nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-022-01082-w

Check for updates

Common and stimulus-type-specific brain representations of negative affect

Marta Čeko[1]✉, Philip A. Kragel[1,2], Choong-Wan Woo[3,4,5], Marina López-Solà[6] and Tor D. Wager[1,7]✉

The brain contains both generalized and stimulus-type-specific representations of aversive events, but models of how these are integrated and related to subjective experience are lacking. We combined functional magnetic resonance imaging with predictive modeling to identify representations of generalized (common) and stimulus-type-specific negative affect across mechanical pain, thermal pain, aversive sounds and aversive images of four intensity levels each. This allowed us to examine how generalized and stimulus-specific representations jointly contribute to aversive experience. Stimulus-type-specific negative affect was largely encoded in early sensory pathways, whereas generalized negative affect was encoded in a distributed set of midline, forebrain, insular and somatosensory regions. All models specifically predicted negative affect rather than general salience or arousal and accurately predicted negative affect in independent samples, demonstrating robustness and generalizability. Common and stimulus-type-specific models were jointly important for predicting subjective experience. Together, these findings offer an integrated account of how negative affect is constructed in the brain and provide predictive neuromarkers for future studies.

Affect is a fundamental property of brain function. The hedonic quality and motivational relevance of sensory stimuli govern the strength of brain responses to sensory cues and drive learning[1,2]. Much attention has been devoted to understanding how affect influences behavior and is disrupted in psychopathology and neurological disorders, but less is known about the neural structure of affective processes themselves—how they are represented in the brain and whether they converge on generalized (common) representations of value.

Affective experiences are often defined in terms of the 'core' dimensions valence and arousal[3,4] or approach-avoidance tendencies[5], implicitly assuming a level of interchangeability among stimulus types. Neuroeconomic theories postulate a 'common currency' for value[6,7], whereby signals from diverse reinforcers are integrated into a common representation that shapes decision-making and behavior. These ideas have shaped clinical research. For example, emotional facial expressions are commonly used as probes of negative affect across clinical conditions[8,9]. Likewise, pain neuroimaging has concentrated on a few types of stimuli, most commonly heat, as probes of pain sensitivity in general[10].

If different types of affective stimuli can be used interchangeably, any aversive stimulus might be suitable for probing 'negative affect' systems (for example, as defined by the National Institutes of Health (NIH) Research Domain Criteria[11]). If they cannot, important basic and clinical effects could be missed, for example, if a stimulus type used is not relevant for the effect or population studied. Theories of affect and computational accounts of learning, predictive coding and active inference might need to be extended to account for reinforcer-specific and stimulus-type-specific brain processes[12].

Evidence for shared neural representations is mixed. On one hand, animal studies have identified cross-modal coding of affective information in single neurons[7,13], and human functional magnetic resonance imaging (fMRI) studies have identified commonalities across aversive stimulus types in ventromedial and orbital prefrontal cortices (vMPFC and OFC, respectively), anterior midcingulate cortex (aMCC) and anterior insula (aINS)[14–16]. Meta-analyses have identified potential neural substrates for 'core' affective dimensions (valence and arousal) in the OFC, MPFC, nucleus accumbens (NAC)/ventral striatum (vStr) and amygdala[17,18]. On the other hand, several theories suggest that the brain is organized into separable neural processes for different types of negative affect[19,20]. Although these are labeled as 'negative' or 'aversive', these labels may be culturally constructed categories rather than fundamental properties of brain organization[21]. fMRI studies have identified distinct activity patterns for different categories of emotion and affective stimulus types[16,22–27] (although these may also reflect constructed categories[21]). Animal studies have identified neuronal ensembles specific to distinct types and aspects of nociception[28–30] and affective states like thirst[31–33].

Evidence for common and stimulus-type-specific representations of negative affect have largely been investigated separately, in different studies and paradigms, rather than comparing them directly. The latter was the goal of the current study.

In study 1, we measured fMRI responses to four types of aversive stimuli—painful heat, painful pressure, aversive images and aversive sounds—and a positive affective control (pleasant images) in $N = 55$ healthy participants (Fig. 1a). Including four intensity levels per type and subjective post-trial ratings on a common rating scale allowed us to identify fMRI patterns that tracked subjective aversiveness across all stimulus types ('common negative affect') or in a stimulus-type-specific fashion. Partial least squares regression (PLS-R) provided a framework for jointly estimating both common and stimulus-type-specific representations[34,35] (Supplementary

[1]Institute of Cognitive Science, University of Colorado, Boulder, CO, USA. [2]Department of Psychology, Emory University, Atlanta, GA, USA. [3]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea. [4]Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea. [5]Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, South Korea. [6]Serra Hunter Programme, Department of Medicine, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain. [7]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. ✉e-mail: marta.ceko@colorado.edu; tor.d.wager@dartmouth.edu

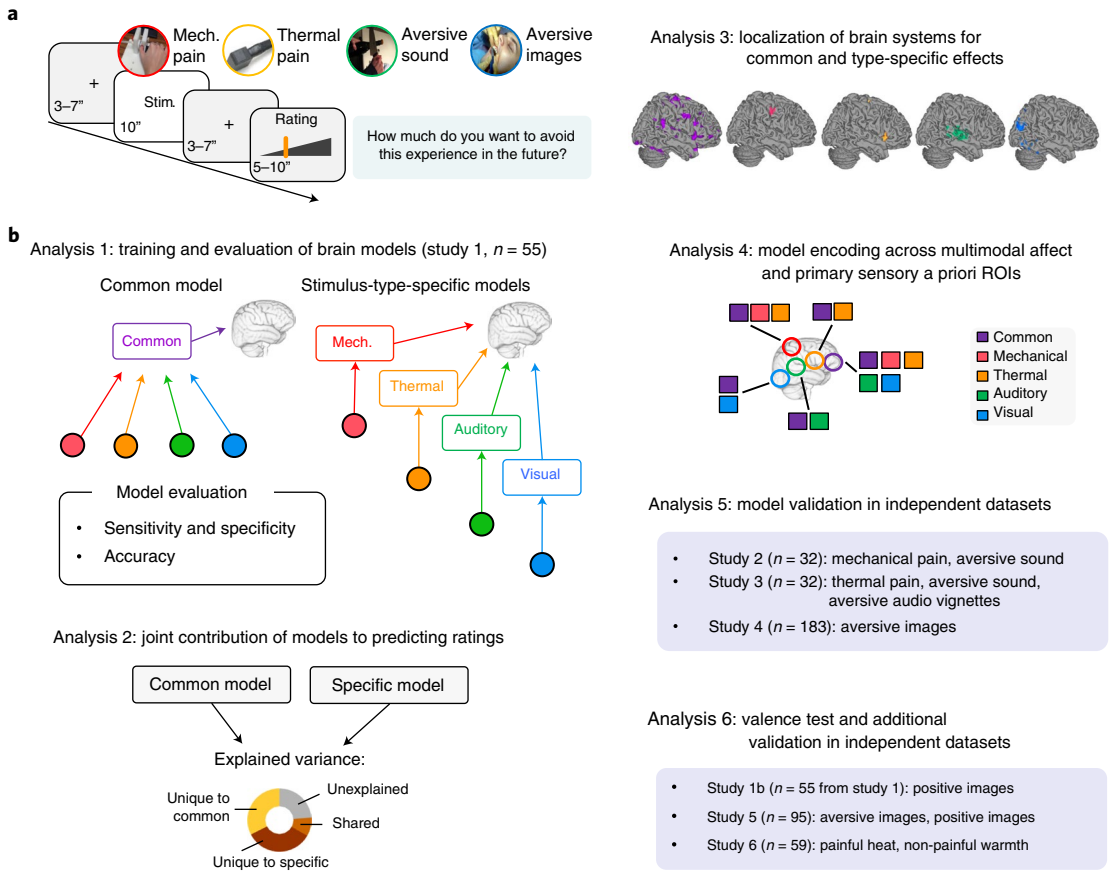


Fig. 1 | Task design and main analyses. a, Multiple aversive experiences task. Four stimulus types (thermal and mechanical pain, aversive sounds and aversive images) at four preselected intensity levels for a total of 96 randomized stimuli over six fMRI runs; aversiveness ('negative affect') rated on a common scale after each stimulus presentation, by measuring 'how much do you want to avoid this experience in the future?'. **b**, Main analyses: (1) brain model development using PLS-R, model evaluation (sensitivity and specificity); (2) variance decomposition analysis to test how much of the predicted variance in ratings is uniquely attributable to each model versus shared among models; (3) identification of core systems for common (purple) and stimulus-type-specific negative affect; (4) analysis of model encoding in selected multimodal and primary sensory ROIs; (5) validation in independent datasets; and (6) valence tests using positive stimuli and additional validation analyses.

Fig. 1), and yielded predictive models whose sensitivity and specificity for negative affect could be quantitatively evaluated in new individuals. In several prospective samples (studies 2–6; $N=401$ participants), we tested the final models to evaluate their predictive accuracy (replicability), generalizability across studies and generalizability across stimulus types versus specificity to a particular type. This framework allowed us to test: (1) whether there is a neural representation of common 'negative affect' across stimulus types that predicts the degree of negative affect experienced in response to any stimulus; (2) whether there are also stimulus-type-specific representations of negative affect; (3) whether these representations are specific to aversive stimuli or also respond to positive stimuli, which could signal encoding of arousal or salience; and (4) the relative importance of common and type-specific representations in jointly predicting negative affect ratings in new individuals, assessing their utility as neuromarkers in future studies[36].

Results

Behavioral results. In study 1, $N=55$ participants (24 females) experienced four types of aversive stimuli, during fMRI at 3T (Methods). Negative affect induced by these stimuli was measured on a uniform scale across stimulus types, allowing direct comparisons across sensory inputs despite variation in stimulus properties. Each stimulus type was rated as moderately to strongly aversive across the four intensity levels, ranging from 0.18 ('moderate') to 0.37 ('strong', general Label Magnitude Scale (gLMS); Supplementary Table 1). Negative affect ratings increased with intensity (Fig. 2a) for mechanical pain ($t(54)=7.68$, $P<0.001$), thermal pain ($t(54)=13.86$, $P<0.001$), aversive sounds ($t(54)=6.47$, $P<0.001$) and aversive images ($t(54)=11.42$, $P<0.001$). Stimulus types were comparably aversive, permitting analysis of variation in brain activity across types while approximately matching on (and statistically controlling for) reported negative affect, and individual differences in sensitivity were correlated across types ($r=0.26$–0.68; Supplementary Table 2). Ratings for each stimulus level (averaged across trials) within participants were used as outcomes for brain model development.

Identifying common and stimulus-type-specific brain models. Negative affect models were developed using PLS-R on study 1 data and were constrained to simultaneously predict common and stimulus-type-specific effects based on brain-wide patterns within gray matter ('PLS-R for brain model development'). This produced five multivariate patterns: one for each of the four stimulus types and one for common negative affect. Models were tested using fivefold leave-whole-participant-out cross-validation. Thus, accuracy statistics are based on models trained on other participants ('Model evaluation') and additionally validated in independent studies (see below).

These models provided evidence for both common and stimulus-type-specific coding of negative affect. All five models were sensitive to their target stimuli, as evidenced by significant

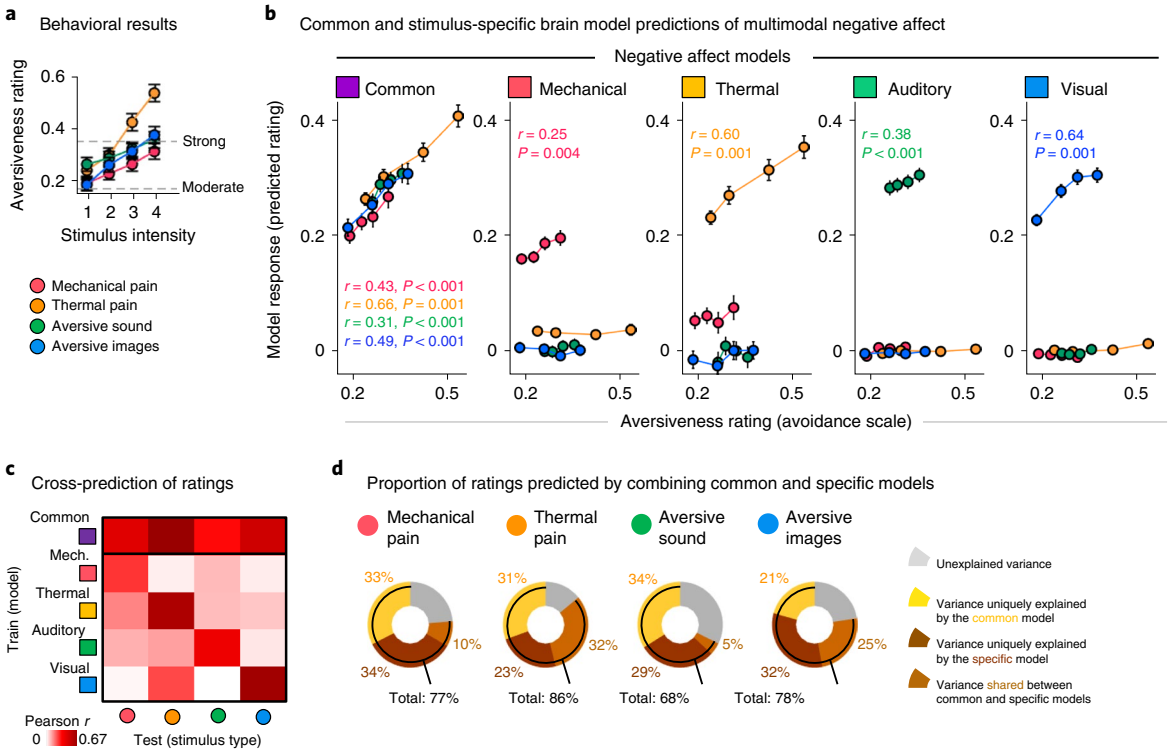


Fig. 2 | Model evaluation and joint contributions to predicting negative affect. a, Ratings significantly scaled with stimulus intensity ($N = 55$ participants; two-sided $P < 0.001$ for each stimulus type; linear regression). Data are shown as mean values across participants for each stimulus type. Error bars reflect within-participant s.e.m. **b**, Relationship between observed and predicted ratings (that is, model response). Data are shown as mean values across participants for each stimulus type. Error bars reflect within-participant s.e.m. The common model, trained on all stimuli, significantly predicted ratings to each stimulus type. Stimulus-type-specific models, optimized for specificity by setting other stimulus types at 0 during training, significantly predicted ratings to target (color-matched) stimulus type, but not to off-target stimulus types. *r*, mean within-participant Pearson correlation between predicted and observed ratings; two-sided *P* values based on a 10,000 samples bootstrap test of within-participant *r* values. **c**, Cross-prediction of ratings across stimulus types tested by Pearson correlation between the predicted and the observed outcomes for each train-test stimulus pair; darker shading indicates a higher Pearson *r* value**. d**, For each stimulus type, the within-participant variance in outcome (ratings) explained by the predictors (common model, specific model) is partitioned between the predictors into unique and shared components by computing the full and reduced regression models: total r^2, mean within-participant total variance explained by common and specific models, unique r^2 for common model = total r^2 – single variance for specific model (yellow slice), unique r^2 for specific model = total r^2 – single variance for common model; dark-brown slice, shared r^2 = total r^2 – unique r^2 specific – unique r^2 common (light-brown slice).

