitations subsection. Practically every study has limitations, and this one is not an exception. Better to tell the limitations to the readers upfront so they can factor them into their evaluation of the relevance of the manuscript and reported evidence.*

We thank the reviewer for this constructive suggestion. While the original manuscript already discussed key limitations in the Discussion section (page 36; e.g., "Although distinct thalamic roles in threat learning have been proposed, fMRI data do not fully capture the complexity of this structure…"), we agree that these considerations would benefit from clearer organization and visibility.

To address this point directly, we have now added a dedicated "Limitations and Future Directions" subsection to the manuscript. This subsection explicitly summarizes the principal limitations of the study—including methodological constraints of fMRI and anatomical resolution—and outlines specific avenues for future research to address them. This change makes the limitations more transparent and allows readers to more easily incorporate them into their evaluation of the findings.

> *(6) Data should be made available to the scientific community. Code too. Even if you just used standard fMRI toolboxes, any code used to run analyses will be helpful to the community, or if someone decides to try to replicate your findings.*

We thank the reviewer for this important suggestion and fully agree with the value of data and code sharing for transparency and reproducibility.

The data supporting the findings of this study are derived from a larger, actively used database that is currently involved in ongoing projects. For this reason, the full dataset cannot yet be publicly released. However, the data underlying the reported analyses are available upon reasonable request from the corresponding author, subject to standard data-use agreements.

To facilitate reproducibility, all analysis scripts and pipelines used in this study—including preprocessing and analysis workflows implemented in SPM12, and CONN—are available upon request and can be shared with researchers seeking to replicate or extend the reported findings.

We have clarified this data and code availability statement in the manuscript (page 46).

> *Despite these weaknesses and what can be derived from them, this manuscript constitutes a valuable contribution to the field to start characterizing and conceptualizing the involvement of thalamic nuclei and their interactions with other brain regions in the associative threat learning circuitries. It also paves the road for further testing of the functional dynamics among these regions and circuitries, and modeling testing.*
>
> ***Recommendations for the authors:***
>
> ***Editor's note:***
>
> *Should you choose to revise your manuscript, if you have not already done so, please include full statistical reporting including exact p-values wherever possible alongside the summary statistics (test statistic and df) and, where appropriate, 95% confidence intervals. These should be reported for all key questions and not only when the p-value is less than 0.05 in the main manuscript.*

We thank the editors for this important note. Full statistical reporting, including test statistics, degrees of freedom, exact (raw and corrected) p-values, effect sizes, and 95% confidence intervals, is provided for all key analyses in Supplementary Tables 1–9. In addition, uncertainty estimates and major statistics tests are now explicitly reported throughout the main text, as recommended by the reviewers, irrespective of statistical significance.

During this revision process, we conducted a comprehensive internal consistency check of all reported statistics and figure annotations. We identified and corrected minor discrepancies between some statistical annotations in the figures and the corresponding results reported in the Supplementary Tables. All figures have now been updated to ensure full consistency with the reported analyses. These corrections do not alter the results or conclusions of the study.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) What is the significance of using two different head coils? Were the data comparable from each coil? How did the authors determine this?*

We thank the reviewer for this important question. Data were acquired using two different receiver head coils across participants. Receiver coils primarily influence signal-to-noise ratio

(SNR) and spatial sensitivity profiles, rather than the physiological basis of the BOLD response itself (Triantafyllou et al., 2011).

Importantly, all analyses were based on within-subject contrasts (CS+ vs. CS−), which are robust to global signal scaling differences that may arise from coil sensitivity variations. In addition, standard preprocessing procedures—including intensity normalization, spatial normalization, and nuisance regression—further minimized potential coil-related variability.

To empirically evaluate whether acquisition differences influenced our results, we conducted a repeated-measures ANOVA testing the Trial × CS × Site interaction (where Site reflects acquisition location and associated scanning setup, including receiver coil configuration) during fear conditioning (N = 293). As shown in Author response table 2, none of the thalamic nuclei demonstrated a significant interaction effect, and all effect sizes were negligible ($\eta^2p \leq .01$)

	F	p values	Effect size (η^2_p)
Anterior pulvinar	0.96	0.4643	0.01
Lateral pulvinar	0.695	0.708	0.007
Inferior pulvinar	0.59	0.798	0.006
Medial pulvinar	1.17	0.306	0.01
Mediodorsal thalamus	1.59	0.189	0.005
Lateral geniculate nucleus	0.93	0.497	0.01

Author response table 2. Repeated-Measures ANOVA results for the Trial X CS X site interaction across all relevant thalamic nuclei during fear conditioning.

