Neuroscience and Biobehavioral Reviews 131 (2021) 211–228

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Predictive processing models and affective neuroscience

Kent M. Lee [a,*], Fernando Ferreira-Santos [b], Ajay B. Satpute [a]

[a] Northeastern University, 360 Huntington Ave, 125 NI, Boston, MA 02118, USA
[b] Laboratory of Neuropsychophysiology, Faculty of Psychology and Education Sciences, University of Porto, Portugal

ARTICLE INFO

Keywords:
Predictive processing
Predictive coding
Subjective experience
Ecological validity
External validity
Valence
Degeneracy
Reverse inference
Experimental design
fMR
Arousal
Emotion
MVPA

ABSTRACT

The neural bases of affective experience remain elusive. Early neuroscience models of affect searched for specific brain regions that uniquely carried out the computations that underlie dimensions of valence and arousal. However, a growing body of work has failed to identify these circuits. Research turned to multivariate analyses, but these strategies, too, have made limited progress. Predictive processing models offer exciting new directions to address this problem. Here, we use predictive processing models as a lens to critique prevailing functional neuroimaging research practices in affective neuroscience. Our review highlights how much work relies on rigid assumptions that are inconsistent with a predictive processing approach. We outline the central aspects of a predictive processing model and draw out their implications for research in affective and cognitive neuroscience. Predictive models motivate a reformulation of "reverse inference" in cognitive neuroscience, and placing a greater emphasis on external validity in experimental design.

1. Introduction

Research in affective neuroscience has spent considerable effort searching for the brain bases of affective dimensions such as valence and arousal (Barrett and Bliss-Moreau, 2009; Baucom et al., 2012; Berridge, 2019; Bush et al., 2017; Colibazzi et al., 2010; Lewis et al., 2007; Mather et al., 2016; Miskovic and Anderson, 2018; Phan et al., 2002; Tye, 2018). However, a growing body of work suggests that there is not a dedicated neural system that consistently and uniquely decodes these dimensions (Chikazoe et al., 2014; Lindquist et al., 2016; Miskovic and Anderson, 2018; Satpute et al., 2019, 2015). Here, we suggest that progress has been slowed because most work in affective neuroscience assumes that functional activation patterns that underlie valence and arousal are uniform (across context), specific (involves unique circuits), and generalizable. We refer to this view as a **simple feature detector model** of affective experience in that certain patterns of activation are expected to be "on" or "off" when pleasure (or displeasure) is present vs. absent. Meanwhile, research in cognitive neuroscience has increasingly adopted predictive processing models (Keller and Mrsic-Flogel, 2018; Köster et al., 2020; Pereira et al., 2019; Ransom et al., 2020; Stawarczyk et al., 2019). In contrast to traditional approaches in affective neuroscience, predictive processing models challenge assumptions of a stable and unique neural signature for affect.

In this paper, we discuss the consequences of examining the relationship between brain activity and affective experience if we assume that the brain is running a predictive processing model. Our goal is to explore the implications of predictive processing models for how research in affective neuroscience is done. It is not our goal to develop a specific model of predictive processing that describes how predictions and prediction errors generate subjective experience; researchers have provided well-developed accounts elsewhere (Allen and Friston, 2018; Barrett, 2017; Hesp et al., 2019; Hutchinson and Barrett, 2019; Smith et al., 2019b). Instead, we focus on discussing the features of predictive models and their practical and concrete implications for how to study the neural basis of affective experience in terms of experimental design, data analysis, and theoretical interpretation. We focus our review on the human functional neuroimaging literature but note that our points may extend to other modalities, too (e.g., EEG/ERP).

We start with a brief review of research on the neural bases of affective experience with respect to macrolevel functional architecture. We then provide a review of predictive processing models in neuroscience while simultaneously weaving in their implications for research practice in affective neuroscience. We further address implications of taking a predictive processing approach for making "reverse inferences"

* Corresponding author at: Department of Psychology, Northeastern University, Boston, MA 02118, USA.
E-mail address: ke.lee@northeastern.edu (K.M. Lee).

https://doi.org/10.1016/j.neubiorev.2021.09.009
Received 30 April 2019; Received in revised form 10 February 2021; Accepted 7 September 2021
Available online 10 September 2021

in cognitive neuroscience (Poldrack, 2015, 2011, 2006) and for understanding internal vs. external validity. While our focus is on the neuroscience of affective experience, our conclusions may also apply more widely across many domains in cognitive neuroscience that aim to link subjective experience with neural activity.

2. The elusive neural signatures of affective experience

Most research in affective neuroscience starts with the premise that there are reliable and selective patterns of neural activation, or **neural signatures,** that underlie affective dimensions. For example, many researchers have attempted to identify a specific pleasure system in the brain (Berridge and Kringelbach, 2015; Costa et al., 2010; Kringelbach and Berridge, 2009; Olds, 1956; Sabatinelli et al., 2007; Wise, 1980; see Table 1 and Fig. 3. Such putative neural signatures of affect are often assumed to be invariant across time, context, and individuals. In doing so, most affective neuroscience studies assume a model in which the brain systems for affective processing operate like a simple feature detector in which there are reliable and specific patterns of activation in certain brain regions that correlate with variation in dimensions such as valence and arousal. As we review in this section, the supposed neural signatures for affective experience have remained elusive, which raises questions about the utility of the simple feature detector model for affect. Most of this work has examined the valence dimension, and so we focus our review accordingly.

2.1. Activation reliability in univariate neuroimaging studies of valence

There are now hundreds of functional neuroimaging studies that have examined the neural basis of valence (Lindquist et al., 2016). In the typical study (see Fig. 1A), participants are presented with dozens of affect inducing images (e.g., Canli et al., 1998; Lang et al., 1998). The images themselves often vary in terms of semantic content (e.g., pictures of bodily injuries, pollution, snakes, remarkable athletic feats, cute animals), and each image is shown for just a few seconds at a time. Functional activity is averaged across trials that evoke pleasure and compared against those that evoke displeasure or a neutral state to identify brain regions with functional activity that varies by valence conditions. Activity is also averaged across participants to identify group-level activation patterns. These analytical choices inherently assume invariance of responses across context (e.g., stimulus content), time (trials), and participants (see Fig. 1B). These choices make sense if one assumes a simple feature detector model for affective processing. If there is a neural signature for valence that operates like a simple feature detector, then there should be a reliable set of brain regions that are active during pleasure or for displeasure, irrespective of the context that triggers these feelings.

Meta-analyses of neuroimaging studies in affective neuroscience do not provide strong support for a simple feature detector view. In a recent meta-analysis that included results from 397 studies, the most reliably activated brain region during pleasant vs. neutral task conditions (the amygdala) involved fewer than 26 % of the contributing contrasts, and similarly for unpleasant vs. neutral contrasts (also the amygdala; see Fig. 2A; Lindquist et al., 2016). Most other areas that met statistical reliability only had half as many contributing contrasts (approximately 10–15 %). This low reliability could be due to low power. Indeed, a comprehensive examination of reliability of many functional neuroimaging paradigms in social, affective, and cognitive neuroscience found that reliability was ubiquitously low even at the level of individual experiments (e.g., average intra-class correlation coefficient across paradigms was $r = .40$; Elliott et al., 2020). Correspondingly, individual experiments in affective neuroscience might also be underpowered resulting in an overall low-reliability in the meta-analysis.

But the causes, and therefore solutions, to low reliability are not straightforward. A knee-jerk response to issues of reliability is to increase statistical power by collecting more data (more participants and time on task) or using better imaging equipment. However, as reported in Elliott et al.'s (2020) review, reliability in fMRI was not fully explained by task length, task type, scanner quality, or even sample size. Another possibility is that higher quality task paradigms (e.g., more evocative affect inductions) might be helpful. This could be the case, but even so, a focus only on these factors assumes that the underlying theoretical model is correct and the only problem is the signal-to-noise ratio. However, another reason for low reliabilitymight be that the traditional theoretical approach, which implies a simple feature detection model for relating mind and brain, does not fully account for how

Table 1
Classification accuracies for fMRI MVPA studies examining valence and arousal.

Study	Stimulus Modality	Valence			Arousal			Brain Space
		Within Subject	Across Subjects	Across Stimuli	Within Subject	Across Subjects	Across Stimuli	
Baucom et al., 2012	V	~75 %	~70 %	–	~75 %	~75 %	–	Whole
Skerry and Saxe, 2014	V	54 %	–	52 %	–	–	–	ROI
Chikazoe et al., 2014	V, G	–	55 %	54 %	–	–	–	ROI
Shinkareva et al., 2014	V, A	~61 %	–	*ns*	–	–	–	Whole
Kim et al., 2016	V, A	66 %	61 %	–	60 %	61 %	–	Mixed
Bush et al., 2017	V	*ns*	59 %	–	*ns*	56 %	–	Whole
Kim et al., 2017	V, A	–	–	~63 % / ~52 %	–	–	–	Whole / Searchlight
Bush et al., 2018	V	56 % / 85 %	–	–	61 % / 78 %	–	–	Whole
Kim et al., 2020	V, A	–	r = 0.158	–		*ns*		Searchlight
Shinkareva et al., 2020	V, A	–	72 %	60–90%	–	–	–	Whole

Note. A summary of classification accuracy findings from current fMRI MVPA studies on valence and arousal. Accuracy ranges between slightly above chance (50 % in most cases) to high accuracy (up to 85 %) depending on the study. Classification accuracies are summarized by whether analyses and cross-validation were conducted within subjects, across subjects, and across stimulus modality. '-' indicates that classification accuracy of that type was not reported, and 'ns' indicates that the analysis was performed but was not statistically significant. Notably, Bush et al. (2017) found nonsignificant classification within subjects but significant classification across subjects, which they suggest occurs when the algorithm learns idiosyncratic features of the training dataset that ultimately do not generalize upon cross validation (referred to as "anti-learning"). Bush et al. (2018) conducted separate analyses on a full set of stimuli and a subset of stimuli (suggesting that the model works well on normative, but not as well on subjective, affective experiences). Kim et al. (2017) conducted MVPA to classify valenced stimuli across stimulus categories across the whole brain (left of slash) and in specific regions identified in a searchlight analysis conducted to localize modality-general representation of valence (right of slash). Shinkareva et al. (2020) reported an average classification accuracy (across subjects) of 72 % and that classification accuracies across studies (which used stimuli of different modalities) ranged from 60 % to 90 % accuracy. For stimulus modality V = visual, A = auditory, G = gustatory.