associations between observed and predicted ratings (Fig. 2b and Supplementary Table 3). Below we report correlations, expressed as mean within-participant $r \pm$ standard error (s.e.), and the out-of-sample prediction root mean squared error (RMSE; rating scale ranged from 0 to 1) for each model and outcome[37]. The common model accurately predicted mechanical pain ($r = 0.43 \pm 0.07$, $P < 0.001$, RMSE = 0.132), thermal pain ($r = 0.66 \pm 0.06$, $P = 0.001$, RMSE = 0.192), aversive sounds ($r = 0.31 \pm 0.07$, $P < 0.001$, RMSE = 0.176) and aversive images ($r = 0.49 \pm 0.06$, $P < 0.001$, RMSE = 0.179). Some variation in correlation values likely reflects range differences across outcomes, but the RMSE values were similar.

Each type-specific model predicted negative affect ratings of the intended type: $r = 0.25 \pm 0.08$, $P = 0.004$, RMSE = 0.156 for mechanical pain; $r = 0.60 \pm 0.06$, $P = 0.001$, RMSE = 0.205 for thermal pain; $r = 0.38 \pm 0.06$, $P < 0.001$, RMSE = 0.185 for aversive sounds; and $r = 0.64 \pm 0.05$, $P = 0.001$, RMSE = 0.172 for aversive images. Type-specific models were specific to target stimuli, as evidenced by poor cross-prediction of off-target ratings (Fig. 2b,c and Supplementary Table 3), which were not significantly different from chance, with two exceptions (Supplementary Table 3). Effect sizes were on average five times lower than for on-target predictions. All five brain models remained significant after controlling for session order ('Study design'), age and sex, and these variables had little impact on ratings (Supplementary Table 4). The brain models also discriminated between the lowest and highest stimulus level (see Supplementary Fig. 2 for classification accuracy matrix across all stimulus intensity-level pairs).

Relative contributions of models to predicting ratings. Next, we asked how common and type-specific brain models combine to predict negative affect. We partitioned the variance explained in ratings into that: (a) unique to the common model, (b) unique to the specific model, and (c) shared across both models. Variance decomposition was computed via full and reduced multiple regression models[38] on cross-validated model outputs for each individual participant ('Variance decomposition analysis') and then averaged across participants.

Negative affect ratings of all types were predicted by a mixture of common and stimulus-type-specific representations in approximately equal parts (mechanical: 33% common, 34% specific; thermal: 31% common, 23% specific; auditory: 34% common, 29% specific; visual: 21% common, 32%, specific) and some variance shared across the common and specific models (mechanical, 10%; thermal, 32%; auditory, 5%; visual, 25%; Fig. 2c). Because negative affect was predicted by a mix of common and stimulus-type-specific representations, these results show that negative affective

experience is not completely reducible to a common dimension such as valence.

Core systems for multimodal and stimulus-type-specific negative affect. To identify important brain features, we interpret both model weights and model encoding maps (or 'structure coefficients'[39]; Supplementary Fig. 3). Consistent model weights in bootstrap tests identify important voxels contingent on other features in the multivariate model. Structure coefficients identify voxels individually associated with each model's output, mapping individual voxels to the overall multivariate model prediction[39]. The voxels significant in both maps (that is, their conjunction) are the most consistently associated with the target outcome, with or without other brain covariates, and can be interpreted as core regions contributing to the predictive model. For type-specific models, we calculated a three-way conjunction to additionally require that identified voxels correlate more strongly with model predictions for the target stimulus than for any other type. Thus, a core stimulus-type-specific (selective) system has voxels that reliably contribute to prediction, encode the respective model and are selective for the target model above all other models.

The core system for common negative affect included OFC, MPFC, MCC, aINS, vStr and amygdala (Fig. 3a and Supplementary Table 5). These regions are broadly associated with affect, motivation and multisensory integration, and encode common properties across multiple negative affective stimuli in previous studies[14–16,18,40,41]. Some voxels in this system also contributed to stimulus-type-specific models, consistent with intermixed neural populations identified in animal studies[2,42]. The MCC, vStr and aINS also included thermal pain-selective voxels (Fig. 4), and the amygdala included visual negative affect-selective voxels.

Core stimulus-type-selective systems (Fig. 3b and Supplementary Table 6) largely mapped onto early sensory pathways cortices, with several exceptions. Bilateral primary somatosensory cortex (S1) was the only area selective for mechanical pain. Portions of right secondary somatosensory cortex (S2) and dorsoposterior insula (dpINS) were selective for thermal pain. Ventroposterior insula (vpINS) and bilateral auditory cortices (areas A1–A3) were selective for auditory negative affect. Bilateral visual cortices (areas V1–V4) were selective for visual negative affect. These findings are in line with recent meta-analyses showing modality-specific processing of aversive input in early sensory cortices[18,40,41]. Importantly, however, these areas were not only selectively activated by a particular stimulus type, but also selectively predicted negative affect ratings for one type.

An extended set of regions was selective for visual negative affect, including anterior occipitotemporal and parahippocampal areas, amygdala and mid-lateral OFC (Fig. 3b). In addition, some 'sensory' thalamic and early sensory cortices also encoded the common model (Figs. 3a and 4), supporting an expanded role for traditionally 'sensory' areas in processing of multimodal affective input[27,43–45].

Individual model contributions to negative affect in regions of interest. To probe local contributions to each model, we examined model encoding in a set of a priori regions of interest (ROIs) previously linked to affective processes and sensory-specific regions (Methods and Supplementary Table 7). All ROIs contained voxels with significant structure coefficients in at least two models (Fig. 3c,d). Each multimodal ROI encoded the common and at least one stimulus-type-specific model; Fig. 3c). Sensory ROIs were largely stimulus-type specific, as shown in Fig. 3d, with some exceptions. Mechanical and thermal pain were encoded in S1 and S2/dpINS cortices. S2/dpINS also encoded auditory negative affect, possibly reflecting their proximity to auditory cortices (Fig. 4 and Discussion). Auditory cortices (A1–A3) encoded auditory negative affect and thermal pain. Visual cortices (V1–V4) primarily encoded visual negative affect, but also auditory negative affect and pain. Each sensory cortex also contributed to encoding the common negative affect model.

For each region, we calculated the ratio of importance for common affect versus stimulus-type-specific negative affect, where importance was defined as the percentages of the ROI encoded by each of the common and predominant type-specific models. A negative correlation in the two importance scores (Fig. 3e) indicated a trade-off between them. Early sensory ROIs were more type selective, with auditory regions (A1–A3) the most selective (that is, highest type-specific/common ratio; Fig. 3d). Visual cortices (V1–V4) were also type selective. Somatosensory cortices (S1 and S2) and dpINS were more mixed, showing overlap with mechanical and thermal pain, auditory and common models.

Among multimodal affect/motivation-related regions, vMPFC, ventrolateral prefrontal cortex (vLPFC) and dorsomedial prefrontal cortex (dMPFC) showed the highest relative importance for the common model. The MCC and vStr were preferentially important for the common and thermal models. The amygdala was preferentially important for the common and visual models. The aINS and vLPFC/OFC were more mixed, with a relatively low common/type-specific ratio. Voxels in these regions encoded mechanical and thermal pain, auditory and common negative affect models. Overall, anatomical regions encoded distinct combinations of common and type-specific affect. No region encoded negative affect in a purely domain-general fashion.

Organization of regions into sensory and cortico-brainstem pathways. Selectivity in early sensory cortices (for example, V1 and A1; ref. [18]) suggested that negative affect may be encoded along related ascending thalamocortical and brainstem sensory pathways as well. To test this, we examined model encoding within an expanded set of thalamic and brainstem regions (Fig. 4). These included: (1) superior colliculus (SC), pulvinar, and lateral geniculate nucleus (LGN), which project to V1 and mediate visual perception; (2) inferior colliculus (IC) and medial geniculate nucleus (MGN), which form an auditory pathway projecting to A1; (3) somatosensory/pain-related thalamocortical pathways including ventroposterior lateral (VPL) and intralaminar (IL) thalamic nuclei; and (4) midbrain rostroventral medulla (RVM) and periaqueductal gray (PAG) and thalamic mediodorsal (MD) nucleus, which (along with IL) are thought to be related to multiple forms of negative affect.

Visual model encoding was indeed reflected in the SC and LGN, along with interconnected V1 and amygdala, indicating encoding of negative affect related to visual images in the classic LGN–V1 and pulvinar–V1 thalamocortical pathways as well as the SC–pulvinar–amygdala subcortical visual pathways[46,47]. Likewise, auditory negative affect was encoded in an RVM–PAG–IC–MGN–A1 pathway[46,48], with additional involvement of S2 and posterior insula (pINS) proximal to the auditory cortex. IC selectively encoded auditory negative affect, and SC selectively encoded visual negative affect.

Thermal and mechanical pain were encoded in differentiable pathways. Thermal pain was encoded in an RVM–PAG–midline thalamic (MD and IL) pathway connected to MCC and insula (aINS and mid-insula (mINS)) in the cortex and vStr in the forebrain[49]. Mechanical pain was more strongly related to an S2–mINS pathway (in addition to mechanical pain-specific portions of S1 described above). pINS also showed a fine-grained distinction, with dorsal pINS contributing to pain and ventral pINS contributing to auditory negative affect[25,43]. We also observed a dorsoventral distinction in S1 and surrounding areas, with dorsal S1 selective for thermal pain and ventral S1 selective for mechanical pain (Fig. 4). These types of pain are often used interchangeably, although they behave differently in different animal strains[50] and clinical populations (for example, ref. [51]).