> *(2) Why were the data smoothed? This could have a negative impact on the specificity of the signals averaged within the pre-defined thalamic ROIs.*

Spatial smoothing was applied to improve signal-to-noise ratio and statistical stability in small, deep thalamic subregions, which are particularly susceptible to noise. We acknowledge that smoothing can reduce spatial specificity. However, our analyses were based on anatomically predefined thalamic ROIs and focused on average activation within each region rather than voxel-wise localization. Under this approach, modest smoothing (i.e., a 6-mm full-width at half-maximum smoothing kernel, rather than the commonly used 8-mm kernel) primarily increases reliability while any signal mixing across adjacent regions would be expected to reduce regional specificity and bias effects toward the null, rather than produce spurious or false-positive differences.

Additionally, we conducted robustness analyses to examine whether spatial smoothing artificially influenced our results. Specifically, we subdivided the mediodorsal thalamus into medial and lateral anatomical regions and compared activation across these subregions. The activation patterns were comparable across both subdivisions, indicating that the observed mediodorsal thalamus effect is unlikely to reflect boundary spillover resulting from smoothing. If smoothing had driven the effect, we would expect differential signal patterns across the subdivisions rather than comparable activation. (See full response to Weakness C, Reviewer 3, as well as Author response image 1 and Author response table 1 in our response).

> *(3) Did the authors consider using any null models to determine whether the observed PPI results could have been observed by chance? E.g., block-resampling nulls scramble temporal order while preserving temporal autocorrelation, and can determine whether*

> *subtle differences in autocorrelation across regions can give rise to the observed signatures.*

We thank the reviewer for this thoughtful suggestion. All PPI analyses were conducted using the default CONN toolbox pipeline. In this framework, PPI effects are estimated within a GLM at the first level following standard denoising procedures that reduce motion- and physiology-related variance and apply temporal filtering. Importantly, PPI effects are modeled as subject-level contrast terms rather than computed from raw timeseries correlations.

Group-level inference was performed on these subject-level contrast estimates using paired t-tests with FDR correction across regions. To further assess whether the observed effects could arise by chance, we additionally performed 10,000 bootstrap resamples of the CS+ vs. CS− differences to evaluate the stability of the effects. While we did not implement explicit block-resampling null models that preserve temporal autocorrelation, the combination of first-level GLM modeling following denoising, large sample size (N ≈ 300), and convergent inferential and resampling procedures provides a rigorous and standard assessment of PPI effects. We have revised the manuscript to clarify these procedures and their rationale.

We added this language to directly address the reviewer's concern and revised the connectivity analyses section to clarify the workflow (page 44):

"Following standard denoising procedures—including regression of motion- and physiology-related confounds and temporal filtering—condition-dependent connectivity effects were inferred from subjectlevel generalized psychophysiological interaction (gPPI) contrast estimates rather than from raw timeseries correlations. This GLM-based framework reduces the likelihood that observed PPI effects reflect differences in temporal autocorrelation or spectral properties across regions rather than genuine task-dependent interactions."

> *(4) The authors may wish to report results in text, as there are currently many demonstrative statements that are not associated with requisite uncertainty estimates, making inference challenging.*

We thank the reviewer for this helpful suggestion. We have revised the Results section to explicitly report statistical outcomes in the main text for all key findings, including appropriate uncertainty estimates (e.g., test statistics, effect sizes, and p-values) alongside demonstrative statements. This ensures that all inferences in the text are directly supported by quantitative evidence.

Additionally, the full statistical details, including test statistics, degrees of freedom, effect sizes, 95% confidence intervals, and both raw and FDR-corrected p-values, are provided in Supplementary Tables 1–9. These changes improve clarity and transparency while avoiding redundancy. Newly added text in the Results section is highlighted in green.

> *(5) I could not find any information about the EBICglasso model in the Methods section, nor information about how the centrality measures were estimated. Given the lack of transparency, I recommend down-weighting the often overly-strong language regarding the conclusions of this analysis.*

We have revised and added these details along with other details to the Statistical tests section on pages 42-44:

"Statistical tests

All statistical tests were conducted using JASP versions 0.18.3 and 0.19.3(JASP Team, 2024).