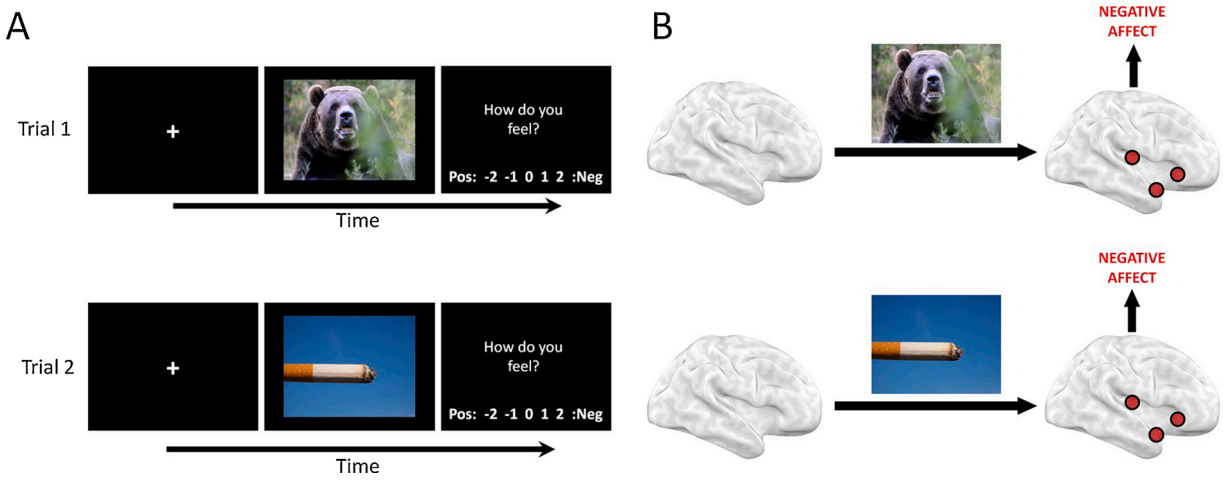


Fig. 1. The typical design and analysis of factorial design experiments in affective neuroscience and their relationship with a simple feature detection model of subjective phenomena.
(A) To examine the neural basis of subjective experiences of affect and emotion, the typical experimental design involves presenting participants with a sequence of evocative images and obtaining self-report ratings of valence or arousal. In a univariate analysis, it is common practice to average BOLD signal across trials that evoke the same affective response (e.g. into positive, neutral, and negative categories) and also across participants in a sample. Doing so assumes that the brain regions supporting affective experience are invariant across time/trials of the same condition, and across participants. **(B)** The assumptions of this approach are presented as a "simple feature detection model" for how brain activity is associated with affective experience. Negative affect evoked by different evocative stimuli at different moments in time assumed to evoke the same activation pattern. The brain state prior to the stimulus is typically assumed to be noise (represented brains without activation patterns on the left) and thus lays relatively dormant. As an aside, when examining subjective experiences of affect, different stimuli may evoke different affective feelings (i.e. not everyone may feel the same way toward images of cigarettes or aggressive bears). Researchers account for this variance by using subjective ratings to group trials together into conditions. Photograph of bear adapted from "grizzly bear" by S. Kringen, 2010 (https://www.flickr.com/photos/18161271@N00/4957154697). CC BY-SA 2.0. Photograph of cigarette adapted from "cigarette" by Fried Dough [screen name], 2011 (https://www.flickr.com/photos/42787780@N04/6447342961). In the public domain.

the brain functions (i.e., in a predictive fashion; see Section 3).

In addition to reliability concerns, another issue remains: there is also little evidence of *selectivity* in the brain to affective valence, at least when examining individual brain regions (Lindquist et al., 2016). In the context of affective (and cognitive) neuroscience, **selectivity** would be reflected in the ability of an indicator (e.g., neural activity) to discriminate between mental states of interest (e.g., pleasure or displeasure). Indeed, brain regions that are engaged during negative (vs. neutral) affect were also often engaged during positive (vs. neutral) affect as shown in a meta-analysis (Lindquist et al., 2016; see Fig. 2A and B) and in individual experiments (Bonnet et al., 2015; Chikazoe et al., 2014). Of course, many of these areas may simply be responsive to arousal irrespective of valence since arousal is typically higher for both positive and negative valence stimuli relative to neutral stimuli in these studies. However, we also failed to find evidence that any areas were selectively engaged during positive valence or negative valence.

There are two common responses to this issue. First, it has been argued that fMRI may not have sufficient resolution to resolve valence. Neurons that are sensitive and specific to valence (if they exist) may be interdigitated within a single region or even voxel (Tye, 2018). However, it remains unclear whether putative neurons are indeed valence specific when it comes to subjective experience (see Rigotti et al., 2013 for a discussion of how individual neurons may encode multiple disparate psychological dimensions), and even so, whether they respond in ways that are invariant across time and context (as is often assumed). Second, it has been argued that valence representations are not confined to activation magnitude in a single or collection of brain regions, but instead involve patterns of activation across voxels that may span across brain regions (Baucom et al., 2012; Bush et al., 2017; Chikazoe et al., 2014; Kassam et al., 2013; Satpute et al., 2015). To address this possibility, researchers have turned to multivariate analysis approaches which examine whether patterns of activation across brain regions carry information about valence, and which also take steps toward addressing selectivity in neuroimaging analysis. We address these methods next.

2.2. Classification accuracy in multivariate neuroimaging studies

Multivariate pattern analysis (MVPA) refers to a family of analytical techniques that tests whether the distribution of activation levels across multiple spatial locations (sets of voxels, brain regions, etc.) that may be used to distinguish between different conditions within a task or between tasks (Haxby, 2012; Kriegeskorte et al., 2006; Norman et al., 2006). An advantage of MVPA is that even when regional "activation based" approaches show no difference in the overall magnitude of activation, the distributed pattern of activation may nonetheless contain information that enables an algorithm to distinguish between task conditions. An algorithm may be trained to classify whether a given activation pattern is more likely to occur during task conditions that induce pleasure, displeasure, or neutral states. For example, the orbitofrontal cortex may not show a global difference in activation by valence when conducting a univariate analysis that averages the signal across all voxels spanning its territory (Chikazoe et al., 2014; Lindquist et al., 2016). However, certain voxels in the orbitofrontal cortex (OFC) may show a heightened response, others no response, and still others a diminished response when experiencing pleasure (Chikazoe et al., 2014). This pattern of activation across voxels of the OFC may be more similar to that observed during other trials that evoke pleasure, and dissimilar to trials that evoke displeasure. As a general caveat to keep in mind, while MVPA can be used to classify distinct task conditions, the underlying reasons for successful classification is subject to experimental interpretation. For example, differences in semantic content between positive and negative stimuli may also drive successful classification yet have little to do with feeling pleasure or displeasure (Hamzani et al., 2020; Itkes et al., 2017).

To test whether patterns of activity contain information about valence, researchers have used two strategies (Haxby, 2012; Kriegeskorte et al., 2006; Weaverdyck et al., 2020). One is to test for