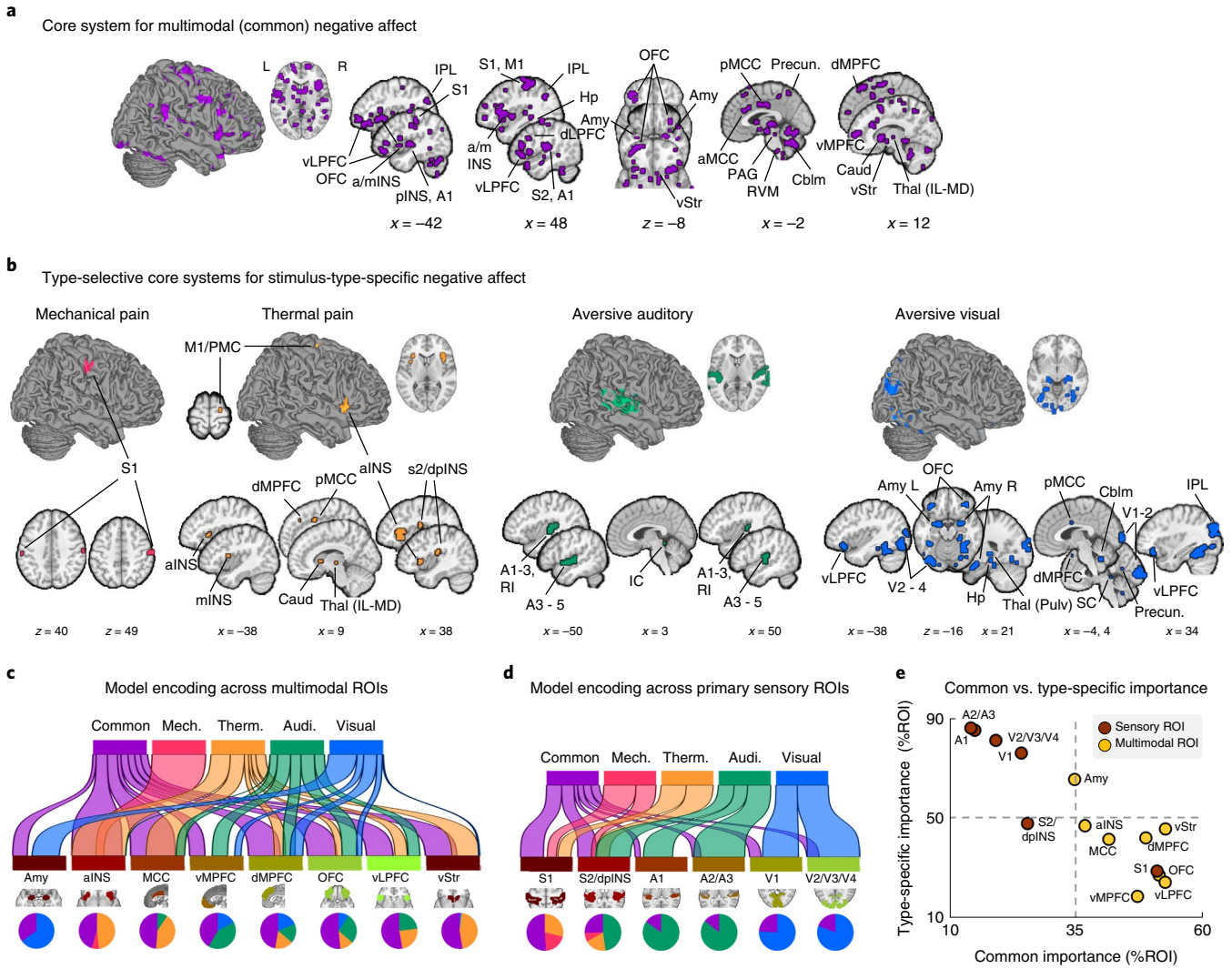


Fig. 3 | Core brain systems for multimodal and stimulus-type-specific negative affect. a, Core system for common processes; defined as the conjunction (retaining positive values) of the PLS weight map and the PLS model encoding (structure coefficient) map, each thresholded at a false discovery rate (FDR) of $q < 0.05$ and retaining positive values, for the common model. **b**, Core type-selective systems; per stimulus type, conjunction of PLS model encoding (structure coefficient) map and the type-selective PLS model encoding map (each thresholded at FDR $q < 0.05$; showing voxels where model encoding values were significantly greater in that model than in any of the other four models). **c**,**d**, River plots show spatial similarity (computed as cosine similarity) between model encoding maps and anatomical parcellations (ROIs), documented in previous neuroimaging studies to show activation to aversive stimuli across stimulus types or preferential activation to individual stimulus types. Ribbons are normalized by the max cosine similarity across all ROIs. Each predictive model is shown in a different color. Maps were thresholded at FDR $q < 0.05$ and positive voxels were retained only for similarity calculation and interpretation. Ribbon locations in relation to the boxes are arbitrary. Pie charts show relative contributions of each model to each ROI (that is, percentage of voxels with highest cosine similarity for each predictive map). **e**, Regional importance scores for common versus specific models. *x* axis, common model importance (percentage of ROI occupied by the common model); *y* axis, type-specific model importance (specific model with highest percentage); S1, primary somatosensory cortex; M1, primary motor cortex; A1, primary auditory cortex; A2/A3, intermediate auditory cortex; V1, primary visual cortex; V2/V3/V4, intermediate visual cortex; Amy; amygdala; Hp, hippocampus; aINS, anterior insula; RI, retroinsular cortex; dLPFC, dorsolateral prefrontal cortex; IPL, inferior parietal lobule; Precun, precuneus; Caud, caudate; Thal, thalamus; Cblm, cerebellum.

Interestingly, common negative affect was encoded in traditionally pain-related neural pathways, from brainstem (RVM–PAG) via thalamic (MD, IL and VPL) nuclei to somatosensory cortices and multimodal cortical and forebrain structures (for example, mINS, aINS and posterior midcingulate cortex (pMCC); Fig. 4). Some regions (MD, IL and PAG) play roles in multiple types of affective behavior[52,53], whereas others are thought to be more specific to somatic processing (S2, VPL and RVM)[49,54] but may reflect affect-related brain–body communication.

Unexpectedly, thermal pain was weakly encoded in RVM and PAG, which are traditionally pain related[49], and strong encoding of visual negative affect in the auditory MGN. This could be due to current limitations of fMRI at this resolution (Discussion), or to complex roles of RVM and PAG in pain representation and modulation.

Prediction of negative affect ratings in new samples. We tested the replicability, robustness and generalizability of the models developed in study 1 by testing their sensitivity and specificity in predicting negative affect ratings in 401 new participants from different cohorts[15,55] (studies 2–6; see Figs. 5 and 6 and Supplementary Table 8 for experimental designs and stimulus types). Each model

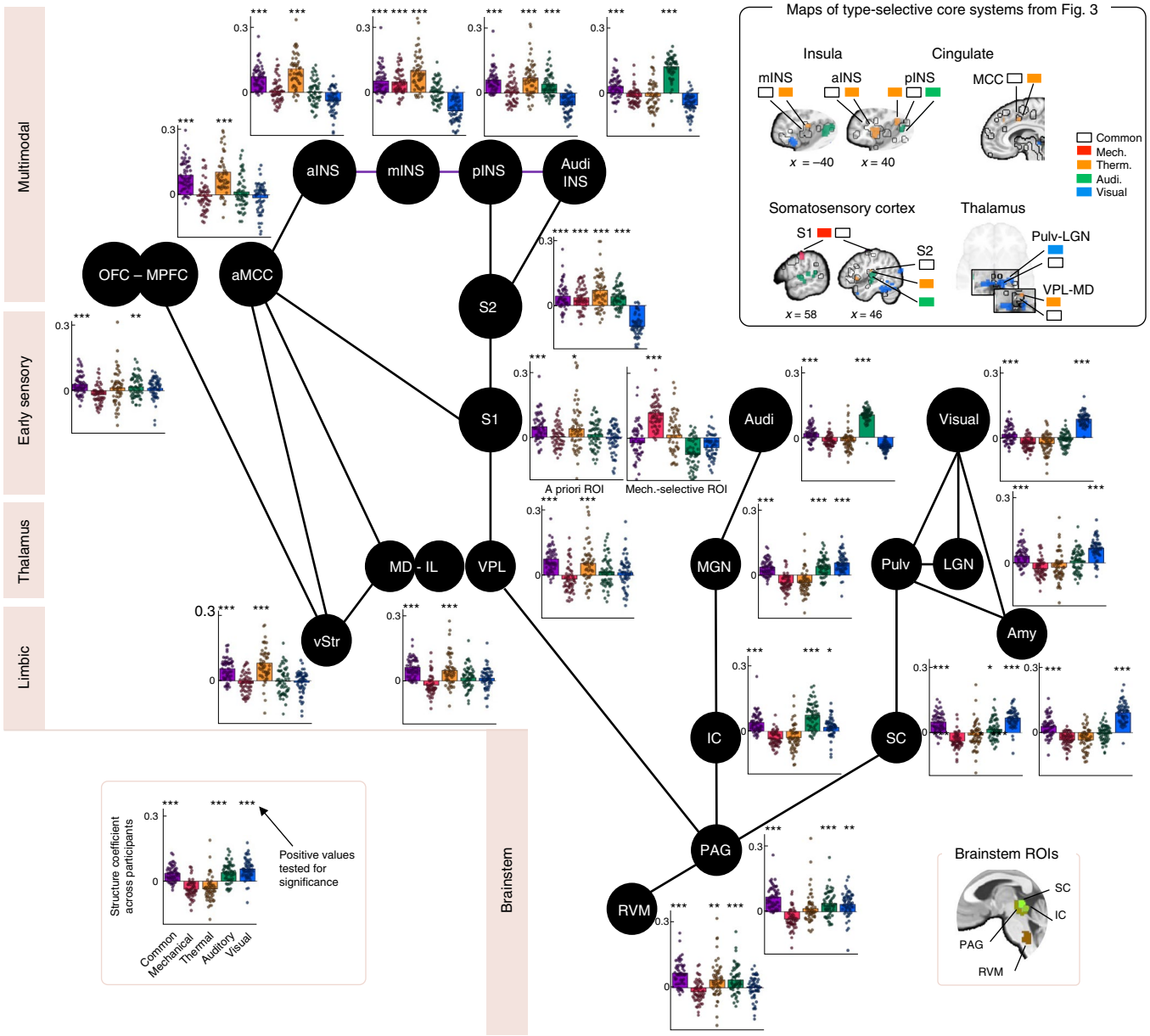


Fig. 4 | The proposed neural architecture of multimodal negative affect. Anatomical ROIs were selected a priori (Supplementary Table 7), with the exception of the mechanical pain-selective S1 ROI, which was localized in the current study (Fig. 3). Plots show structure coefficients for each model. Structure coefficients were averaged across voxels within ROI tested for significance for each model and across participants ($N=55$); one-sample t-test treating subject as random effect; $^{*}P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$ (for exact P values and the associated t-statistics, see Supplementary Table 10); t-tests were performed on all data (positive and negative), but only P values associated with positive t-values are marked and interpreted; each dot is a participant, bars reflect group means and error bars reflect within-participant s.e.m. The upper-right inset shows co-localization of processing in the insula, cingulate, somatosensory cortex and thalamus. The lines display selected major anatomical projections based on previous literature. Individual OFC and MPFC regions, MCC subregions, V1–V4 visual cortex and A1–A3 auditory cortex showed similar profiles thus were combined here. Thalamic nuclei: Pulv, pulvinar; Audi, A1–A3, primary and intermediate auditory cortices; Visual, V1/V2/V3/V4, primary and intermediate visual cortex; Audi INS, ventroposterior (auditory) insula (area 52, retroinsular cortex).

was applied only once to the dataset, without refitting or otherwise altering the model (that is, no model degrees of freedom). Participants made avoidance ratings in study 2, as in study 1, and unpleasantness ratings in studies 3 and 4. Studies 5 and 6 tested sensitivity and specificity to negative versus positive images and pain versus non-painful warmth, respectively.

Negative affect models developed in study 1 generalized across new samples: they significantly predicted negative affect ratings in most cases, and stimulus-type-specific models were largely sensitive and specific to their target stimulus types (Figs. 5 and 6).

The common model predicted ratings for all stimulus types: (a) mechanical pain ratings in study 2 ($r=0.34\pm0.12$, $P=0.006$, classification accuracy for the highest versus lowest stimulus levels = 72%, $P=0.02$, Fig. 5a); (b) thermal pain ratings in study 3 ($r=0.55\pm0.10$, $P=0.001$, accuracy = 91%, $P<0.001$, Fig. 5b); (c) aversive sound ratings in study 2 ($r=0.33\pm0.09$, $P<0.001$,

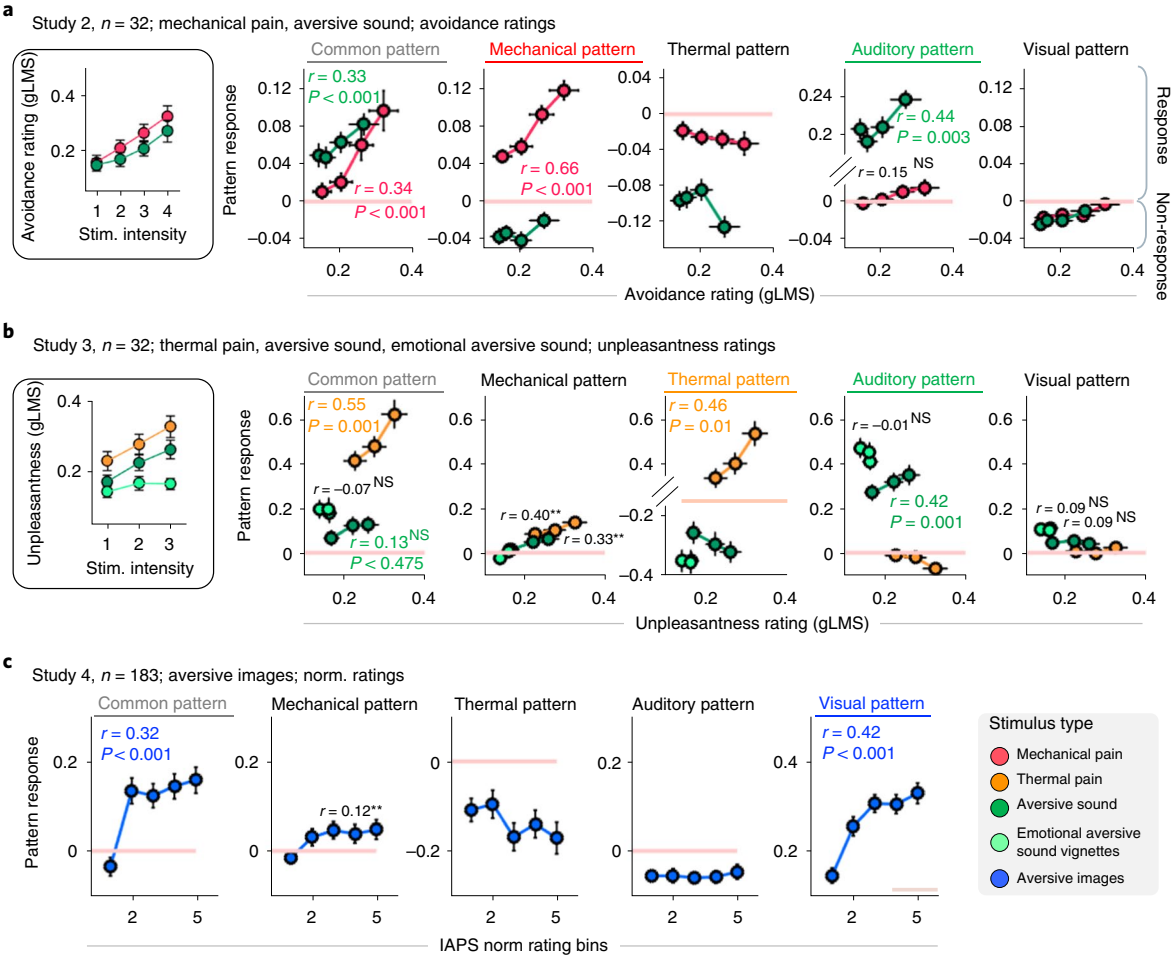


Fig. 5 | Prediction of negative affect in new datasets. For each study in **a**–**c**, plots 1–5 show the relationship between observed (self-report) and predicted (brain model response) negative affect and insets show subjective ratings per stimulus level. Model response was calculated as the dot product of the PLS pattern weights and fMRI activation maps. Model responses shown in each plot are indicated on top of the plot; those expected to show a significant response given the stimulus type and the associated 'true-positive' *r* values are bolded and color coded; *r* values of all other above-zero predictions are displayed in black. Broken plot axes are used to enable visualization on the same scale in the same plot. Model response values are derived from full-sample weight maps so that the model weights and test data are in each case independent. *r*, mean within-participant Pearson correlation between predicted and observed ratings; two-sided *P* values are based on a bootstrap test using 10,000 samples of within-participant *r* values. Negative affect ratings: avoidance rating scale (general Labeled Magnitude Scale (gLMS); study 2), unpleasantness rating scale (gLMS; study 3), IAPS normative ratings (study 4). Insets, data are shown as mean rating values across participants for each stimulus type, and error bars reflect within-participant s.e.m.; main plots, data are shown as mean model response values across participants for each stimulus type, and error bars reflect within-participant s.e.m. (horizontal error bars for *x* values, where applicable; vertical error bars for *y* values). NS, not significant.