Activation Differences across all phases of threat learning

In each threat learning phase, we used paired t-tests to examen the differences in activation of the thalamic nuclei in response to CS+ vs. CS- at the block level (average activation across trials), and 2x2 RM-ANOVA to estimate the differences in activation at the trial-wise level. Assumptions of sphericity were checked, and Greenhouse-Geisser corrections were applied where necessary. This model was followed by post hoc tests to estimate the differences at the trial level and False discovery rate (FDR) correction was applied for each question.

Network analyses of the within pulvinar relationships during conditioning

The network analyses examined functional relationships between pulvinar divisions. Nodes corresponded to block-level activation estimates of the CS+ minus CS− contrast for each pulvinar division, yielding four nodes (one per division). Networks were estimated using a Gaussian graphical model with EBICglasso (LASSO regularization) based on Pearson correlation matrices, with the EBIC tuning parameter set to $\gamma = 0.5$. Edge weights represent partial correlations.

Three centrality measures were computed on the estimated weighted partial-correlation network: node strength, defined as the sum of the absolute edge weights directly connected to a node; closeness, defined as the inverse of the average shortest path length from a node to all other nodes; and betweenness, defined as the proportion of shortest paths between all pairs of nodes that pass through a given node. Shortest paths were computed using inverse edge weights, consistent with standard practice for weighted networks. Centrality indices were normalized.

Network accuracy and centrality stability were assessed using nonparametric bootstrapping (10,000 iterations) to estimate confidence intervals for edge weights and centrality measures. All analyses were conducted in JASP (versions 0.18.3 and 0.19.3) using default settings unless otherwise specified, following the procedures described in Epskamp, Borsboom, and Fried (2018).

Mediation analyses of within pulvinar relationships during conditioning

Mediation models of the relationships between the activations in pulvinar divisions were estimated using the lavaan package (Rosseel, 2012) with maximum likelihood estimation. All variables were zstandardized prior to analysis. Block-level activation estimates from the inferior and lateral pulvinar were entered as predictors, activation in the medial pulvinar was specified as the mediator, and activation in the anterior pulvinar was specified as the outcome variable.

To assess the robustness and generalizability of the mediation effects, we conducted 3-fold crossvalidation. The full sample (N = 293) was randomly partitioned into three non-overlapping sub-samples (n = 91, 96, and 106). In each iteration, the mediation model was estimated in one sub-sample, while the remaining sub-samples were used to assess the stability of parameter estimates and indirect effects. This procedure resulted in six cross-validation iterations, allowing evaluation of whether the direction and magnitude of the indirect effect were consistent across independent subsets of the data. Mediation models were estimated using the lavaan package (Rosseel, 2012) with maximum likelihood estimation. Indirect effects were evaluated using bias-corrected percentile bootstrap confidence intervals based on 10,000 resamples, as recommended by Biesanz, Falk, and Savalei (2010). An indirect effect was considered significant when the 95% confidence interval did not include zero ($p < 0.05$).”

> *(6) Bar plots are not effective ways to report group-level data. I recommend replacing all bar plots with visualisations that expose the distribution of the data, such as a violin plot or a raincloud plot.*

We thank the reviewer for this suggestion. In general, we agree that visualizations exposing the full data distribution can be highly informative, and we therefore present distribution-based plots for several analyses (e.g., connectivity results). However, for the activation analyses, our primary goal was to highlight trial-to-trial changes and overall patterns across conditions, rather than the distribution of individual data points per se. For this purpose, bar plots provide a clearer representation of the directional effects and facilitate comparison across trials and conditions.

> *(7) The thought bubbles are atypical of scientific figures.*

The figure has been revised to remove the thought bubbles.