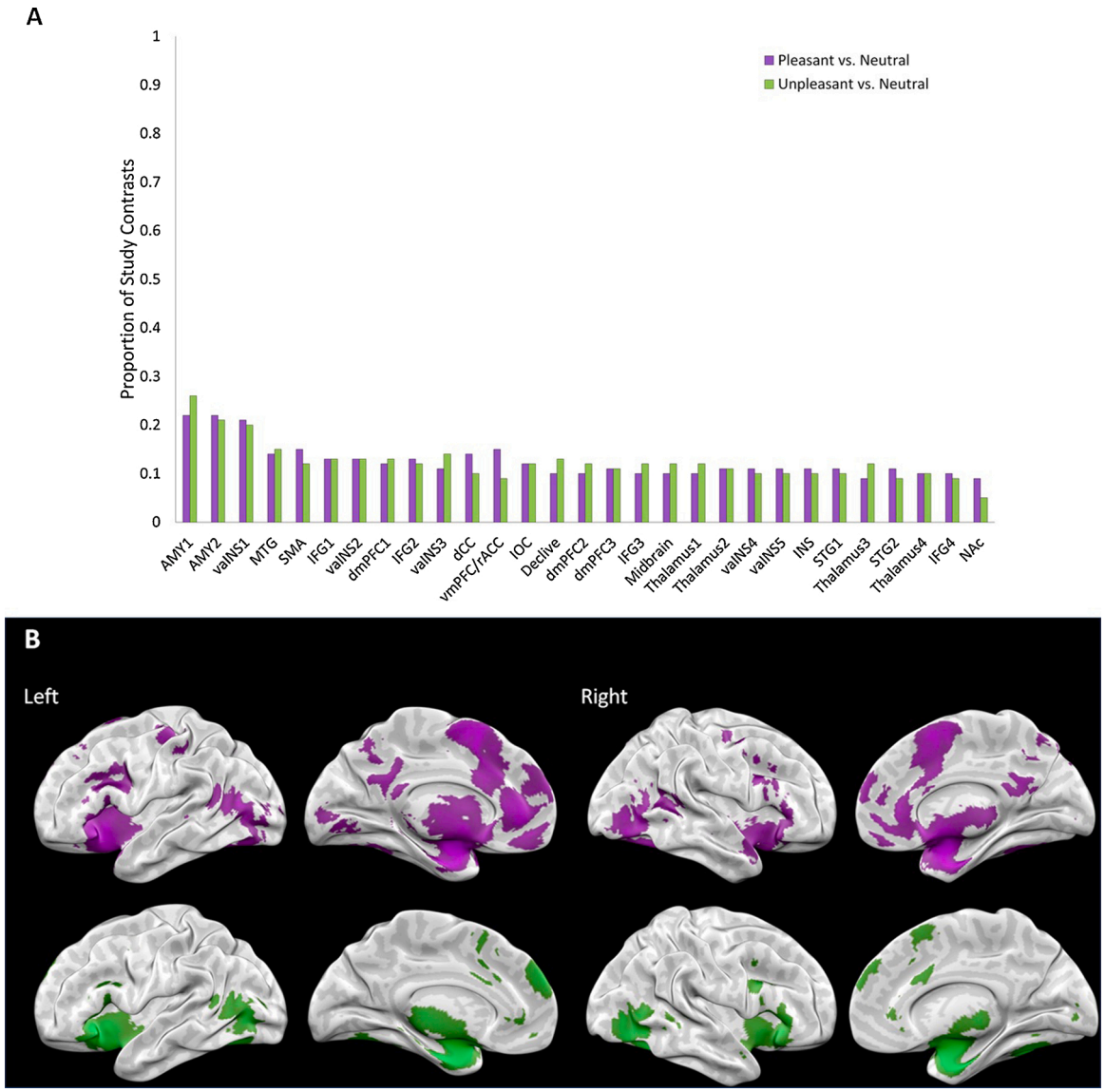


Fig. 2. Meta-analytic results from Lindquist et al. (2016).
(A) shows the proportion of study contrasts showing activation in each brain region for pleasant (purple) and unpleasant (green) valence. Areas that showed multiple peaks are numbered (see Supplementary Table 1 for coordinates of global and local maxima). Inconsistent with a simple feature detector view, the areas most frequently associated with affective valence showed little reliability across studies. Activation in the most reliable area, the amygdala, was observed in 26 % of study contrasts at most. AMY = amygdala, vaINS = ventral anterior insula, dmPFC = dorsomedial prefrontal cortex, vmPFC = ventromedial prefrontal cortex, dCC = dorsal cingulate cortex, SMA = supplementary motor area, MTG = middle temporal gyrus, IFG = inferior frontal gyrus, IOC = inferior occipital cortex, INS = insula, STG = superior temporal gyrus, NAc = nucleus accumbens. **(B)** shows brain areas consistently associated with pleasant (purple; top row) and unpleasant (green; bottom row) valence. Top row shows areas consistently activated across 110 positive > neutral contrasts. The bottom row shows areas consistently activated across 255 negative > neutral contrasts. There is notable overlap between areas associated with unpleasant and pleasant valence. In particular, there is a high degree of overlap in temporal and some limbic areas.

representational similarity (or dissimilarity), by examining whether how similar or distinct one neural pattern is from another (e.g., by calculating a Pearson correlation of voxel activities between patterns) and comparing that neural similarity metric with behavioral ratings of valence. The other is to test for pattern classification by using a machine learning algorithm that learns to classify valence from the activation pattern. Scaling up, the same analysis could be conducted across multiple brain regions, or even looking at similarity when including all voxels in the brain.

The simple feature detector model assumes that there will be a pattern of activation that reliably and specifically decodes affective dimensions. Consistent with this model, MVPA studies in affective neuroscience typically assume that activation patterns are invariant across trials (time) and to an extent, across contexts and participants. However, MVPA offers some flexibility in testing these assumptions. Researchers can test invariance across participants by assessing classification performance when the model is trained on data from one sample of participants and tested on another set of participants (Table 1, "across subjects"). Researchers can also test invariance across contexts by training the model on one set of stimuli (e.g., pictures) and testing on another set of stimuli (e.g., other pictures not in the training set, or even stimuli from another stimulus modality; Table 1, "across stimulus types"; Chikazoe et al., 2014; Kim et al., 2017; Shinkareva et al., 2020).

Most of the studies summarized in Table 1 do show above chance classification accuracies for valence and arousal. However, the findings also vary considerably. In terms of the brain regions identified in these studies, there are some consistent findings such as the anterior medial prefrontal cortex and limbic areas, but also many scattered findings that

were not consistent across most studies (see Fig. 3). Moreover, accuracy of the classifications was highly variable (See Table 1). Three studies show relatively higher levels of accuracy (>70 %; Baucom et al., 2012; Bush et al., 2018; Shinkareva et al., 2020) but others are only slightly above chance levels (52–66 %, with chance being 50 %; Bush et al., 2017, 2018; Chikazoe et al., 2014; Kim et al., 2017, 2016; Skerry and Saxe, 2014). At first glance, one might assume that high classification accuracy is attributable to study quality or power, but a closer look suggests otherwise. For example, a relatively high classification accuracy (~75 %) was observed in one study, but the experimental design also used repeated stimulus presentations (Baucom et al., 2012). Affective feelings may habituate with just three to four repetitions in fMRI task paradigms (Satpute et al., 2016), and so it is unclear whether classification is related to feelings *per se*. Indeed, affective judgments can be equally driven by the semantic evaluation (Itkes et al., 2017) or memory that something is pleasant (Robinson and Clore, 2002), rather than the feeling of pleasantness.

In another study, valence classification accuracy was 85 % but only for certain stimuli that had high classification performance in the training sample (Bush et al., 2018). However, restricting the stimulus set in this way runs counter to the presumed goal of identifying stable patterns of activation that predict latent, subjective experiences of affect that generalize across stimuli. When attempting to cross-validate the model without constraining the stimulus set, accuracy dropped to 56 %. In a study optimized to examine subjective affective experience, in which stimuli were only presented once and classification used participants' subjective responses, classification levels were on average 55 % (Chikazoe et al., 2014). Taken together, these findings suggest that obtaining much higher classification accuracies in studies of valence requires using more artificial constraints in the design or analysis.

In the most robust examination of this question to date, Shinkareva and colleagues conducted a cross-study analysis using data from six experiments. They were able to address lingering questions about generalization by including data from different samples, scanning sites, and affect induction tasks (auditory or visual affect inductions, different stimuli, etc.; Shinkareva et al., 2020). MVPA models were trained using subject-level maps of valence and arousal from five of the studies and tested on the left out study. Consistent with the assumptions of the simple feature detection model, the analysis assumed invariance across trials and stimulus contents (trials were averaged within pleasant and unpleasant conditions to generate one activation pattern map per participant per condition), and participants (cross-validation was implemented across participant groups). But even so, they reported a fairly high level of classification accuracy of 72 %. These findings are perhaps some of the most encouraging in support of the notion that a distributed pattern of activation may serve as a simple feature detector for valence.

Yet, there remains some limitations; the study design and modeling assumptions are conducted in a highly constrained experimental context. Positive or negative valence are treated categorically and are the only two categories the classification algorithm must contend with. Thus, it is unclear how the model would perform in a more naturalistic and dynamic context in which valence fluctuates more continuously between positive, negative, and also more neutral moments. Taking strides in that direction, Kim and colleagues (2020) have begun to address this question by examining neural predictors of valence ratings when participants watched a lengthy TV show. They observed a significant, albeit modest, prediction of valence ratings from brain activity (average $r = .16$; using leave one subject out validation). This predictive validity is difficult to interpret since normative (rather than subjective) valence ratings were used which themselves had generally low reliability in the study (average interrater correlation, $r = 0.30$). However, this is an important direction of work and among the first of its kind; it would be interesting to see whether the MVPA model trained in more rigid experimental settings similar to those used in the former cross-study analysis, are able to predict valence ratings in more naturalistic experiments similar to this latter study.

There is also one other study that bears mention. Chang et al. (2015) only examined classification of neutral and negative affective states. They obtained a high cross-validated classification accuracy of >90 %, but it is unclear whether this classification is for valence or arousal. Moreover the authors make the point that negative arousal vs. neutral classification is modality dependent, and that they were unable to classify across stimulus modalities (i.e., from using picture stimuli to using pain stimuli) – a point that we return to later when discussing context dependency in predictive processing models below.