accuracy = 78%, $P = 0.002$, Fig. 5a); and (d) aversive image ratings in study 4 ($r = 0.32 \pm 0.04$, $P < 0.001$, Fig. 5c). The common model did not significantly predict aversive 'knife on bottle' sound ratings in study 3, however ($r = 0.13 \pm 0.13$, $P = 0.39$, accuracy = 59%, $P = 0.38$, Fig. 5b; see also below).

Stimulus-type-specific models were sensitive and largely specific to ratings of target stimuli for mechanical pain in study 2 ($r = 0.66 \pm 0.07$, $P < 0.001$, accuracy = 91%, $P < 0.001$; Fig. 5a), for thermal pain in study 3 ($r = 0.46 \pm 0.11$, $P = 0.001$, accuracy = 91%, $P < 0.001$; Fig. 5b) and for aversive images in study 4 ($r = 0.42 \pm 0.03$, $P < 0.001$; Fig. 6c). However, the mechanical pain-specific model was also sensitive to ratings of off-target stimuli in study 3 ($r = 0.40$, $P = 0.004$ for thermal pain and $r = 0.33$, $P = 0.006$ for aversive sound) and study 4 ($r = 0.12$, $P = 0.008$ for aversive images; Fig. 5 and Supplementary Table 9).

Study 2 included the inherently aversive 'knife on bottle' sound stimuli used in study 1, whereas study 3 included this sound and also emotionally charged 'auditory vignettes' from the International Affective Digitized Sounds whose aversiveness was driven mainly by contextual information (people crying, gunshots). The aversive auditory model significantly predicted ratings to 'knife on bottle' sounds in both study 2 ($r = 0.44 \pm 0.10$, $P = 0.003$, accuracy 88%, $P < 0.001$; Fig. 5a) and study 3 ($r = 0.42 \pm 0.11$, $P = 0.001$, accuracy 75%, $P = 0.007$, Fig. 5b). However, it did not predict ratings to context-driven emotional sounds in study 3 ($r = -0.01$, $P = 0.92$) nor did the common model ($r = -0.07$, $P = 0.192$; Fig. 5b). Thus, the auditory negative affect model captured negative affect driven by intrinsic sound qualities rather than contextually driven emotional content.

Valence specificity and additional validation analyses. Our affect models may reflect factors that are not specific to negative valence, such as attention and salience[56]. The use of multiple aversive controls for each stimulus type, as reported above (Fig. 2b,c), provides evidence that type-specific models are selective to a particular

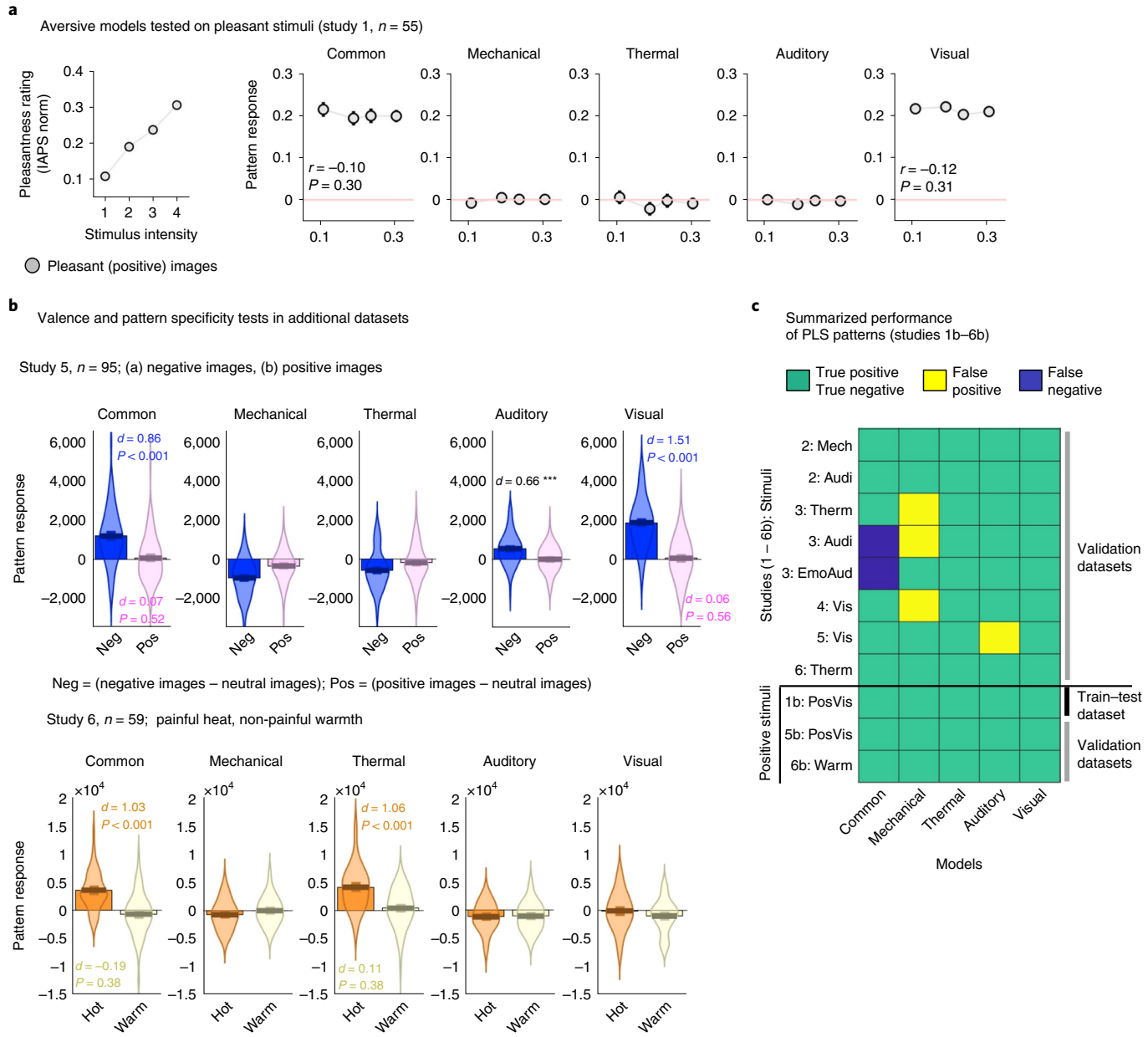


Fig. 6 | Valence specificity and additional validation tests. a, Aversive visual models (common, specific) did not significantly track normative ratings to pleasant images (from the IAPS) in the same participants (study 1b; cross-validated pattern responses). *r*, mean within-participant Pearson correlation between predicted and observed ratings; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; *P* values are based on a 10,000-sample bootstrap test of within-participant *r* values. **b**, Validation tests of patterns in two additional independent datasets. In study 5, aversive visual models (common, specific) responded to (a) negative visual stimuli (model sensitivity) but not to (b) positive visual stimuli (valence specificity). In study 6, thermal pain models (common, specific) responded to (a) painful heat (model sensitivity) but not to (b) non-painful warm stimuli (valence specificity). Other stimulus-specific model responses were plotted for completeness, and as expected, did not respond to off-target aversive stimuli nor to positive images or warm stimuli. *d*, Cohen's *d* effect size; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; one-sample *t*-test treating the subject as a random effect; bars reflect mean values across participants, and error bars reflect within-participant s.e.m. **c**, Performance of PLS-R patterns developed in study 1 is summarized across all tested datasets (studies 1b–6b); true-positive responses (significant positive pattern response for on-target stimuli) and true-negative responses (nonsignificant or negative pattern response for off-target stimuli) are displayed in green; false-positive responses (significant pattern response for off-target stimuli) are displayed in yellow; false-negative responses (nonsignificant or negative pattern response for on-target stimuli) are displayed in blue.

type of negative affect. This rules out a general attention, arousal or salience explanation for the type-specific models, but does not address this issue for the common model.

We tested specificity to negative affect in another way, by testing whether the models predicted normative pleasantness ratings of positive stimuli. In study 1, neither the common nor visual type-specific models predicted normative pleasantness (mean within-participant $r = -0.10$, $P = 0.30$ for the common model and $r = -0.12$, $P = 0.31$ for the visual model; Fig. 6a). Positive and negative stimuli from the International Affective Picture System (IAPS) were largely matched on normative arousal, excluding the possibility that negative images were simply more arousing. Additional cross-validated PLS-R models

trained to predict normative ratings for negative and positive IAPS images (Supplementary Methods and Extended Data Fig. 1a) ruled out the possibility that pleasantness ratings might simply be poorly predicted by brain activity. These analyses revealed a double dissociation in brain representations for negative and positive affect (see pattern responses in Extended Data Fig. 1b, and PLS-R weight and model encoding maps in Extended Data Fig. 1c). The negative-specific model predicted normative aversiveness ratings ($r=0.49$, $P=0.001$), but not normative pleasantness ratings (negative predicted slope, $r=-0.41$), and the positive-specific model predicted pleasantness ratings ($r=0.75$, $P=0.002$), but not aversiveness ratings (negative predicted slope, $r=-0.11$). An additional 'arousal' model trained to predict both positive and negative ratings combined predicted rating intensity for both types ($r=0.58$, $P=0.001$ for negative images and $r=0.69$, $P=0.001$ for positive images). This analysis shows that normative pleasantness is encoded in brain patterns distinct from our negative affect models.

We further tested the negative affect models' specificity for negative valence in two additional independent datasets. In study 5 (Fig. 6b), aversive visual models (common, specific) showed a significant pattern response to aversive IAPS images (Cohen's $d=0.86$, $P<0.001$ for the common model and $d=1.51$, $P<0.001$ for the aversive visual-specific model), showing model sensitivity. Neither the common nor visual-specific model responded to positive IAPS images ($d=0.07$, $P=0.52$ and $d=0.06$, $P=0.56$, respectively), showing valence specificity. Similarly, in study 6, thermal pain models (common, thermal-specific) responded significantly to painful heat ($d=1.03$, $P<0.001$ and $d=1.06$, $P<0.001$, respectively), showing model sensitivity. Neither model responded to non-painful warm stimuli ($d=-0.19$, $P=0.15$ and $d=0.11$, $P=0.38$, respectively), showing valence specificity. Other stimulus-type-specific models did not respond to positive images or warm stimuli, as expected. Thus, the brain models in study 1 were specific to negatively valenced stimuli in these test datasets and not sensitive to positive pictures. However, other types of appetitive stimuli may activate similar regions and patterns to our models (Extended Data Fig. 1), and more comprehensive tests of specificity across appetitive tasks are needed (Discussion).

Summary of predictive performance in independent samples. Figure 6c summarizes the models' performance across studies 1b–6. Across 55 model generalization tests (5 models × 11 tests), the models' decisions were correct (that is, true positives on-target and true negatives off-target) in 49 cases. Fifty of these tests were performed on independent study cohorts (studies 2–6). We found two false-negative cases (failures of sensitivity) and three false-positive cases (failures of specificity), as reported above. Together, the models developed in study 1 were sensitive to negative affect and specific to negative versus positive affect, providing preliminary evidence for generalizability that should be tested on a wider array of aversive and appetitive tasks in future studies. Such tests should carefully consider whether conditions tested are positive or negative (for example, affective responses to sexual images are complex[57]).

Discussion

Understanding how human affect is encoded in the brain is a central neuroscientific question. Psychological theories[3–5] and frameworks like the NIH Research Domain Criteria[11] have focused on generalized negative valence as a cross-modal construct, and the existence of generalized 'negative affect' in the brain is widely assumed. On the other hand, different affective stimuli activate differentiable neural populations and pathways in animal studies[28,30,33,58] and many clinical effects are found only with specific aversive stimulus types (for example, ref. [51]). Studies have thus provided evidence for both common and stimulus-type-specific negative affect. However, by focusing on these accounts separately—and typically with only one or two stimulus types tested in the same individuals—inferences about the nature of affect coding in the human brain have been limited.