> *(8) Figure 7 - there are many connections not shown in this figure, suggesting that it is sufficiently oversimplified as to be potentially misleading. For instance, the authors offer no anatomical connections between pulvinar and the cortical hierarchy; however, these connections are ample and (likely) highly important for the functionality assessed here. Similarly, there is no room in the figure for the integration of the shock stimuli (presumably via the spinothalamic tract) and the visual stimuli (via the retina/LGn).*

We agree that the pulvinar has extensive cortical and sensory input/output connections that are not depicted in Figure 7. Our intention was not to provide a complete anatomical wiring diagram, but rather a simplified functional model derived from observed statistical dependencies. We have revised the figure and added an explicit note to the legend clarifying that pulvinar–cortical and sensory pathways (e.g., retina/LGN and spinothalamic inputs) are intentionally omitted due to incomplete subnuclear-level anatomical characterization, and that their omission should not be interpreted as a lack of importance. We added this to Figure 7 legend:

“Note (panel a):

Known pulvinar–cortical connections, as well as sensory input pathways (e.g., visual inputs via the retina/LGN and nociceptive inputs via the spinothalamic tract), are not explicitly shown. These connections are well established anatomically but were omitted due to their heterogeneity and incomplete characterization at the level of pulvinar subnuclei. Their absence should not be interpreted as a lack of anatomical or functional relevance.”

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) It's somewhat confusing that Figures 1,4,5 D and E are not in the text until later in the results section. Perhaps these should be presented in the figures in the same order they are discussed in the text, although this is a stylistic issue.*

We thank the reviewer for this comment. To improve clarity and align the figures with the structure of the Results section, we reorganized the figures. Specifically, we added a new figure (Figure 7) that consolidates all connectivity analyses. Figures 1, 4, and 5 now focus exclusively on activation results, while Figure 7 presents connectivity results only. This reorganization allows the figures to follow the flow of the text more closely and makes the narrative of each figure clearer.

> *(2) Stylistic: I would strongly recommend adding n numbers and describing the basics of statistical tests used and how multiple comparisons were accounted for in the legend for Figures 1,4, and 5.*

We thank the reviewer for this recommendation. We have added the sample sizes (n) and brief descriptions of the statistical tests used, including how multiple comparisons were handled, to the legends of Figures 1, 4, and 5. In addition, we direct the reader to the Supplementary Tables, which were submitted with the original manuscript and provide full

statistical details, including test statistics (t, F), degrees of freedom, effect sizes, 95% confidence intervals, raw p values, and corrected p values. Finally, we further elaborated on the statistical tests on pages 42–44, as detailed in our response to Recommendation 5 (Reviewer 1).

> ***Reviewer #3 (Recommendations for the authors):***
>
> *As previously indicated, please note that no information is included in the manuscript about data and code availability. Although you mainly use toolboxes for data analyses, any script(s) that you have used to run things would be great to upload for reproducibility purposes.*
>
> *Also, it would be good to include a limitations subsection in the manuscript.*

Thank you for these recommendations. We added limitations subsection to the manuscript. See our responses under Comments 5 and 6 (Reviewer 3, Public Review).

> *In terms of data analyses:*
>
> *(1) It would be ideal if you quantify in-scanner motion for the different conditions to see if there were no differences in motion due to the task.*

Head motion was estimated at each time point as part of standard preprocessing, and motion parameters were included as nuisance regressors in all first-level models. Because motion estimates are defined per volume rather than per experimental condition, condition-specific motion metrics were not explicitly computed. Importantly, this approach removes motion-related variance uniformly across the time series and therefore controls for potential motion effects across all task conditions. Any residual motion would be expected to increase noise rather than systematically bias condition contrasts.

> *(2) You also may want to indicate if normalization followed the SPM 12 default and the data was resampled to 2 x 2 x 2 mm, or kept the same. It is not stated in the data preprocessing subsection of the methods.*

We thank the reviewer for this suggestion. We have now clarified this point in the manuscript (page 41):

“In addition, spatial normalization was performed with data normalized to Montreal Neurological Institute (MNI) space and resampled to a $2 \times 2 \times 2$ mm^3 voxel grid, followed by spatial smoothing with a 6-mm full-width at half-maximum Gaussian kernel.”

> *(3) It is important to indicate how many subjects went into each analysis. Also, it is not clear, based on the current methods section, how many observations per condition were used. That can be reported in the text or the figures.*

We thank the reviewer for this comment. This information has now been added to the Methods section and the relevant figure legends, as described in our response to Comment 2 (Reviewer 3, Public Review).

References

Triantafyllou C, Polimeni JR, Wald LL. 2011. Physiological noise and signal-to-noise ratio in fMRI with multi-channel array coils. NeuroImage 55:597–606. DOI: *https://doi.org/10.1016/j.neuroimage.2010.11.084*, PMID: 21167946

https://doi.org/10.7554/eLife.108043.2.sa0