2.3. Summary

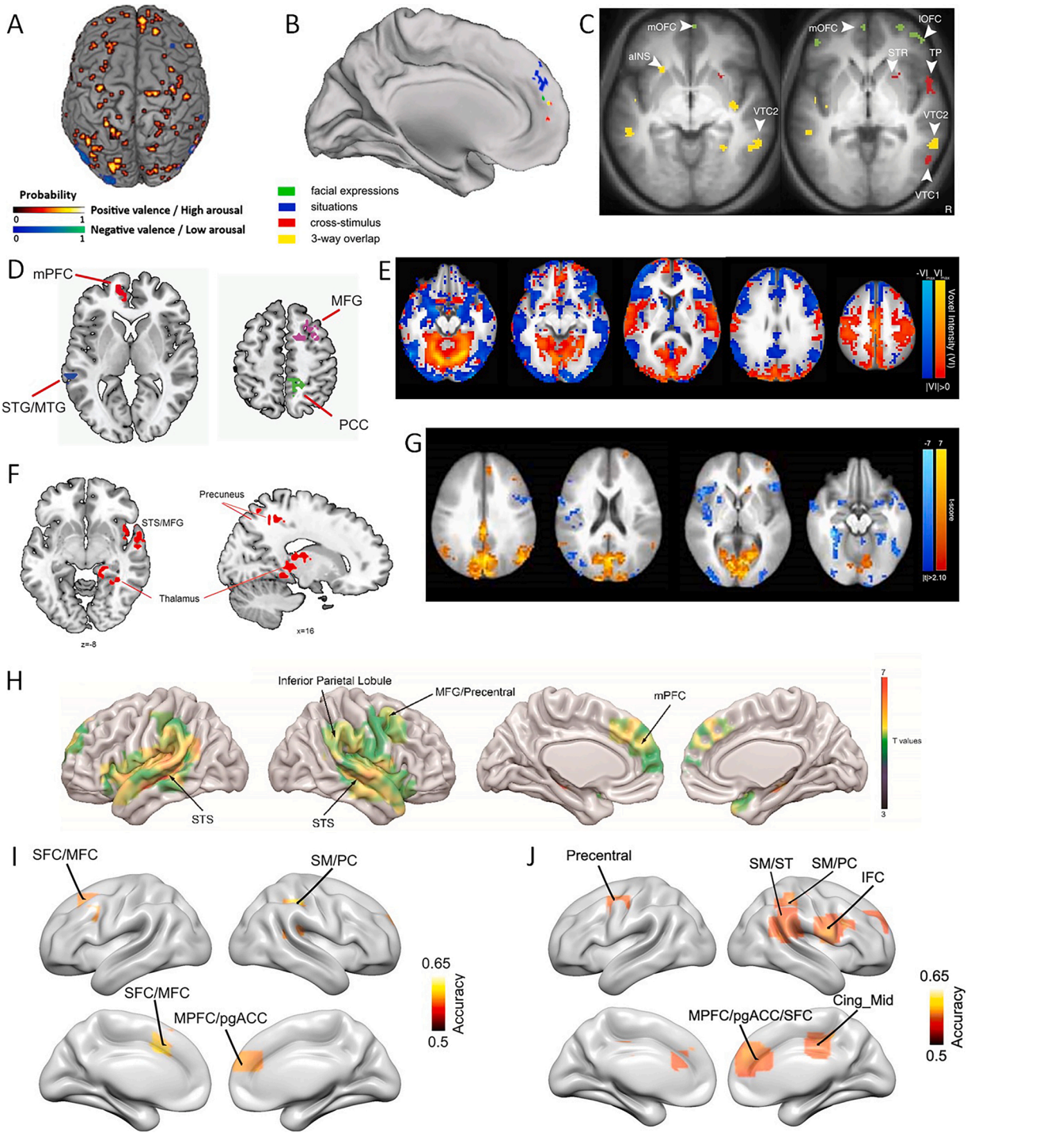
The simple feature detection model of affective experience suggests that there are reliable and specific neural representations for affective dimensions of valence and arousal, and these representations are, by and large, invariant across context (e.g., stimulus content), time (trials), and participants. Studies that use univariate analyses typically makethese assumptions of invariance by averaging findings across these dimensions and assuming residual variance is error. However, these findings have not shown evidence of high reliability or specificity in the neural representations of valence. Multivariate analyses are more flexible in the assumptions they must make, but often make similar assumptions of invariance of functional activity across context and time/trials, and in some instances, invariance across individuals as well (but see Table 1, within vs. across subject columns). Overall, the findings are currently mixed and it is unclear whether this path of work will eventually support something akin to a simple feature detection model of valence. The more successful studies use highly constrained experimental contexts and more rigid analytical assumptions leaving it unclear whether the findings will generalize to modeling the neural basis of affective experiences in everyday life situations.

We suggest that the reason for low reliability may not just be due to low signal-to-noise issues or statistical power. Rather, the underlying theoretical model that guides most analysis in fMRI studies may also be too rigid. Predictive processing models challenge assumptions about invariance of functional activity implied by the simple feature detector model. In doing so, predictive processing offers new directions for affective neuroscience that may enable the field to overcome its current hurdles (e.g., low reliability, mixed evidence for specific neural signatures). In the next section, we provide an overview of predictive processing models in neuroscience and discuss their implications for research in affective neuroscience specifically.

3. Predictive processing models

Predictive processing models originated in part as a computationally efficient data compressing technique in computer science (i.e., predictive coding; Atal, 2006; Elias, 1955). Applied to neuroscience, predictive processing models offer accounts for how information is shared in the brain in a metabolically efficient (Friston, 2010; Rao and Ballard, 1999; Sterling and Laughlin, 2015) and neurobiologically plausible (Bastos et al., 2012) manner. These models posit that the brain does not passively wait to receive stimulation, but instead is continuously making predictions about the future. The brain generates predictions about its future. To the extent that there is a mismatch between those predictions and what occurs, this mismatch is transmitted as prediction error.

In one neural instantiation of this view, the activity in one type of neuron (**prediction units**) encodes the expected activity of other neurons. Prediction units communicate this expected activity to those neurons via top-down (i.e., feedback) connections as **predictions** (see Fig. 4). A second type of neuron (**prediction error units**) computes the mismatch between the predictions and signals generated by sensory inputs as **prediction error**. The prediction error unit relays this signal back to the prediction unit via bottom-up (i.e., feedforward) connections. The arrival of a prediction error indicates that the prediction was



(caption on next page)

Fig. 3. Overview of prior studies using MVPA to examine valence classification.
The brain regions that support valence classification vary considerably from study to study. Certain areas, such as the anterior medial prefrontal cortex may be implicated in many studies, but there are also many areas that are implicated in valence classification for a given study but that are not consistently observed across studies. **(A)** Baucom et al. (2012, Figure 6B) showed a highly distributed pattern of informative voxels for valence classification spread across cortical lobes using a whole brain approach. **(B)** Skerry and Saxe (2014, Figure 6) show significant valence classification across stimulus types (faces and situations) using voxels in a predefined ROI of the anterior medial prefrontal cortex. **(C)** Chikazoe et al. (2014, Fig. 5C) also shows significant valence classification across modalities in predefined ROIs in the anterior medial and lateral prefrontal cortex (green), but also modality-specific valence classification for evocative visual (red) and gustatory (yellow). **(D)** Kim et al. (2016, Fig. 3B) showed valence classification in four brain regions including the perigenual cingulate cortex, superior temporal gyrus, middle frontal gyrus, and precuneus. Kim et al. (2016) examined pre-determined voxels at a group level identified by an initial whole brain searchlight across individual participants. **(E)** Bush et al. (2017, Figure 3) illustrates encoding parameters from a whole-brain SVM of valence. **(F)** Kim et al. (2017, Figure 7) used a whole brain searchlight approach and identified voxels informative for classifying valence located in the insula, superior temporal cortex, precuneus, and thalamus. **(G)** Bush et al. (2018) used a whole brain approach showing valence informative voxels along the cortical midline including the anterior medial prefrontal cortex, poster cingulate and precuneus, and also in the lateral prefrontal and parietal cortex. **(H)** Unlike other studies, Kim et al. (2020, Figure 5) had participants watch a lengthy movie that may preserve more spatiotemporal continuity (albeit used normative valence ratings). Using a searchlight approach, Kim et al. (2020) found that the dorsomedial PFC, insula, and superior temporal cortex and temporoparietal cortex carry information correlating with valence. (**I** and **J**) Shinkareva et al. (2020, Figures 2 and 3, respectively) was the only study to examine MVPA across multiple studies spanning task paradigms, two induction modalities (auditory and visual), different participant groups, and different scanning sites. Informative voxels, identified in a whole brain analysis, were located in the superior frontal gyrus, dorsomedial and cingulate prefrontal cortex **(I)**. They also systematically tested for brain regions that provided information about valence across-modalities and observed activity in some of the same areas **(J)**. Note: Results from Shinkareva et al. (2014) are not depicted since they were conducted only at the subject level, but areas that appear to be informative for classifying valence across participants included the anterior temporal cortex and occipital cortex.

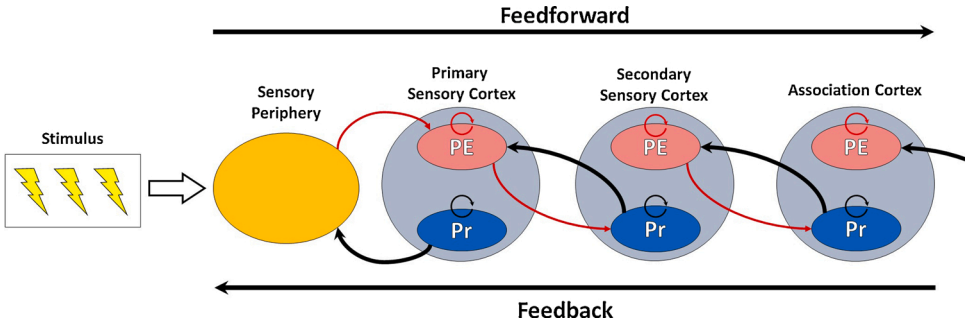


Fig. 4. A schematic depiction of a predictive processing neural architecture.
Contemporary models suggest that predictions and prediction errors are hierarchically organized. Prediction units (**Pr**; blue nodes) located primarily in deep cortical layers send predictions down the cortical hierarchy via feedback connections (black curved arrows). Prediction error units (**PE**; light red nodes) located in superficial layers send prediction errors up the cortical hierarchy via feedforward connections (red curved arrows). Circular arrows above each node reflect precision-weighting. Adapted from Goulas et al. (2018; Figure 1B).

incorrect. The prediction unit can then use this input to update its own activity in order to derive a more accurate prediction of neuronal activity that is conveyed back to the prediction-error unit. If the sensory input remains unchanged, then the updated prediction will effectively minimize prediction-error and no further updates will be required until changes in sensory input elicit novel prediction-errors. Finally, the influence of prediction or prediction error signals is also weighted by the reliability or precision (inverse variance) of those signals. Thus exchanges of predictions and prediction errors enable prediction units to update their activity so as to minimize overall prediction error (Arnal and Giraud, 2012; Friston, 2005).

Predictive processing is thought to be implemented within a loosely hierarchical (or more accurate heterarchical; see Pessoa, 2019) structure. Each level of the hierarchy predicts the activity of the adjacent lower level, while relaying prediction-errors to the immediate upper level (see Fig. 4; Friston, 2008). Predictions and prediction errors are thought to be broader over space and time and more abstract further up versus lower down. Over time, the minimization of sensory prediction-errors may support perceptual learning by entraining prediction units to anticipate the statistical regularities of the environment (Friston, 2005), including the internal environment or the body (Barrett and Simmons, 2015; Seth, 2013). Predictions do not directly predict the statistical regularities of the environment *per se*, rather they reflect statistical regularities in the form of a generative model that predicts future neural activity, ultimately enabling the brain to develop a representation of (its neural activity with respect to) the state of the body and the world (Gładziejewsk, 2016).