Here we combined a multimodal experimental paradigm with a predictive modeling approach designed to uncover whether brain representations of negative affect are generalizable across types, specific to stimulus types, or a combination of both. In study 1, we jointly estimated common (general) and stimulus-type-specific representations of negative affect (subjective ratings) across four types of aversive stimuli. The findings indicate that negative affect is encoded in a combination of general (common) and stimulus-type-specific representations. Studies 2–6 provided further evidence that these representations are generalizable across individuals and cohorts and can be applied separately or jointly to predict negative affect in new studies. Applications to new studies include (a) characterizing neural differences related to disorders and subgroups, (b) predicting or monitoring the progression of mental health disorders over time, and (c) providing targets for behavioral, pharmacological and neural interventions. Further validation will help refine the use cases and boundary conditions for such applications.

Model encoding maps revealed several principles underlying the architecture of negative affect. First, some voxels encoded a general aversion signal across stimulus types, and this common negative affect representation was distributed across cortical, forebrain and brainstem regions. Extending recent work on cross-modal affect[16,59], the common model was related to a set of 'core' affect areas, including midline aMCC, thalamus, brainstem, cerebellum, amygdala and lateral OFC. Second, affect was represented in a stimulus-type-specific manner in sensory thalamocortical and corticobulbar pathways. Third, the models predicted the intensity of negative affect and did not respond to positive stimuli, at least among tasks studied here. The models therefore probably do not code for arousal, salience, or other unsigned processes related to affective intensity (for example, associability)[56,60]. Their response patterns, and the double dissociation observed between negative and positive affective images, are also inconsistent with bipolar (single-dimension) encoding but consistent with the idea that negative and positive affect are encoded in separable systems[1,18,41]. Fourth, several key affect-related regions (for example, the amygdala) contained overlapping populations of voxels that contributed to multiple types of negative affect. This provides a substrate for common and type-specific valence systems to interact within the same anatomical regions.

A commonly held perspective is that negative affect is encoded chiefly in multimodal forebrain regions such as OFC/vMPFC and amygdala[16,59,61]. However, we found that negative affect ratings were also linearly encoded by activity in early sensory cortices, with pINS and somatosensory cortex predicting pain, auditory cortex predicting auditory negative affect and visual cortex predicting visual negative affect. These findings align with previous findings that affective associations with visual stimuli are embedded in the visual cortex[27,45], and, more broadly, that category-specific affective information is embedded in early sensory areas[18,43,44,62]. Our results, however, extend beyond classification of stimulus types to suggest that subjective ratings of stimuli are also encoded in sensory cortices.

The co-localization of different types of affective representations also provides a basis for multimodal integration and interactions (for example, between pain and emotion[63,64]). Although sensory pathways were largely modality specific, some (for example, S1, S2, pINS, SC, IC and LGN) also encoded general negative affect, suggesting integrative processing in these areas. In other areas, including the amygdala and all multimodal cortical regions tested, distinct stimulus-specific representations were encoded in close anatomical proximity to each other.

The pattern-based approach we used, which shows promise for disentangling general and stimulus-type-specific signals, may help address discrepant findings in previous studies. For example, some studies have reported that the amygdala encodes nonspecific arousal

or salience-related signals (or 'unsigned valence'; for example, refs. [56,60]), but others have shown valence-specific effects in the amygdala matching for arousal[65]. Here, the amygdala contributed to both general and vision-specific (but not auditory or somatosensory) negative affect, but an overlapping population of voxels encoded an 'arousal' signal (Extended Data Fig. 1d,e). Overlapping patterns of fMRI brain response (with different population vectors) could, in principle, separately predict negative affect, positive affect[16,66,67], and general 'arousal'. Positive affect might be processed in some of the regions we identified, including vStr, albeit in separable neural populations (Extended Data Fig. 1d,e). This approach can be extended to other tasks as well. For example, vStr might encode more rewarding (appetitive) positive stimuli[1,68] not investigated in the present work.

Brain measures that track distinct types of affect may be important for identifying disorder-relevant affective brain responses. For example, we found a double dissociation in mechanical and thermal pain encoding in S1 and pINS/S2, respectively. Responses to thermal, mechanical and chemical pain are relatively uncorrelated (for example, ref. [51]) and clinical hypersensitivity is often stimulus-type specific[69]. Some disorders are characterized by increased sensitivity to mechanical pain[70,71] and others to thermal pain[72]. We also identified dissociable pINS/S2 regions encoding pain versus auditory affect, and a double dissociation between auditory and visual affect in the IC and SC, respectively. Our approach thus identifies dissociations that have been difficult to separate using standard fMRI methods.

Finally, our findings provide evidence that accounting for multiple common and stimulus-type-specific representations is important for developing accurate predictive models. Accurate predictions of subjective affective experience required jointly considering common and type-specific measures in all cases we tested. This principle is consistent with a growing literature showing that even basic affective judgments are complex processes that involve coordination of multiple brain systems. Future studies must consider a range of contextual factors and individual differences, which likely influence the nature of affect representations and how they combine to create subjective experience. Nevertheless, present findings demonstrate a substantial degree of consistency across individuals and studies, and establish a baseline for future context-dependent and subgroup-dependent models.

This paper has several limitations. First, we only included one type of positive stimuli (visual images) and thus, our valence specificity tests are largely limited to visual stimuli, although we provide preliminary evidence for valence specificity for thermal stimuli. Second, not all models performed equally well when tested in new samples; the mechanical pain-specific model in particular might not generalize as well as other models. Third, conventional 3T fMRI cannot be precisely localized to midbrain and brainstem nuclei. However, we and others have found reasonable localization in these areas at 3T, for example, PAG, RVM, SC, IC and other nuclei[53,73], in some cases validated with high resolution (~1 mm) 7T fMRI[46,53].

In conclusion, we show that negative affect is encoded in the brain in multiple distributed representations, some generalizable across affective stimulus types and others specific to negative affect elicited by a particular stimulus type. Negative affect is embedded in sensory pathways, and integrative regions represent distinct combinations of negative affect types. The resulting models provide a set of measures that can serve to understand disorders, track the progression of disorders and treatments over time, and serve as targets for interventions. They also lead to further basic and translational research questions. One area for future development concerns how common and distinct representations are integrated, how 'cross-talk' across sensory modalities occurs and how 'cross-talk' may be enhanced in disorders, helping to explain, for example, pain hypersensitivity after emotional trauma. Another area concerns hierarchical coding, including whether common and type-specific affective codes are parallel or hierarchical, and how context-based predictive signals are integrated info affect representations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01082-w.

Received: 24 September 2021; Accepted: 25 April 2022;
Published online: 30 May 2022

References

1. Berridge, K. C. & Kringelbach, M. L. Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Curr. Opin. Neurobiol.* **23**, 294–303 (2013).
2. Tye, K. M. Neural circuit motifs in valence processing. *Neuron* **100**, 436–452 (2018).
3. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
4. Russell, J. A. & Barrett, L. F. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *J. Pers. Soc. Psychol.* **76**, 805–819 (1999).
5. Gray, J. A. *The Psychology of Fear and Stress*. (CUP Archive, 1987).
6. Montague, P. R. & Berns, G. S. Neural economics and the biological substrates of valuation. *Neuron* **36**, 265–284 (2002).
7. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
8. Hariri, A. R. et al. A susceptibility gene for affective disorders and the response of the human amygdala. *Arch. Gen. Psychiatry* **62**, 146–152 (2005).
9. Bishop, S., Duncan, J., Brett, M. & Lawrence, A. D. Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. *Nat. Neurosci.* **7**, 184–188 (2004).
10. Duerden, E. G. & Albanese, M.-C. Localization of pain-related brain activation: a meta-analysis of neuroimaging data. *Hum. Brain Mapp.* **34**, 109–149 (2013).
11. Insel, T. et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* **167**, 748–751 (2010).
12. Hayden, B. Y. & Niv, Y. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* **135**, 192–201 (2021).
13. Gehrlach, D. A. et al. Aversive state processing in the posterior insular cortex. *Nat. Neurosci.* **22**, 1424–1437 (2019).
14. Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P. & Singer, T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nat. Commun.* **7**, 10904 (2016).
15. Kragel, P. A. et al. Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018).
16. Chikazoe, J., Lee, D. H., Kriegeskorte, N. & Anderson, A. K. Population coding of affect across stimuli, modalities and individuals. *Nat. Neurosci.* **17**, 1114–1122 (2014).
17. Kober, H. et al. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* **42**, 998–1031 (2008).
18. Satpute, A. B. et al. Involvement of sensory regions in affective experience: a meta-analysis. *Front. Psychol.* **6**, 1860 (2015).
19. Ekman, P. An argument for basic emotions. *Cognition Emot.* **6**, 169–200 (1992).
20. Friedman, B. H. Feelings and the body: the Jamesian perspective on autonomic specificity of emotion. *Biol. Psychol.* **84**, 383–393 (2010).
21. Barrett, L. F. Solving the emotion paradox: categorization and the experience of emotion. *Pers. Soc. Psychol. Rev.* **10**, 20–46 (2006).
22. Saarimäki, H. et al. Discrete neural signatures of basic emotions. *Cereb. Cortex* **26**, 2563–2573 (2016).
23. Kragel, P. A. & LaBar, K. S. Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cogn. Affect. Neurosci.* **10**, 1437–1448 (2015).
24. Stephens, C. L., Christie, I. C. & Friedman, B. H. Autonomic specificity of basic emotions: evidence from pattern classification and cluster analysis. *Biol. Psychol.* **84**, 463–473 (2010).
25. Horing, B., Sprenger, C. & Büchel, C. The parietal operculum preferentially encodes heat pain and not salience. *PLoS Biol.* **17**, e3000205 (2019).

26. Wager, T. D. et al. A Bayesian model of category-specific emotional brain responses. *PLoS Comput. Biol.* **11**, e1004066 (2015).
27. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).
28. Corder, G. et al. An amygdalar neural ensemble that encodes the unpleasantness of pain. *Science* **363**, 276–281 (2019).
29. Hua, T. et al. General anesthetics activate a potent central pain-suppression circuit in the amygdala. *Nat. Neurosci.* **23**, 854–868 (2020).
30. Chiang, M. C. et al. Divergent neural pathways emanating from the lateral parabrachial nucleus mediate distinct components of the pain response. *Neuron* **106**, 927–939 (2020).
31. Allen, W. E. et al. Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149–1155 (2017).
32. Allen, W. E. et al. Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364**, 253 (2019).
33. Pool, A.-H. et al. The cellular basis of distinct thirst modalities. *Nature* **588**, 112–117 (2020).
34. McIntosh, A. R. & Lobaugh, N. J. Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage* **23**, 250–263 (2004).
35. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
36. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
37. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
38. Nimon, K., Lewis, M., Kane, R. & Haynes, R. M. An R package to compute commonality coefficients in the multiple regression case: an introduction to the package and a practical example. *Behav. Res. Methods* **40**, 457–466 (2008).
39. Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* **87**, 96–110 (2014).
40. Hayes, D. J. & Northoff, G. Identifying a network of brain regions involved in aversion-related processing: a cross-species translational investigation. *Front. Integr. Neurosci.* **5**, 49 (2011).
41. Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J. & Barrett, L. F. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cereb. Cortex* **26**, 1910–1922 (2016).
42. Panzeri, S., Macke, J. H., Gross, J. & Kayser, C. Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* **19**, 162–172 (2015).
43. Mouraux, A., Diukova, A., Lee, M. C., Wise, R. G. & Iannetti, G. D. A multisensory investigation of the functional significance of the 'pain matrix'. *Neuroimage* **54**, 2237–2249 (2011).
44. Liang, M., Mouraux, A., Hu, L. & Iannetti, G. D. Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nat. Commun.* **4**, 1979 (2013).
45. Shuler, M. G. & Bear, M. F. Reward timing in the primary visual cortex. *Science* **311**, 1606–1609 (2006).
46. Kragel, P. A. et al. A human colliculus-pulvinar-amygdala pathway encodes negative emotion. *Neuron* https://doi.org/10.1016/j.neuron.2021.06.001 (2021).
47. Pessoa, L. & Adolphs, R. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* **11**, 773–783 (2010).
48. Chen, C., Cheng, M., Ito, T. & Song, S. Neuronal organization in the inferior colliculus revisited with cell-type-dependent monosynaptic tracing. *J. Neurosci.* **38**, 3318–3332 (2018).
49. Tan, L. L. & Kuner, R. Neocortical circuits in pain and pain relief. *Nat. Rev. Neurosci.* **22**, 458–471 (2021).
50. Mogil, J. S. The genetic mediation of individual differences in sensitivity to pain and its inhibition. *Proc. Natl Acad. Sci. USA* **96**, 7744–7751 (1999).
51. Baron, R. et al. Peripheral neuropathic pain: a mechanism-related organizing principle based on sensory profiles. *Pain* **158**, 261–272 (2017).
52. Price, D. D. Psychological and neural mechanisms of the affective dimension of pain. *Science* **288**, 1769–1772 (2000).
53. Satpute, A. B. et al. Identification of discrete functional subregions of the human periaqueductal gray. *Proc. Natl Acad. Sci. USA* **110**, 17101–17106 (2013).
54. Craig, A. D., Bushnell, M. C., Zhang, E. T. & Blomqvist, A. A thalamic nucleus specific for pain and temperature sensation. *Nature* **372**, 770–773 (1994).
55. Woo, C.-W. et al. Quantifying cerebral contributions to pain beyond nociception. *Nat. Commun.* **8**, 14211 (2017).
56. Anderson, A. K. & Sobel, N. Dissociating intensity from valence as sensory inputs to emotion. *Neuron* **39**, 581–583 (2003).
57. Wehrum, S. et al. Gender commonalities and differences in the neural processing of visual sexual stimuli. *J. Sex. Med.* **10**, 1328–1342 (2013).
58. Neugebauer, V. Amygdala pain mechanisms. *Handb. Exp. Pharmacol.* **227**, 261–284 (2015).
59. Kim, J., Shinkareva, S. V. & Wedell, D. H. Representations of modality-general valence for videos and music derived from fMRI data. *Neuroimage* **148**, 42–54 (2017).
60. Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A. & Daw, N. D. Differential roles of human striatum and amygdala in associative learning. *Nat. Neurosci.* **14**, 1250–1252 (2011).
61. Belova, M. A., Paton, J. J. & Salzman, C. D. Moment-to-moment tracking of state value in the amygdala. *J. Neurosci.* **28**, 10023–10030 (2008).
62. Hayes, D. J. & Northoff, G. Common brain activations for painful and non-painful aversive stimuli. *BMC Neurosci.* **13**, 60 (2012).
63. Villemure, C. & Bushnell, M. C. Mood influences supraspinal pain processing separately from attention. *J. Neurosci.* **29**, 705–715 (2009).
64. Roy, M., Piché, M., Chen, J.-I., Peretz, I. & Rainville, P. Cerebral and spinal modulation of pain by emotions. *Proc. Natl Acad. Sci. USA* **106**, 20900–20905 (2009).
65. Anders, S., Eippert, F., Weiskopf, N. & Veit, R. The human amygdala is sensitive to the valence of pictures and sounds irrespective of arousal: an fMRI study. *Soc. Cogn. Affect. Neurosci.* **3**, 233–243 (2008).
66. Woo, C.-W. et al. Separate neural representations for physical pain and social rejection. *Nat. Commun.* **5**, 5380 (2014).
67. Peelen, M. V. & Downing, P. E. Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends Cogn. Sci.* **11**, 4–5 (2007).
68. Tomova, L. et al. Acute social isolation evokes midbrain craving responses similar to hunger. *Nat. Neurosci.* **23**, 1597–1605 (2020).
69. Woolf, C. J. Central sensitization: implications for the diagnosis and treatment of pain. *Pain* **152**, S2–S15 (2011).
70. Ceko, M., Bushnell, M. C., Fitzcharles, M.-A. & Schweinhardt, P. Fibromyalgia interacts with age to change the brain. *Neuroimage Clin.* **3**, 249–260 (2013).
71. López-Solà, M. et al. Towards a neurophysiological signature for fibromyalgia. *Pain* **158**, 34–47 (2017).
72. Grothusen, J. R., Alexander, G., Erwin, K. & Schwartzman, R. Thermal pain in complex regional pain syndrome type I. *Pain Physician* **17**, 71–79 (2014).
73. Eippert, F. et al. Activation of the opioidergic descending pain control system underlies placebo analgesia. *Neuron* **63**, 533–543 (2009).

Methods

Data from six studies were used in this paper. Study 1 was the primary study used for brain model development and all main analyses. Studies 2, 3 and 4, 5a and 6a contained different aversive stimuli in novel participants and were used for prospective validation of brain models developed in study 1. Study 1b contained pleasant visual stimuli tested in the same participants as study 1 and was used to test whether the aversive brain models are sensitive to other attentionally demanding, salient stimuli (that is, positive images). Studies 5b and 6b contained positive pictures and non-painful warm stimuli and were used for additional valence specificity tests. Experimental parameters for each study are listed in Supplementary Table 8.

Participants. Study 1 included 55 adult participants (mean ± s.d. age: 24.9 ± 6.8 years; 31 males, 24 females; 9 left-handed; 47 white and 8 non-white (1 Hispanic, 5 Asian, 1 Black and 1 American Indian)). All participants were healthy, with normal or corrected to normal vision and normal hearing, and with no history of psychiatric, physiological or pain disorders and neurological conditions, no current pain symptoms, and no MRI contraindications. Eligibility was assessed with a general health questionnaire, a pain safety screening form and an MRI safety screening form. Participants were recruited from the Boulder/Denver Metro Area. The institutional review board of the University of Colorado Boulder approved the study, and all participants provided written consent. Participants were compensated at a rate of $12 per hour for behavioral sessions (that is, tasks outside the fMRI scanner) and $24 per hour for fMRI sessions. Data collection and analysis were not performed blind to the conditions of the experiments. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[23–25,74]. No participants were excluded from the analysis.

Study design. The study 1 dataset included in the current paper is part of a larger study, in which participants completed two fMRI sessions in a counterbalanced order. In each session, we administered the 'multimodal aversiveness task'. In the 'experience' session, presented here, participants were instructed to 'experience aversive stimuli as they come'. In the 'regulate' session (not presented here), participants were instructed to downregulate aversive stimuli using a cognitive self-regulation strategy. Physiological recordings were collected throughout each session (not presented here).

fMRI task design. The 'multimodal aversive experience task' was developed for this study to test brain responses to multiple instances of negative affect in the same individuals (Fig. 1a). The task comprised a series of aversive stimuli of different types (painful pressure, painful heat, aversive sounds, aversive images) and a pleasant (positive) stimulus type (pleasant images). In total, participants received 96 aversive stimuli (4 stimulus types × 4 intensities × 6 runs presented in random order) and 24 pleasant stimuli (4 intensities × 6 runs) over six fMRI runs and rated their experience after each stimulus (including the positive stimuli) using a uniform rating scale, detailed below.

Negative affect rating. We obtained rating scores of negative affect on the same scale across different stimulus types. Participants rated negative affect ('aversiveness') in terms of avoidance ('how much do you want to avoid this experience in the future?') on a gLMS scale. The gLMS was chosen over the more common visual analog scale as it might be better suited for cross-modal comparisons and for capturing subjective variance in the high-intensity stimulus range[75]. The gLMS was anchored at 'not at all' (0) to 'most' (1), with intermediate labels spaced quasi-logarithmically ('a little bit' (0.061), 'moderately' (0.172), 'strongly' (0.354) and 'very strongly' (0.533). Before to the fMRI session, participants were instructed in the meaning of scale labels: 'not at all' = I feel no need to avoid this experience, 'a little bit' = I would put a little effort into avoiding this experience, 'moderately' = I would prefer to avoid this experience in the future, 'strongly' = I strongly want to avoid this experience in the future, 'very strongly' = I very much want to avoid this experience and 'most' = I never want to experience this again in my life. During the fMRI experiment, the intermediate labels were removed to minimize clustering of ratings around the labels[76].

Calibration of pain stimuli. Before the fMRI experiment, all participants underwent a short pain calibration session to assure normal pain sensitivity. Participants experienced different levels of thermal and pressure pain stimuli in a random order (thermal: between 43 and 49 °C, pressure: between 4 and 7 kg/cm^2, maximum duration of 10 s). The highest stimulus level was chosen based on previous studies as tolerable yet painful for most participants. All participants included in the study were able to tolerate all stimuli.

Aversive stimuli. Mechanical pain stimuli were administered using an in-house pressure pain device. The pressure pain device is an MRI-safe device with dynamic pressure delivery controlled by LabView (National Instruments). Four stimulus levels were delivered to the left thumbnail for 6 s each (level 1: 4 kg/cm^2; level 2: 5 kg/cm^2; level 3: 6 kg/cm^2; level 4: 7 kg/cm^2).

Thermal pain stimuli were administered using an ATS Pathway System (Medoc) with a 16-mm Peltier contact thermode (that is, hot plate). Four stimulus intensity levels were delivered to the thenar eminence of the left hand (level 1: 45 °C; level 2: 46 °C; level 3: 47 °C; level 4: 48 °C) and each stimulus lasted 10 s (1.5–2-s ramp-up, 1.5–2-s ramp-down, 6–7 s at target temperature).

Aversive sound was administered using MRI-compatible headphones. We used the sound of a knife scraping on a bottle (sound file retrieved from YouTube), which is a reliable aversive auditory stimulus[77,78]. Four stimulus intensity levels were delivered at 2,000 Hz for 6 s each (level 1: level 4 minus 8 dB; level 2: level 4 minus 4 dB; level 3: level 4 minus 1 dB; level 4: original YouTube sound file).

Aversive images were presented on the MRI screen and included normed images from the IAPS database[79]. To induce four stimulus intensity levels, we selected four groups of seven images each in a two-step process: (1) preliminary selection based on normed aversiveness ratings (averaged across male and female raters) available in the IAPS database, and (2) final selection based on ratings by $N = 10$ laboratory members (5 males, 5 females) in response to 'how aversive is this image? 1–100'. Selected images included photographs of animals (7), bodily illness and injury (12), industrial and human waste (9). Four stimulus levels were delivered to participants for 6 s each.

fMRI data acquisition and preprocessing. Whole-brain fMRI data were acquired on a 3T Siemens MAGNETOM Prisma MRI scanner at the Intermountain Neuroimaging Consortium facility at the University of Colorado, Boulder. Structural images were acquired using high-resolution T1 spoiled gradient recall images and were used for anatomical localization and warping to the standard Montreal Neurological Institute (MNI) space only. Functional images were acquired with a multiband EPI sequence (repetition time = 460 ms, echo time = 27.2 ms, field of view = 220 mm, multiband acceleration factor = 8, flip angle = 44°, 64 × 64 matrix, 2.7 × 2.7 × 2.7 mm voxels, 56 interleaved ascending slices, phase encoding posterior » anterior). In total, six runs of 7.17 min in duration (= 934 measurements) were acquired. Stimulus presentation and behavioral data acquisition were controlled using Psychtoolbox (MATLAB, MathWorks).

fMRI data were preprocessed using an automated pipeline based on AFNI, FSL and SPM5, and implemented by the Mind Research Network. Briefly, the preprocessing steps included: distortion correction using FSL's top-up tool (https://fsl.fmrib.ox.ac.uk/fsl/), motion correction (affine alignment of first EPI volume (reference image) to T1, followed by affine alignment of all EPI volumes to the reference image and estimation of the motion parameter file (sepi_vr_motion.1D, AFNI; https://afni.nimh.nih.gov/), spatial normalization via the participant's T1 image (T1 normalization to MNI space (nonlinear transform), normalization of EPI image to MNI space (3dNWarpApply, AFNI; https://afni.nimh.nih.gov/), interpolation to 2 × 2 × 2 mm^3 voxels and smoothing with a 6-mm FWHM kernel (SPM 8; https://www.fil.ion.ucl.ac.uk/spm/software/spm8/). Spatial smoothing improves interindividual functional alignment without impairing the sensitivity of multivariate pattern analyses[80]. Before first-level analysis, in each run we removed the first 15 volumes of the fMRI data to allow for image intensity stabilization and identified severe global motion outliers (spikes) in the data. Spikes were defined as time points in the data with either the absolute global signal intensity or the Mahalanobis distance across slice-specific global means and spatial standard deviations exceeding ten median absolute deviations.

Behavioral data analysis. Behavioral data were analyzed using 'glmfit_multilevel.m', a multilevel generalized linear model implemented in custom MATLAB (2019a, MathWorks) code available from the authors' website (https://github.com/canlab/CanlabCore/). For each stimulus modality (mechanical, thermal, auditory and visual), the outcome variable was the average rating for each stimulus level. The within-participant predictor at the first-level model was stimulus intensity. Data distribution was assumed to be normal but was not formally tested.

fMRI analysis. fMRI data were analyzed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/) and custom MATLAB (2019a, MathWorks) code available from the authors' website (https://github.com/canlab/CanlabCore/). A univariate general linear model (GLM) was used to create images for the prediction analyses. Data distribution was assumed to be normal but was not formally tested. The participant-level GLM analyses were conducted in SPM12. The six runs of the fMRI task were concatenated for each participant. Boxcar regressors, convolved with the canonical hemodynamic response function, were constructed to model the 6–10-s stimulation and 4–7-s rating periods. The fixation cross epoch was used as an implicit baseline. A high-pass filter of 180 s was applied. Nuisance variables included: (a) 'dummy' regressors coding for each run (intercept for each run); (b) linear drift over time within each run; (c) the six estimated head movement parameters (x, y, z, roll, pitch and yaw), their mean-centered squares, their derivatives, and squared derivative for each run (total = 24 columns); and (d) motion outliers (spikes). Contrasts of interest (beta images) included stimulation periods averaged across six trials for each stimulus intensity (each against implicit baseline). The resulting beta maps computed for each stimulus level of each aversive stimulus type were used for brain model development (study 1). Beta maps computed for each stimulus level of the pleasant (positive) stimulus were used for brain model development in supplementary analyses (Supplementary Information).