It is important to clarify that predictive processing models in neuroscience describe communication between neurons; they are not models of the mind (Clark, 2013; Friston, 2010; Gilbert and Li, 2013; Kveraga et al., 2007). This can lead to some ambiguity when using the terms "prediction," "expectation," "surprise," etc., which have been used to describe both mental experience and neural processes. The psychological state of "having an expectation" or of "being surprised" are not the same as prediction and prediction error at the neural level of analysis (Clark, 2013). Yet, despite not being about subjective experience, predictive processing models can offer insights into how representational content emerges from the brain. For example, most theories suggest that predictions give rise to subjective experiences, whereas prediction errors are presumably processed unconsciously until the error is incorporated in a new prediction (i.e., model updating; Barrett and Simmons, 2015; Chanes and Barrett, 2016; Seth et al., 2012; Seth and Friston, 2016).

In this section, we review three aspects of predictive processing models including and their implications for affective neuroscience: (i) neural activity reflects a dynamic flow of predictions (including the precision of predictions) and prediction errors that occur throughout the brain (Bar, 2007; Clark, 2013; Friston, 2010; Hutchinson and Barrett, 2019; Rao and Ballard, 1999), (ii) neural activity is temporally dependent such that current brain activity is dependent on prior brain activity (Kiebel et al., 2008; Ploner et al., 2010), and (iii) the brain is hierarchically organized in a way that provides a structural foundation for a predictive processing architecture (Bastos et al., 2012; Chanes and Barrett, 2016; Clark, 2013; Friston, 2010, 2005; Goulas et al., 2018; Rao and Ballard, 1999; Sterling and Laughlin, 2015). Collectively, these features challenge the assumptions of invariance by context, time, and person that underlie the simple feature detection model.

3.1. *Neural activity is a function of prediction and prediction error*

A large body of work suggests that neural activity consists of

predictions and prediction errors that are ubiquitous throughout the brain. Prior information serves as predictions that drive activity across multiple domains including visual (Alink et al., 2010; den Ouden et al., 2010; Egner et al., 2010; Kok et al., 2017, 2012, 2011; Meyer and Olson, 2011; Murray et al., 2002; Summerfield et al., 2008) auditory (Blank and Davis, 2016; Wacongne et al., 2012), somatosensory (Atlas and Wager, 2014; Bingel et al., 2011; Freeman et al., 2015; Kong et al., 2009; Lui et al., 2010; Wagner et al., 2011; Watson et al., 2009), and affective processing (Belova et al., 2007; Johansen et al., 2010; Pessoa et al., 2002; Sussman et al., 2017). Given the ubiquity of predictive processing across multiple domains, it is striking that the vast majority of fMRI experiments in affective neuroscience are actually designed to limit predictability.

In a typical fMRI study in affective neuroscience (see Fig. 1A), evocative stimuli are presented with minimal context, in a randomized order, and with jittered temporal intervals. If one assumes a simple feature detector model for valence, then these choices make perfect sense — they serve to isolate and investigate the functioning of the simple feature detector for affect by eliminating putative confounding or moderating influences (see Table 2). However, from a predictive processing view, the task is processed by a brain that is nevertheless generating predictions and prediction errors throughout. Presenting evocative stimuli devoid of context, in a randomized order, and with unpredictable timing does not eliminate prediction and prediction error. Instead, it only precludes their systematic investigation and relative contributions in constructing subjective experience. As a consequence, the findings reflect the neural basis of affective processing in the unique case of highly unpredictable and atypical environments. It calls into question the validity of these findings for understanding the neural basis of affective experiences in everyday life.

If predictions are constitutive of affective experience, then experimental design may benefit by systematically investigating how predictability and unpredictability contribute to affective experience and its neural bases across multiple domains. Importantly, the ubiquity of predictions and prediction errors across the brain suggest that what predictions are "about" can span many dimensions and features (spatial location, stimulus content and temporal sequence, etc.). In turn, this raises the question of how predictions (and prediction errors) across these different dimensions and features contribute to affective experience. To provide examples of how affective neuroscientists might approach these research questions, we provide a more detailed review below of a small set of cognitive neuroscience studies to illustrate how predictive processing models have been used in the study of sensory perception (where most of this work is conducted). We also highlight relevant work in affective neuroscience that has not typically been construed as research on predictive processing, but nonetheless incorporates many features of predictive processing models (e.g., stimulus predictability, prediction error).

In their seminal study, Rao and Ballard (1999) used computational simulations to show that predictive processing models could better account for neural activity in early visual cortex in comparison to a simple feature detector model. They computationally modeled the interactions between V1 and V2 assuming a predictive or non-predictive processing architecture by including or excluding predictive signals from V2 in accounting for V1 activity in their simulation. In particular, they examined "endstopping" behavior of neurons. End stopping is a phenomenon in which a neuron that fires robustly to a stimulus in its receptive field (e.g., a line segment that is only as long as the neuron's receptive field) shows a markedly diminished response if the line is simply extended in space (i.e., a line segment that traverses through but also beyond the neuron's receptive field). According to a simple feature detection model, there should be no difference in response since the portion of the line segment that is within the neuron's receptive field is identical in both cases. However, from a predictive processing perspective, such shortened line segments are rare in natural scenes. While V2 neurons drive the prediction of spatial continuity of line segments, and the firing of V1 neurons actually reflects prediction error in the case when the V2 prediction is violated. These findings suggest that neurons in V1 are not necessarily firing in proportion to visual stimuli in their receptive fields (i.e., patches of retina that are innervated by retinal neurons). Rather, these findings suggest that V1 activity is a function of (spatial) predictions and prediction error.

Table 2
Study design choices viewed under simple feature detector vs. predictive processing models.

	Simple Feature Detector View	Predictive Processing View
Study Design Elements		
	Assumes expectancy about stimulus content (e.g., valence) is a confound.	Assumes expectations (or predictions) about stimulus content are constitutive of the phenomena.
Randomized Stimulus Presentations	Randomizing stimulus presentation order isolates the effect of the affective stimulus from the effect of expectancy.	Randomizing stimulus presentation order introduces a confound by inflating prediction errors and reducing the influence of predictions about stimulus content in affective processing.
	Assumes anticipation of stimulus onset is a confound.	Assumes expectations (or predictions) are part of a trajectory that is perturbed by the stimulus.
Jittered Interstimulus Intervals	Jittering stimulus onset isolates the effect of an affective stimulus from the effect of anticipation.	Jittering stimulus onset introduces a confound by inflating prediction errors and reducing the influence of predictions about stimulus onset in affective processing.
	Assumes that an unequal number of stimuli from categories of interest (e.g., positive vs. negative vs. neutral valence) introduces a confound.	Assumes that predictions are informed by the base rates of different kinds of stimuli observed in everyday life.
Equal Frequency of Stimulus Categories (i.e., Uniform Baserates)	Equal frequency of stimuli from different categories of interest isolates the effect of an affective stimulus from unbalanced frequency of stimuli.	Equal presentation of stimuli introduces a confound by conflating unnatural base rates of valenced stimuli with the effect of the stimuli.
	Assumes that contextual factors such as goal-relevance, situational, embedding of stimuli, sensory modalities of stimuli, etc. are confounds.	Assumes that contextual factors influence predictions about a stimulus.
Decontextualized Stimuli	Presenting decontextualized stimuli isolates the effect of an affective stimulus from contextual factors.	Decontextualizing stimuli introduces a confound by conflating inflated prediction errors by removing information relevant for generating predictions about the stimuli.

Note: The table outlines how theoretical assumptions of simple feature detection and predictive processing models influence experimental design choices. When studying the neural basis of affective experience in particular, expectancies about the stimulus content, timing, and frequency, are often viewed as sources of potential confounds when adopting a simple feature detection model. Care is taken to distribute their impact evenly across experimental conditions in a factorial design. However, in a predictive processing model, predictions (and even the precision of predictions) are considered to be constitutive of many psychological phenomena including affective and emotional experience. The design choices from the simple feature detection model introduce untenable challenges to external validity.

Work by Emberson et al. (2015) and Egner et al. (2010) further demonstrate that the brain may generate predictions about stimulus expectations on a temporal dimension. For example, Emberson et al. (2015) found evidence that functional activity in the occipital cortex is a function of prediction and prediction error signals in six-month old infants. Infants were presented with an auditory stimulus (sound of a rattle) that predicted the onset of a visual stimulus (picture of a cartoon face) while functional activity in the occipital cortex was measured using functional near infrared spectroscopy (fNIRS). On the key trials, a visual stimulus was not shown even though it was expected to be shown on the basis of the sounds. A simple feature detection model would suggest that activity in occipital cortex should track with the presentation of visual stimuli. However, the results were more consistent with a predictive processing account: there was greater activity in occipital cortex even when the visual stimulus was not presented, so long as it was predicted to be shown. Similarly, Egner et al. (2010) found that semantic cues that predicted or did not predict a stimulus modulated neural activity. In their study, participants were shown expected and unexpected face stimuli while tracking activity in the "fusiform face area" (FFA; Kanwisher et al., 1997). The FFA was sensitive to expectation: it showed greater activation when a face was expected to be shown regardless of whether face was actually shown. And the FFA was sensitive to prediction error. In particular, the FFA responded robustly when participants did not expect to see a face but were shown one anyway. When comparing model predictions, the pattern of findings was overall more consistent with a predictive processing model than a simple feature detection model for face stimuli (Egner et al., 2010; see also Apps and Tsakiris, 2013).

Finally, many studies have examined how expectations about affective stimuli shape neural responses during pain (Atlas and Wager, 2014; Benedetti et al., 2011; Lui et al., 2010; Ploghaus et al., 1999; Wager et al., 2004) and reward learning (Bayer and Glimcher, 2005; Hampton et al., 2007; O'Doherty et al., 2003, 2001; Ribas-Fernandes et al., 2011; Rolls et al., 2008; Schultz et al., 1993; Suri and Schultz, 2001). .In these studies on pain and reward learning, induced expectations are often learned as a part of the experimental procedure (e.g., via classical or operant conditioning). For example, a participant might undergo a conditioning procedure to associate a stimulus with a subsequent increased or decreased pain intensity (e.g., Benedetti et al., 2011). Similarly, in reward learning, participants might undergo conditioning to associate a neutral stimulus (e.g., a shape) with some reward or punishment (e.g., financial gains or losses; e.g., Hampton et al., 2007). Researchers then examine neural activity when these expectations are violated such as when pain intensity is not reduced or an expected reward does not materialize. Using this type of paradigm, studies on pain have found that expectations about pain strongly modulate activity in response to a nociceptive stimulus (Atlas and Wager, 2014; Freeman et al., 2015; Lui et al., 2010). Similarly, expectations of reward strongly modulate activity in response to a rewarding stimulus (e.g., de la Fuente-Fernández et al., 2002; O'Doherty et al., 2004; Pagnoni et al., 2002).