PLS-R for brain model development. *Statistical approach.* In study 1, we developed common and stimulus-type-specific predictive models of aversiveness ratings from brain activity across the four stimulus types using PLS-R[35]. PLS-R estimates a set of latent brain components (voxel-wise spatial maps) and a set of latent negative affect rating factors that are optimized to be maximally intercorrelated (that is, maximal variance in ratings explained by brain patterns). Compared to standard predictive brain models that typically characterize a single outcome at a time, PLS-R jointly estimates multiple solutions (that is, separate brain patterns for common and stimulus-type-specific outcomes) simultaneously, which is why it is capable of predicting multiple (correlated) stimulus types, as is the case with our data. The predictors (brain activity) are stored in the input matrix **X** and the outcome variables (ratings) are stored in the input matrix **Y**. By an iterative application of a singular value decomposition algorithm, which factorizes (decomposes) the cross-product matrix of the two input matrices, PLS-R finds latent variables, also called component scores, that model **X** (for example, brain activity) and simultaneously predict **Y** (for example, ratings). Each run of the singular value decomposition algorithm produces orthogonal latent variables and corresponding regression weights for predictions. By estimating different latent sources, PLS-R can provide improved estimates of common and specific patterns (versus single-outcome models such as PCR), but these are not necessarily fully independent of each other.

Implementation. PLS-R was conducted using 'plsregress.m' in MATLAB (version R2019b). Predictors (**X**) constituted 880 whole-brain activation maps associated with participants (55) *x* stimulus intensity level (4) *x* stimulus type (4), aggregated into an images *x* voxels matrix (stacked across participants), and split into training and test sets using fivefold blocked cross-validation (leaving out all images associated with test-set participants together). The activation maps were derived via univariate GLM analysis in which we modeled stimulation periods against implicit baseline and averaged across trials of each stimulus level of each stimulus type to obtain 16 contrast images for each participant. We constructed the outcome (**Y**) matrix to include negative affect ratings across all stimulus types ($\mathbf{Y}_1$) as well as stimulus-type-specific negative affect ($\mathbf{Y}_2-\mathbf{Y}_5$, ratings for each stimulus type separately, with values of 0 for other stimulus types). By setting the **Y** value of other stimulus types to 0 we constrained each pattern to be stimulus-type specific. The linear combination of latent brain factors that explains $\mathbf{Y}_1$ reflects a common model of negative affect across stimulus types. Likewise, brain patterns predictive of $\mathbf{Y}_2-\mathbf{Y}_5$ are models optimized to be selective to mechanical pain, thermal pain, aversive sounds and aversive images. Each model was then projected into a single predictive spatial map.

To estimate their predictive accuracy, and specificity for the target outcome, and generalization to new individual participants, these patterns were applied to fMRI activation maps obtained from new participants in cross-validation test sets (fivefold; leaving out all data for each test participant) and prospectively applied to independent studies ($N=247$; see 'Validation in independent datasets' below). Model (pattern) responses were calculated using the dot product of each pattern weight map with the univariate GLM-derived fMRI activation map for each participant for each stimulus level. Weight maps applied to study 1 and study 1b were based on data from out-of-sample individuals (cross-validated estimates), and the final pattern weights applied to studies 2–6 were based on the full study sample (full samples estimates).

Model evaluation. To evaluate the models' performance, we assessed in each participant the RMSE for each model's predictions of its target outcome (that is, four average ratings per stimulus type) and the ability to significantly predict increasing negative affect within a participant. To provide an interpretable effect size metric, we estimated in each participant the Pearson correlation (r) between observed and cross-validated predicted ratings and tested whether the output of one model predicted the specific type of negative affect it was trained to predict (sensitivity) and not other types (specificity).

Classification between stimulus levels. For each brain model, we computed the classification accuracy between each pair of stimulus intensity levels (1 versus 2, 2 versus 3, 3 versus 4, 1 versus 3, 2 versus 4 and 1 versus 4) from receiver operating characteristic curves using forced-choice classification. Forced-choice classification uses the maximum value of a relative comparison within a participant and is therefore 'threshold free'. P values were calculated using a two-sided independent binomial test.

Cross-prediction. To test how well each PLS-R brain pattern predicted other stimulus types, we used a cross-prediction procedure on cross-validated estimates. In this procedure, a Pearson correlation was calculated for each participant separately between the predicted and the observed outcomes (that is, negative affect ratings) for each stimulus type. The mean within-participant correlation coefficient for each train–test stimulus pair is visualized in a cross-correlation matrix.

Variance decomposition analysis. For each stimulus type separately, full and reduced regression models (commonality analysis, implemented in R[38]) were used in each participant to partition the variance in outcome (ratings) explained by the predictors (common model, specific model) into unique and shared components. First, total r^2 was defined as the mean total variance explained by common and specific models in a multiple regression. The unique r^2 for the common model (UVC) was computed as: total r^2 − single variance for the specific model, whereas unique r^2 for the specific model (UVS) was computed as: total r^2 − single variance for the common model. Shared variance between common and specific models was computed as: total r^2 − UVS − UVC. Proportions of variance explained were computed on cross-validated model outputs.

Identifying core systems involved in negative affect. *General approach.* Core regions for multimodal (common) negative affect processes were defined as having voxels that reliably contribute to model prediction (that is, model weights) and are related to model predictions in an interpretable way (that is, model encoding voxels where the prediction correlates with fMRI activation). Core regions for stimulus-type-specific processes were defined as above, with an additional 'type-selectivity filter' applied to identify the most important regions. Thus, a core stimulus type-specific system was defined as having voxels that reliably contribute to prediction, encode the model and are selective for this model above all other models (3-way conjunction).

Step 1: model weight maps. To determine which brain areas made reliable contributions to the prediction and to threshold voxel weights for interpretation and display, we constructed 10,000 bootstrap samples (with replacement) consisting of paired brain and outcome data and performed PLS-R on each. The z-scores at each voxel were estimated based on the mean and standard error of the bootstrap distributions, and the statistical map was thresholded based on the corresponding P values. The maps were thresholded voxel-wise at $q<0.05$ (FDR corrected)[81]. Uncorrected maps thresholded at $t>3$ were used for display purposes in Supplementary Fig. 3a.

Step 2: model encoding maps. Model encoding ('structure coefficient') maps were computed for each participant by regressing the PLS-R model predictions on voxel-wise fMRI activation maps (four maps per person for each condition, corresponding to averages for each stimulus level). Structure coefficients identify voxels individually associated with each model's output, mapping individual voxels to the overall multivariate model prediction[39,82–84]. The analysis was performed using a standard summary statistics-based mixed-effects GLM[85], with robust regression at the second level, thresholded at FDR $q<0.05$ corrected for multiple comparisons. Uncorrected maps thresholded at $t>3$ were used for display purposes in Supplementary Fig. 3b.

Step 3: core system. The core system map for each model was derived via a conjunction of model weight maps thresholded at FDR $q<0.05$ created in step 1 and model encoding maps thresholded at FDR $q<0.05$ created in step 2 (Supplementary Fig. 4a). The conjunction was restricted to preserve positive values.

Step 4a: type-selective model encoding maps. Voxels where model encoding values were significantly greater in one model than in any of the other four models (that is, type-selective voxels) were calculated for each participant as a supremum statistic image for the target model encoding map minus the maximum of the four remaining model encoding maps (Supplementary Fig. 4b). A second-level robust t-test, thresholded at FDR $q<0.05$ corrected for multiple comparisons and preserving positive statistics only, identified regions selective for the respective model on participant group level.

Step 4b: core stimulus type-specific systems. Core stimulus-type-specific systems were derived via conjunction of the core system maps thresholded at FDR $q<0.05$ created in step 3 and the type-selective encoding maps thresholded at FDR $q<0.05$ created in step 4a.

Spatial similarity between model encoding maps and a priori regions of interest. River plots were created to depict spatial similarity between model encoding maps and a set of anatomical parcellations (ROIs), documented in previous neuroimaging studies as regions showing preferential activation to somatic pain stimuli, aversive auditory and aversive visual stimuli or regions showing activation to aversive stimuli across stimulus types (meta-analyses in refs. [18,41,62] and confirmed using Neurosynth in $N=238$ fMRI studies with the search term 'aversive'). Spatial similarity was calculated as cosine similarity between the ROI and the gray matter masked model encoding group map thresholded at $q<0.05$ FDR and retaining positive values for interpretation.

Regional model importance scores. The common/type-specific ratio was derived for each ROI separately from model encoding values. The common model importance score was defined as the relative percentage of the ROI encoded by the common model. The type-specific model importance score was defined as the percentage of ROI encoded by the predominant specific model (that is, the model encoded by the maximum number of voxels).

Testing model-selectivity along afferent pathways. To test the relative contribution of each model to negative affect representation in selected ROIs

along the afferent processing pathways[86], we extracted from each anatomically defined ROI and for each model separately the mean structure coefficients across participants. Significance of each model was tested using a one-sample *t*-test.

Validation in independent datasets. We tested whether brain models derived in study 1 ($N=55$) could predict negative affect ratings in new individuals by applying PLS-R-derived brain patterns to three independent test datasets (study 2, $N=32$; study 3, $N=32$; study 3, $N=183$; see Supplementary Table 8 for sample characteristics and study design). Pattern response was estimated for each test participant in each test condition by computing the dot product of each PLS-R-derived full-sample pattern with the participant's brain activation map, yielding a single scalar value. We estimated the sensitivity and specificity of each stimulus-type-specific pattern by testing whether the pattern responds significantly to the increasing level of self-reported negative affect for the respective (target) stimulus type (sensitivity to change in behavior; significant positive response with a positive slope) but has a nonsignificant response to other types of stimuli (specificity; nonsignificant or negative response or negative slope). For the common model, we tested whether the pattern responds significantly to the increasing level of self-reported negative affect across different aversive stimulus types. As in our main analyses, we calculated the mean within-participant Pearson correlation coefficient and the RMSE between the observed and predicted ratings as measures of model performance.

Valence specificity test and additional validation analyses. First, we tested the common and stimulus-type-specific patterns derived in study 1 on a set of pleasant visual stimuli collected in the same participants (labeled as study 1b; Supplementary Table 8). Pattern response was estimated as explained above. We used the cross-validated PLS-R pattern—testing in the same participants—to compute the dot product with each participant's brain activation map.

Second, we tested the PLS-R patterns derived in study 1 on two additional datasets, study 5 ($N=95$ participants, fMRI brain responses to negative and positive images; Supplementary Table 8) and study 6 ($N=59$; fMRI brain responses to painful and non-painful thermal stimulation; Supplementary Table 8). Pattern response was estimated as explained above. We used the full-sample PLS-R pattern to compute the dot product with each participant's brain activation map.

Third, we summarized the performance of PLS-R patterns derived in study 1 across all tested datasets (studies 1b–6b) in a single matrix showing (1) true-positive responses (significant positive pattern response for on-target stimuli) and true-negative responses (nonsignificant or negative pattern response for off-target stimuli) in green; (2) false-positive responses (significant pattern response for off-target stimuli) in yellow; and (3) false-negative responses (nonsignificant or negative pattern response for on-target stimuli) in blue.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Brain patterns generated and analyzed during the current study, as well as source data for figures are freely available via . The dataset used in study 6 is available at https://github.com/cocoanlab/interpret_ml_neuroimaging/.

Code availability

Code for analysis and for generating figures is openly shared at https://github.com/canlab/2021_Ceko_MPA2_Aversive/. Analyses reported in this paper were performed using code release v1.0.1 (https://doi.org/10.5281/zenodo.6452244).

References

74. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
75. Bartoshuk, L. M. et al. Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. *Physiol. Behav.* **82**, 109–114 (2004).
76. Hayes, J. E., Allen, A. L. & Bennett, S. M. Direct comparison of the generalized visual analog scale and general labeled magnitude scale. *Food Qual. Prefer.* **28**, 36–44 (2013).
77. Kumar, S., Forster, H. M., Bailey, P. & Griffiths, T. D. Mapping unpleasantness of sounds to their auditory representation. *J. Acoust. Soc. Am.* **124**, 3810–3817 (2008).
78. Kumar, S., von Kriegstein, K., Friston, K. & Griffiths, T. D. Features versus feelings: dissociable representations of the acoustic features and valence of aversive sounds. *J. Neurosci.* **32**, 14184–14192 (2012).
79. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. International affective picture system (IAPS): affective ratings of pictures and instruction manual. *Technical Report A-8. University of Florida*, *Gainesville* (2008).
80. Op de Beeck, H. P. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage* **49**, 1943–1948 (2010).
81. Genovese, C. R., Lazar, N. A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870–878 (2002).
82. Thompson, B. & Borrello, G. M. The importance of structure coefficients in regression research. *Educ. Psychol. Meas.* **45**, 203–209 (1985).
83. Parra, L. C., Spence, C. D., Gerson, A. D. & Sajda, P. Recipes for the linear analysis of EEG. *Neuroimage* **28**, 326–341 (2005).
84. Kohoutová, L. et al. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* https://doi.org/10.1038/s41596-019-0289-5 (2020).
85. Mumford, J. A. & Nichols, T. Simple group fMRI modeling and inference. *Neuroimage* **47**, 1469–1475 (2009).
86. Bota, M. & Swanson, L. W. BAMS Neuroanatomical Ontology: design and implementation. *Front. Neuroinformatics* **2**, 2 (2008).

Acknowledgements

We thank D. Ott, J. Griffin, E. Biringen and T. Wilkes for assistance with data collection; P. Gianaros for sharing data included in study 4 and R. Stark for sharing data included in study 6; and R. Botvinik-Nezer, K. Zorina-Lichtenwalter and B. Petre for helpful comments on earlier versions of the manuscript. This work was funded by NIH R01DA035484 (to T.D.W.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.C., C.-W.W., M.L.-S. and T.D.W. conceived and designed the experiment for studies 1 and 2, and P.A.K. and T.D.W. conceived and designed the experiment for study 3. M.C. and C.-W.W. collected and preprocessed the data for studies 1 and 2, and P.A.K. collected and preprocessed the data for study 3. M.C., P.A.K. and T.D.W. analyzed the data and interpreted the results. M.C. created the figures, with intellectual input from all other authors. M.C. and T.D.W. wrote the manuscript. All authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

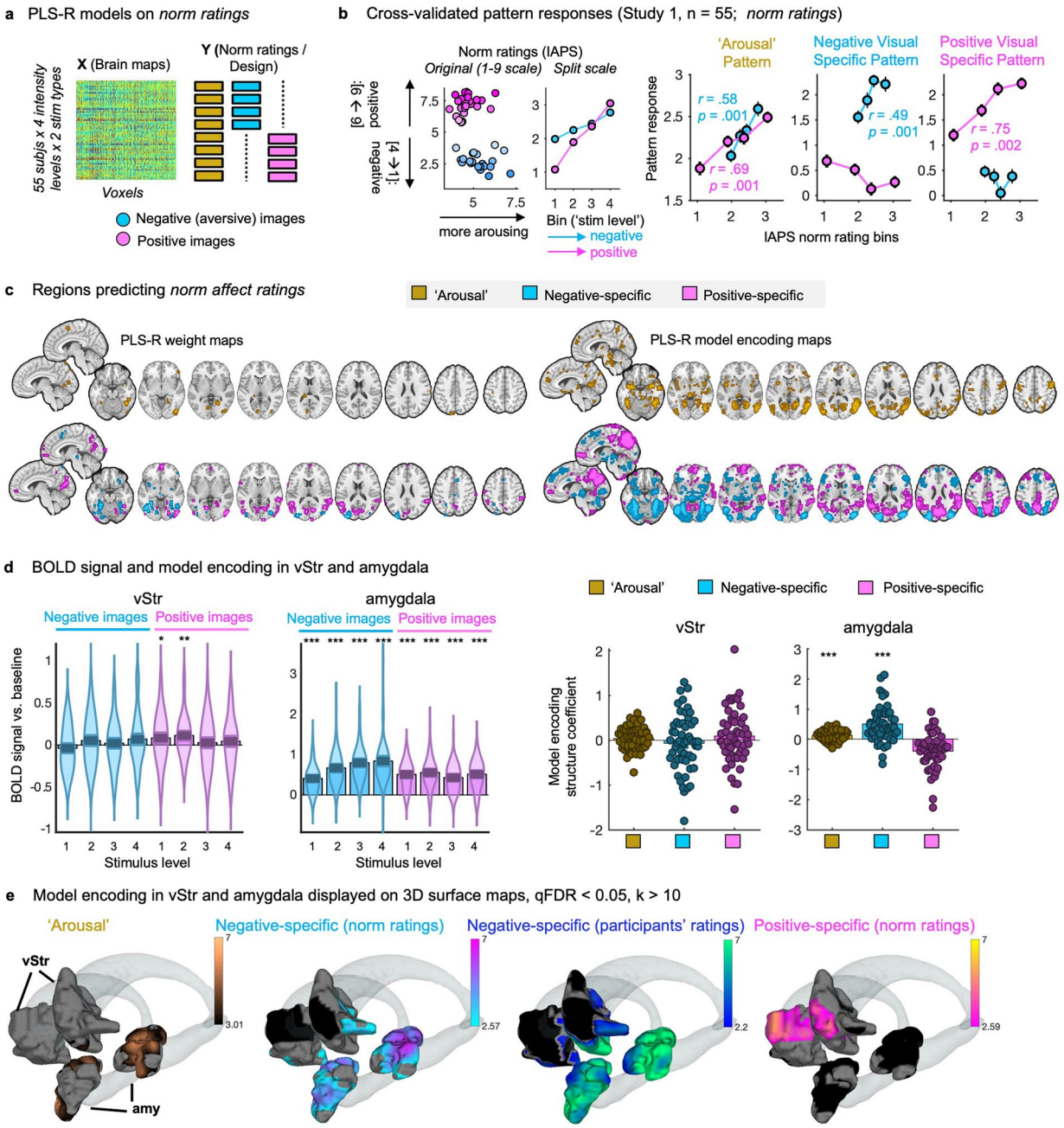
Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01082-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01082-w.

Correspondence and requests for materials should be addressed to Marta Čeko or Tor D. Wager.

Peer review information *Nature Neuroscience* thanks Junichi Chikazoe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | PLS-R models trained to predict normative ratings to negative (aversive) and positive (pleasant) IAPS images. (a) PLS-R procedure to estimate brain patterns for 'arousal' (common across stimuli) and for stimulus type-specific outcomes (IAPS norm ratings) simultaneously **(b)** *Behavior plots*. **Left:** normative ratings shown for each individual stimulus (that is, IAPS image); original IAPS scales (1–9 scales for Valence (higher score = less negative / more positive; 0 is neutral) and Arousal (higher score = more arousing). **Right**: norm ratings averaged per bin ('stimulus intensity level', used for PLS-R training) and shown on a 0–4 split scale (higher score = more negative / more positive; 0 is neutral); *Pattern response plots*. Relationship between observed and predicted ratings. Circles reflect mean values across participants for each stimulus type, error bars reflect within-participant SEM. 'Arousal' model (panel 1), trained on all stimuli, significantly predicted ratings across stimulus types. Stimulus type-specific models (panels 2–3) significantly predicted ratings to target (color-matched), but not off-target stimulus type. *r*, mean within-participant Pearson correlation between predicted and observed ratings; two-sided P-values based on a 10,000 samples bootstrap test of within-participant *r* values**.** **(c) Left**: PLS-R model weight maps showing which brain areas make a reliable contribution to each model's prediction (based on bootstrapping with 10,000 samples and displayed here at $t > 3$, retaining positive values). **Right**: Model encoding maps showing where in the brain voxel-wise activity correlates with PLS model outcomes, corrected for multiple comparisons using $q < 0.05$ FDR and thresholded at $t > 3$, retaining positive values. **(d)** Violin plots showing average BOLD response per stimulus intensity (x-axis) in bilateral ventral striatum (vStr) and amygdala ROIs (Supplementary Table 7), * $p = 0.047$, ** $p = 0.002$, * $p < 0.001$ (left panels); Mean structure coefficient values for each model, averaged across in-ROI voxels across both hemispheres, * $p < 0.001$, only p-values associated with positive t-values are marked and interpreted, each dot is a participant (right panels); one-sample t-test on $n = 55$ participants, treating participant as random effect, bars reflect mean values across participants for each stimulus type, error bars reflect within-participant SEM.**(e)** 3D surface maps of vStr and amy are displaying FDR-corrected model encoding maps for PLS 'norm' models of positive and negative images, and for the PLS model trained on participants' ratings of negative images (Analysis 1).

nature portfolio

Corresponding author(s): Marta Ceko, Tor Wager

Last updated by author(s): April 12, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Studies 1, 2, and 3: Psychtoolbox (version 3.0, http://psychtoolbox.org/) running on Matlab (version 2017a, Mathworks). For Studies 4 - 6 please refer to the original publications listed in Supplementary Table 8.
Data analysis	FMRI data were analyzed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm) and custom Matlab (MATLAB 2019a, The MathWorks, Inc., Natick, MA) code available at https://github.com/canlab/2021_Ceko_MPA2_Aversive. A univariate general linear model (GLM) was used to create images for the prediction analyses. The subject-level GLM analyses were conducted in SPM12.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Brain patterns generated and analyzed during the current study, as well as source data for figures are freely available via https://github.com/canlab/2021_Ceko_MPA2_Aversive. The dataset used in Study 6 is available at https://github.com/cocoanlab/interpret_ml_neuroimaging.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A recommended sample size for the prediction of within-person effects using group-level data is > 50 participants (Lindquist et al. 2017).
Data exclusions	No data were excluded from the analyses.
Replication	The main findings reflect effects that are reliable across 5 different independent studies.
Randomization	All subjects completed two fMRI sessions (only one session is reported in the current paper). We pseudo-randomized (counterbalanced) the order of experimental conditions and of sessions to minimize order effects. There was no significant effect of session order, as reported in Supplementary Table 4.
Blinding	No blinding was necessary for the study because there were not between-subject factors in our study, therefore investigators were not blinded to group assignment. No blinding was performed for data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Study 1: N = 55 participants, Study 2: N = 32 participants, Study 3: 32 participants, Study 4: 182 participants, Study 5: 95 participants, Study 5: 59 participants. Age and gender of participants for each Study are detailed in Supplementary Table 8.
Recruitment	Participants for Studies 1, 2 and 3 were recruited from the Boulder/Denver Metro Area via flyers posted around the University of Colorado Boulder campus, local Boulder coffeeshops and the library, and on-line (university bulletin, Craigslist). The majority of the study sample thus stem from the university undergraduate population. Participants were healthy, with normal or corrected to normal vision and normal hearing, and with no history of psychiatric, physiological or pain disorders and neurological conditions, no current pain symptoms, and no MRI contraindications. Eligibility was assessed with a general health questionnaire, a pain safety screening form, and an MRI safety screening form.
Ethics oversight	The institutional review board of the University of Colorado Boulder approved the study, and all participants provided written consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Block design

Design specifications	Task details for each of the four studies are listed in Supplementary Table 8, and include: stimulus types, intensity levels, stimulus duration, and number of trials per stimulus type.
Behavioral performance measures	Behavioral outcome was the rating of aversiveness (avoidance) for each stimulus type, rated on a general Labeled Magnitude Scale (gLMS) scale ranging from 0-10. Mean within-participant ratings ± SE for each stimulus level are listed in Supplementary table 1 and in Figure 2A.

Acquisition

Imaging type(s)	Functional
Field strength	3 Tesla
Sequence & imaging parameters	Multiband EPI sequence (TR = 460 ms, TE = 27.2 ms, field of view = 220 mm, multiband acceleration factor = 8, flip angle = 44°, 64 x 64 matrix, 2.7 x 2.7 x 2.7 mm voxels, 56 interleaved ascending slices, phase encoding posterior >> anterior).
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	FMRI data were preprocessed using an automated pipeline based on AFNI, FSL, and SPM5, and implemented by the Mind Research Network, Albuquerque, NM. Preprocessing steps included: distortion correction using FSL's top-up tool (https://fsl.fmrib.ox.ac.uk/fsl/), motion correction (affine alignment of first EPI volume (reference image) to T1, affine alignment of all EPI volumes to the reference image and estimation of the motion parameter file (sepi_vr_motion.1D, AFNI, https://afni.nimh.nih.gov/), spatial normalization, interpolation to 2 × 2 × 2 mm3 voxels and smoothing with a 6 mm FWHM kernel (SPM 8, https://www.fil.ion.ucl.ac.uk/spm/software/spm8/).
Normalization	Spatial normalization via participant's T1 image (T1 normalization to MNI 152 space (nonlinear transform), normalization of EPI image to MNI 152 space (3dNWarpApply, AFNI, https://afni.nimh.nih.gov/)
Normalization template	MNI 152, group standardized space
Noise and artifact removal	Prior to first level analysis, in each run we removed the first 15 volumes of the fMRI data to allow for image intensity stabilization and identified severe global motion outliers (spikes) in the data. Spikes were defined as time points in the data with either the absolute global signal intensity or the mahalanobis distance across slice-specific global means and spatial standard deviations exceeding 10 median absolute deviations.
Volume censoring	No censoring, instead inclusion of motion spikes as regressors in the first-level model

Statistical modeling & inference

Model type and settings	Brain models predicting stimulus-type specific aversiveness ratings and common (across all 4 stimulus types) negative affect ratings were developed using the multivariate partial least squares regression (PLS-R) on Study 1 data.
Effect(s) tested	To test the predictive performance of each PLS brain model, we calculated in each participant the root mean squared error (RMSE) for each model's predictions of its target outcome (i.e., four average aversiveness ratings per stimulus type) and the ability to significantly predict increasing aversiveness within-participant. To provide an interpretable effect size metric, we estimated in each participant the Pearson correlation (r) between observed and cross-validated predicted ratings and tested whether the output of one model predicted the specific type of negative affect it was trained to predict (sensitivity) and not other types (specificity). To determine which brain areas made reliable contributions to the prediction and to threshold voxel weights for interpretation and display, we constructed 10,000 bootstrap samples (with replacement) consisting of paired brain and outcome data and performed PLSR on each. The z-scores at each voxel were estimated based on the mean and standard error of the bootstrap distributions, and the statistical map was thresholded based on the corresponding p-values. The maps were thresholded voxel-wise at q < FDR-corrected.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	Anatomical parcellations and the corresponding atlases used to select ROIs are detailed in Supplementary Table 7.
Statistic type for inference (See Eklund et al. 2016)	Voxel-wise
Correction	FDR correction

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

PLSR was conducted using plsregress.m in MATLAB (version R2016b). Predictors (X) constituted 880 whole-brain activation maps associated with participants (55) x stimulus intensity level (4) x stimulus type (4), aggregated into an images x voxels matrix (stacked across participants), and split into training and test sets using 5-fold blocked cross-validation (leaving out all images associated with test-set participants together). The activation maps were derived via univariate GLM analysis in which we modeled stimulation periods against implicit baseline and averaged across trials of each stimulus level of each stimulus type to obtain 16 contrast images per participant. To estimate their predictive accuracy, and specificity for the target outcome, and generalization to new individual participants, these patterns were applied to fMRI activation maps obtained from new participants in cross-validation test sets and prospectively applied to independent studies (N = 247). Model (pattern) responses were calculated using the dot product of each pattern weight map with the univariate GLM-derived fMRI activation map for each participant for each stimulus level. Weight maps applied to Study 1 and Study 1b were based on data from out-of-sample individuals (cross-validated estimates), and the final pattern weights applied to Studies 2-4 were based on the full study sample (full samples estimates).