Of note, studies on pain and on reinforcement learning have not typically been understood through the lense of predictive processing models (for a few notable exceptions in pain see Büchel et al., 2014; Hechler et al., 2016; Saab and Barrett, 2017 and for reward learning, see Friston et al., 2009, 2015). While they share conceptual similarities with predictive processing models they also have important differences. **Reinforcement learning models** are commonly used in affective science to explain how agents make choices based on feedback from rewards and punishments (Bush and Mosteller, 1951a, 1951b; Dayan and Niv, 2008; Pearce and Hall, 1980; Rescorla and Wagner, 1972; Sutton and Barto, 2018). As such, these models also have traditionally had difficulty explaining responses to non-valenced stimuli or explaining how agents differentiate between stimuli that might have similar reward values (e.g., ice cream vs. chocolate; Niv and Schoenbaum, 2008).

In contrast, the main focus of predictive processing models is to explain brain activity by modeling it as neurons predicting activity in other neurons across scales in the neural hierarchy (Clark, 2013; Friston, 2010; Gilbert and Li, 2013; Kveraga et al., 2007; we elaborate on this idea in Sections 3.2 and 3.3). Therefore, predictive processing models do not presume explicit rewards or punishments. Rather, "rewards" in predictive processing are derived from increased metabolic savings (Sterling and Laughlin, 2015) or minimized differences between the expected and actual sensory inputs (that are relevant for survival; Friston, 2010) when these predictions are accurate. The gist is that evolution would favor metabolically efficient neural communication, and predictive processing provides a parsimonious model of such communication (Sterling and Laughlin, 2015).

Thus, whereas there are reinforcement learning *tasks* in congitive neuroscience, in predictive processing models predictions and prediction errors occur constantly and thus *during all tasks*. To be sure, to study predictive processing models experimenters often design tasks that make stimulus events more or less predictable, as illustrated above in studies by Egner et al. (2010) and Emberson et al. (2015), but it is not necessary to organize studies using temporally predictable sequences of trials. In Rao and Ballard's model of V1 and V2 interactions, for example, prediction and prediction error about lines extending in space occur when presented with any image — including images that were never seen before. In work on pain (Atlas and Wager, 2014; Benedetti et al., 2011; Lui et al., 2010; Ploghaus et al., 1999; Wager et al., 2004) and reward learning (Bayer and Glimcher, 2005; Hampton et al., 2007; O'Doherty et al., 2003, 2001; Ribas-Fernandes et al., 2011; Rolls et al., 2008; Schultz et al., 1993; Suri and Schultz, 2001), the predictions might be derived from semantic cues (as in the case of verbal instruction) or recently learned associations stored in memory (as in the case of conditioning). These examples underscore the point that during any experimental fMRI study, predictions are occurring ubiquitously. They may concern stimulus attributes (from low level sensory features to object perception to complex social interactions), stimulus timing (even when "jittered"), expected motor responses (button presses, eye movements), expected somatosensory stimulation (e.g., feeling of making a button press), predicted stimulus sampling given motor responses (e.g., from eye movements, pupil dilation), and also visceral changes (e.g., changes in heart rate, respiration), etc. A simple feature detection model for affective processing effectively ignores these sources of systematic variance. In contrast, a predictive processing account suggests that modeling the distribution of prediction and prediction error is critical to understanding how the brain creates affective experience.

3.2. *Neural activity is temporally dependent*

From a predictive processing view, the neural response to a stimulus reflects a "perturbation" from ongoing neural activity rather than an "elicitation" of activity by the stimulus, which has the mistaken connotation that a brain region lies dormant until its activation is elicited by a stimulus (see Fig. 1B for an example of this assumption of dormancy) and Fig. 5. Predictions and prediction errors occur continuously as the brain receives new information and updates its predictions accordingly (see Fig. 5). As such, there are temporal dependencies that should be considered when conducting experiments in affective neuroscience. In particular, neural activity in response to a stimulus must be examined with respect to the pre-stimulus **brain state**. Here, we define a brain state as neural activity throughout the brain at a given instance. This brain state is a function of continuously updated predictions and ensuing prediction errors that occur over time.

fMRI studies in affective neuroscience often ignore temporal dependence by assuming that individual trials are independent and identically distributed events (as is commonly assumed when using standard general linear models to fit the fMRI data). The neural response "elicited" during stimuli presented in a given trial is assumed to be independent from the response elicited by previous trials, and thus, responses are effectively averaged across trials. Moreover, expectations,

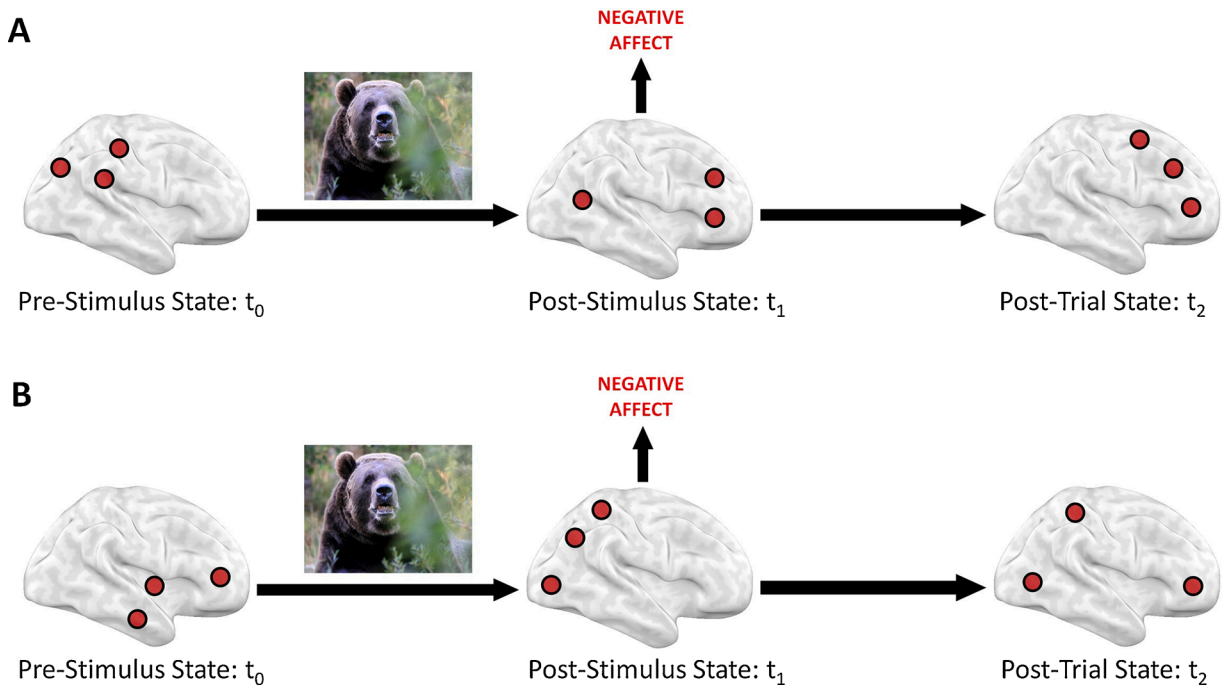


Fig. 5. Brain state transitions and degeneracy in predictive processing models.
Figures **(A)** and **(B)** depict two different brain state trajectories perturbed by the same stimulus. The red dots depict patterns of neural activation. In predictive processing models, an evocative stimulus (e.g., the bear) introduces a perturbation to an ongoing trajectory of brain activation (represented by brains on the left). The pattern of neural activity that follows stimulus onset (the post-stimulus state, t_1) will depend both on the stimulus and the pre-stimulus brain state (t_0). Note that patterns of neural activity in the pre-stimulus and post-stimulus brain states differ in Figures **(A)** and **(B)**, but they still relate with an experience of negative affect following the same stimulus. This many-to-one relationship between brain activation patterns and subjective experience categories is referred to as degeneracy. Finally, in predictive processing models, the brain does not return to baseline, but rather continues along the new trajectory introduced by the perturbation (the stimulus) and endogenous ongoing activity. This results in different post-trial (t_2) brain states. Photograph of bear adapted from "grizzly bear" by S. Kringen, 2010 (https://www.flickr.com/photos/18161271@N00/4957154697). CC BY-SA 2.0.

non-independence, and "carry-over" effects that occur across trials in an experiment are not only ignored analytically, they are viewed as a nuisance variable. To control for them, experimenters typically randomize (and jitter) the stimulus presentation order. These design choices are reasonable when assuming a simple feature detection model, but from a predictive processing view, the natural environment of the brain is one in which current neural activity is contingent on prior neural activity (see Table 2 for a comparison of the two views). This means that the history of past events (e.g., previous experimental trials) ought to be accounted for when modeling the pattern of neural activity during any given event (e.g., the current experimental trial).

There are two main sets of work indicating that temporal dependencies should not be overlooked. The first set of work challenges the idea that there is a stable "return to baseline" after presentation of an evocative stimulus. According to the simple feature detection model, the brain returns to baseline at which point a fresh response can be elicited that is independent of the prior trial. Despite the ubiquity of this notion in guiding much current research, these baselines can actually be highly variable. For instance, the duration of influence of evocative stimuli on behavioral measures, peripheral physiology, and neural measures varies considerably by person and experimental paradigm (e.g., from a few hundred milliseconds to several seconds, or even several minutes; Garrett and Maddock, 2001; Lapate and Heller, 2020; Walter et al., 2009). These findings indicate that a so-called return to baseline may be person and situation dependent, but rarely are these factors taken into consideration when analyzing fMRI data.

One can go even further by questioning the assumption of a stable baseline (Spivey, 2008). Indeed, it is unclear what a "return to baseline" ought to look like, and whether a stable baseline actually exists. Historically, the idea that the baseline state involved low-levels of activity spread uniformly throughout the brain gave way to the finding that a certain set of "default mode" brain regions were actually more active during rest than during task engagement (Raichle, 2015; Raichle et al., 2001). Some of the key regions of the default mode network include portions of the anterior medial prefrontal cortex, ventrolateral prefrontal cortex, posterior cingulate complex, hippocampus and other medial temporal lobe structures, and lateral temporal and parietal areas (Yeo et al., 2011). These areas are reliably activated during fixation in comparison to task engagement (and meaningfully activated insofar as they are consuming metabolic energy as measured in PET; Raichle et al., 2001). However, this "default mode" is not an inert baseline or "at rest" with respect to psychological function. While default mode areas often have reduced overall activity during certain cognitive and attentional performance tasks, they also have relatively greater activity during certain social cognitive tasks (even with respect to a fixation baseline; Davey et al., 2016; Iacoboni et al., 2004). Moreover, the default mode network also traverses through multiple functional connectivity states over time (as do all other large-scale functional networks; Ciric et al., 2017; Reinen et al., 2018). These findings suggest that the brain does not return to a stable functional baseline, at least when it comes to the spatiotemporal resolution of fMRI data.

The second set of work suggesting that temporal dependencies should not be overlooked are findings concerning the "pre-stimulus brain state". Given that the pre-stimulus period likely consists of different brain states in different moments (rather than a stable baseline), it raises the question of whether the pre-stimulus brain state has implications for the brain-behavior relationship. Research across multiple domains and measurement modalities are consistent with this notion. In memory research, incidental prestimulus fluctuations in neural activity correlate with the likelihood of remembering of the stimulus in studies using EEG (Otten et al., 2006), fMRI (Addante et al., 2015), and even intracranial recording (Sweeney-Reed et al., 2016). In

social neuroscience, incidental prestimulus fluctuations in BOLD signal in the dorsomedial prefrontal cortex increases the ease with which people process the social qualities of a stimulus (Spunt et al., 2015). In pain research, pre-stimulus functional connectivity of the BOLD signal correlates with whether a subsequently presented stimulus is experienced as painful or not (Ploner et al., 2010). According to predictive processing models, pre-stimulus activity may contain informative predictions that anticipate incoming information (Brodski-Guerniero et al., 2017). To test if that is the case, a particularly noteworthy study found that pre-stimulus MEG activity decoded the content of an expected stimulus, specifically, the orientation of lines in a visual stimulus, and was associated with improved behavioral performance (Kok et al., 2017).

Taken together, a predictive processing approach orients researchers to examine neural activity as a function of how a stimulus perturbs an ongoing brain state (see Fig. 5). The brain does not lie quiescent awaiting an evocative stimulus that will elicit a specific and unique neural response. Rather, the brain is in motion (Spivey, 2008), a stimulus perturbs its trajectory, and the observed functional activity in response to a stimulus reflects a deviation in the trajectory (Friston, 2008). The research we summarized above focuses primarily on relatively short time-scales on the order of a few seconds. However, predictive processing models are thought to "scale up" to much longer time intervals. The cumulative set of predictions are developed across a person's lifespan (Barrett and Simmons, 2015; Pereira et al., 2019; Sterling and Laughlin, 2015) and may be used at different time scales (Baldassano et al., 2018, 2017; Clark, 2013; Foster et al., 2016; Hasson et al., 2015; Honey et al., 2017, 2007; Seth et al., 2012; Zacks et al., 2007). Predictions developed over years will vary between individuals as a function of their personal developmental history and will likely be hard to modify (for a more in-depth take on the relationship between development and predictive processing, see Pereira et al., 2019).

The nature of temporal dependence in predictive processing models suggests several important directions for research in affective neuroscience. Future task-based fMRI studies might model neural activity during evocative stimuli as perturbations (rather than elicitations) from ongoing activity by incorporating the pre-stimulus brain state into the model. There have also been developments in using **state space models**. These models formally examine transitions in brain states over time and how the trajectory of brain states is perturbed by a stimulus (see McIntosh & Jirsa, 2019; Najafi et al., 2017; Pessoa, 2019). With respect to individual difference variables in social and affective neuroscience (Dubois and Adolphs, 2016), from a predictive processing view, individual differences stem in part from subject-specific predictions that have been developed over a long period of time. Such predictions may be more influential when unchecked by sources of prediction error stemming from the environment — in other words, predictions are more likely to be influential under conditions of reduced environmental constraint (e.g., uncertainty, ambiguity, reduced structure, or dearth of stimulus inputs). Rather than examining how individual difference measures relate with neural activity in response to relatively more concrete stimuli as many studies do (e.g., how trait anxiety relates with neural activity during "fear" faces), research may focus instead on cases in which the brain may rely on more deeply rooted predictions. That is, research may focus on how individual difference measures relate with pre-stimulus activity (e.g., Cuthbert et al., 2003; Ploner et al., 2010), activity during task conditions involving reduced environmental constraints (e.g., Hertel, 2000; Petro et al., 2018; Sussman et al., 2020), and the heterogeneous functional dynamics that occur during "rest" (e.g., Ran et al., 2017; Rashid et al., 2014; Reinen et al., 2018; Sakoğlu et al., 2010; Wu et al., 2015).

3.3. Predictive processing and the organization of the cortex

Decades of research have shown that many connections between cortical regions are hierarchically organized (Barbas, 2015; Barbas and Rempel-Clower, 1997; Felleman and Van Essen, 1991; Markov et al., 2014; Maunsell and Van Essen, 1983). Technically, it may be more accurate to describe the organization of the brain as heterarchical (Pessoa, 2019; see also Bruni & Giorgi, 2015; McCulloch, 1945; and Norman et al., 2011). Unlike a hierarchy, a **heterarchy** does not assume a fixed or static top-down relationship between brain regions. Rather, in a heterarchy the functional relationship between brain regions is flexible and bidirectional. In a heterarchy, whether one brain area is superordinate, subordinate, or equal in ranking to another may be determined by context (Pessoa, 2019). Further, brain regions near the bottom may have direct connections on areas near the top without going through intermediate levels and vice versa (Norman et al., 2011; Pessoa, 2019).

The loosely hierarchical (or heterarchical) organization is well suited to support a predictive processing architecture (Bastos et al., 2012; Chanes and Barrett, 2016; Clark, 2013; Friston, 2010, 2005; Sterling and Laughlin, 2015). Rao and Ballard's (1999) original predictive coding model, for example, was predicated on the hierarchical relationship between V1 and V2, with predictions flowing down from V2 to V1 and prediction errors flowing in the opposite direction. Predictive processing models have extended this principle to the cortex more broadly (Bastos et al., 2012; Huang and Rao, 2011). Brain regions lower in the hierarchy implement more spatiotemporally narrow predictions (Hasson et al., 2015; Kiebel et al., 2008; see also Finlay and Uchiyama, 2015) – for example, predictions of low-level activity in receptive fields of visual neurons on the order of milliseconds. In contrast, predictions originating in higher levels of the hierarchy are implementing spatiotemporally wider predictions, and may concern processes such as those involved in object recognition. These predictions may last over longer intervals of time (e.g., seconds) compared to predictions from lower levels in the hierarchy.

The organization of psychological processes are thought to parallel this structure. For example, when reading a book, predictions in early visual cortex may concern the expected sizes, shapes, and colors of ink on the page, and slightly further up the hierarchy, the letters and words that make up the text; even further up, predictions may be related to the narrative flow or expectations regarding the meanings conveyed by each sentence (Price and Devlin, 2011). The act of reading is therefore thought to involve a multilayer hierarchy of predictions: predictions at higher levels are not independent of predictions made at relatively low levels, and prediction errors signals from lower levels may eventually travel upward to much higher levels.

A similar hierarchical arrangement has been proposed in social and affective neuroscience (Barrett, 2017; Satpute et al., 2019; Spunt et al., 2016, 2011). Given the same input (e.g., sensory inputs regarding the actions of another person), the input may be conceptualized in a more lower-level, or concrete way (e.g., the person is shaking their fist back and forth) or in a higher-level, more abstract way (e.g., the person is angry, or the person is threatening someone). More abstract representations tend to accommodate a wider array of lower-level sensory possibilities at the cost of more precise details (i.e., a person may perceive shaking a first, yelling, or even smiling inappropriately as sensory inputs conveying threat or anger). It is likely that on a neural level, too, areas higher up the hierarchy may be less precise in their predictions by creating gist-level representations and discarding finer grain details (Bar, 2004, 2003; Bar et al., 2006; Kveraga et al., 2007; Finlay and Uchiyama, 2015). To consider an example in affective neuroscience, a more abstract, and less specific, conceptual representation of fear here may actually be advantageous in reducing prediction error in the long-run. If the representation of "fear" involves a highly specific prediction, for example to expect freezing, then moments of fear that do not include freezing would evoke considerable amounts of prediction error. A representation of fear that is not tailored to a specific behavior might result in more prediction error for specific instances of fear involving freezing, but it is also robust to instances of fear involving other behaviors (e.g., fleeing, or even attacking, smiling, or yelling in fear).

This hierarchical, predictive architecture suggests a different

mapping for the neural bases of affect and emotion than proposed by more traditional models. For emotion categories, traditional models have long assumed that discrete emotions like anger and fear reside in "low level" subcortical and limbic/paralimbic cortical structures (Dalgleish et al., 2009; Damasio and Carvalho, 2013; Kober et al., 2008; MacLean, 1990; Panksepp, 1982; Papez, 1937; Rolls, 1990; Smith and DeVito, 1984). Alternatively, in predictive processing models, representations of discrete emotions are considered to involve a multilevel set of predictions that span across the brain (Barrett, 2017; Satpute and Lindquist, 2019). Low-level predictions may drive specific motor actions (including organized motor actions such as freezing or attack) and their anticipated sensory consequences (i.e., active inferences; Adams et al., 2013; Smith et al., 2019a,b). However, these low-level predictions alone are non-diagnostic of an emotion category (i.e., people may attack in anger, fear, or other emotional states; Barrett, 2006; LeDoux, 2014) and insufficient for a neural basis of emotion representation. Discrete emotion representations are also abstract in that they model variation across diverse and heterogeneous sensory inputs (Barrett, 2006; Pereira et al., 2019). As such, a predictive processing view suggests that representations of discrete emotion categories like anger and fear may also rely on brain regions that are higher up in the hierarchy that are more capable of abstract information processing, such as heteromodal cortical areas of temporal and prefrontal cortex (Barrett, 2017; Satpute and Lindquist, 2019).

For affective experience dimensions, current predictive processing models propose that activity underlying an affective experience may span throughout the neural hierarchy (Allen and Friston, 2018; Chanes and Barrett, 2016; Seth and Friston, 2016) depending on the distribution of predictions and ensuing prediction errors that occur when presented with evocative stimulus. Whereas simple feature detection models assume there is a specific and reliable activation pattern that underlies affective experience, perhaps localized to limbic and paralimbic structures, predictive processing models assume that the neural basis of affective experience is contingent on the relationship between activity reflecting ongoing, hierarchically-organized predictions, and perturbations to this activity by an evocative stimulus. Since predictions are non-stationary (as discussed in Section 3.2) and since prediction errors are naturally contingent on predictions, neural activity underlying affective experience will depend on which aspects of an evocative stimulus are consistent with, or in violation of, ongoing predictions. As such, predictive processing models suggest there may be many different brain states that give rise to feelings of (dis)pleasure or arousal (or to a given emotion category such as fear; see Fig. 5). This many-to-one, brain states-to-behavior relationship has been variously referred to as degeneracy or multiple solutions in computational biology (Edelman and Gally, 2001; Marder and Taylor, 2011; Mason et al., 2015; Price and Friston, 2002).

It has been proposed that degeneracy is a natural consequence of a predictive processing architecture (Sajid et al., 2020). Consistent with this idea, research by Mather and colleagues suggests that heightened arousal may involve different brain regions depending on the brain state prior to receiving an input: areas that are on a trajectory of activation are further amplified, and those that are not are reduced (Mather et al., 2016; see also Shimaoka et al., 2018), which may be explained by a predictive processing account. For example, we (Ferreira-Santos, 2016) and others (Owens et al., 2018) proposed that this amplification could be modeled by the uncertainty of predictions (i.e., precision) in predictive processing models. Similar ideas have been proposed in cognitive neuroscience. For example, theoretical models of attention have proposed that attention is not localized to a particular brain region or network but instead is accounted for by the uncertainties in predictions that are distributed throughout the brain (Lupyan and Clark, 2015). These uncertainties are non-stationary, and thus, may involve different brain regions in different moments.

While degeneracy has not been formally tested in affective neuroscience, perhaps due to analytical challenges (see Khan et al., 2020), extant work is consistent with the idea that different neural pathways underlie affective experience in different situations. Coming back to the studies reviewed in Section 2, both the univariate and also multivariate studies reviewed suggest that brain regions supporting affective experience depend on the stimulus induction modality. A meta-analytic summary showed that early sensory areas are reliably engaged during affect inductions involving the corresponding sensory modality (even when using neural stimuli with similar sensory properties), and no brain regions were consistently engaged across them (Satpute et al., 2015; also see, Royet et al., 2000; Sambuco et al., 2020; Woo et al., 2014; see Fig. 6). Studies using MVPA have shown that affect-predictive neural patterns are also situation dependent. Chikazoe et al. (2014) reported several brain regions that were valence-predictive for either visual or gustatory valence dimensions. Chang et al. (2015) found that brain regions predicting negative affect evoked by cutaneous stimulation (i.e., pain) were distinct from those evoked by graphic pictures. Notably, research in non-human animals, too, has shown that early sensory areas exhibit plasticity related to valence processing (Blake et al., 2006; David et al., 2012; Gavornik et al., 2009; Polley et al., 2006).

One could retain the basic assumptions of the simple feature detection model by modifying the model to propose that there are specific and reliable circuits for modality-dependent affect systems (Miskovic and Anderson, 2018). From this view, activity in early visual areas would arise during affective visual stimulus inductions, but not affective auditory inductions. Alternatively, a predictive processing account suggests that greater activity in these early areas may be driven by prediction error in relatively lower levels of the hierarchy. In the typical image-viewing task designs in affective neuroscience, low level predictions may concern when and where to position the eyes, how much to dilate the pupils, and the expected changes in visual sensory information given these motor movements (a.k.a. "active inferences", Adams et al., 2013; Allen and Friston, 2018; Friston et al., 2009, as generated from eye-movements). Sensory predictions may also arise from prior experience of previously processed stimuli in the study, which may also relate to overall frequency and familiarity of certain content in stimuli. From a predictive processing view, insofar as affect involves a multilevel cascade of predictions and prediction errors, it is quite possible that predictions are generated in sensory cortical areas that are not the same as those receiving direct inputs. For example, an evocative image of a child screaming may drive predictions in auditory cortex, and those predictions may play a constitutive role in affective experience. Indeed, there is evidence of cross-modal affective prediction: when auditory emotional prosody is incongruous with a preceding facial expression of emotion there is an increase in neural activity in the auditory cortex, which may be interpreted as prediction error (Garrido-Vásquez et al., 2018).

3.4. Summary

Predictive processing models diverge sharply from simple feature detector models that have long guided study design in affective neuroscience. Rather than treating each trial in an experiment as an independent and identically distributed event, predictive processing models argue that experiments should account for the role of prediction and prediction error, temporal dependence, and the cortical hierarchy. Rethinking experimental design to concord with predictive processing models would require potentially radical shifts. Yet, in exchange, predictive processing models offer affective neuroscience new approaches with which to understand how neural activity relates with dimensions of affective experience.

4. Implications for inferring mental states in neurosciencetions for inferring mental states

The implications that predictive processing models have for affective neuroscience extend to the foundational, theoretical assumptions that

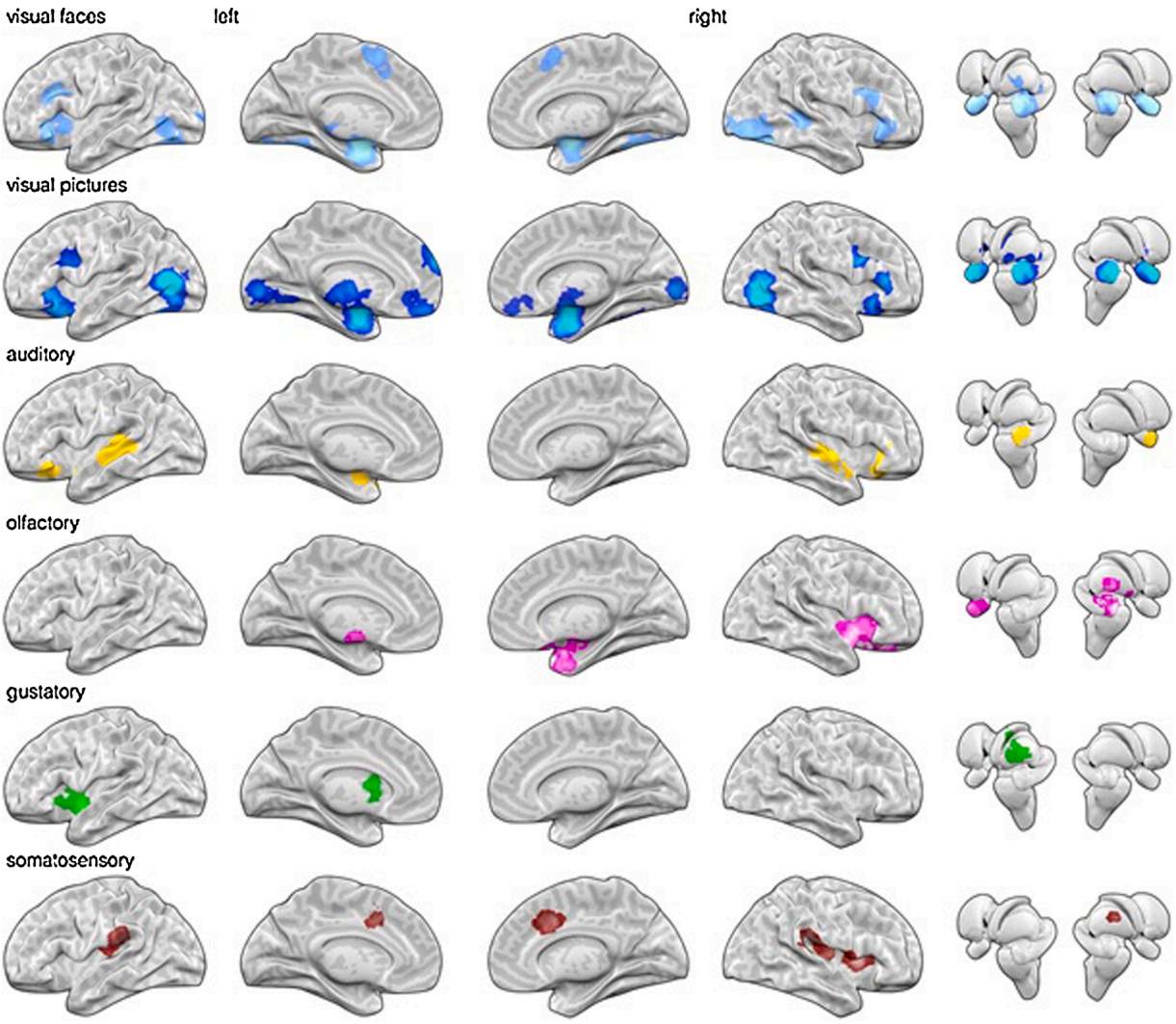


Fig. 6. Brain regions engaged during affective experience depend on the stimulus content.
The figure shows results from a meta-analysis of functional neuroimaging studies comparing evocative (positive and negative) v. neutral stimulus conditions (affective v. neutral faces, affective vs. neutral tastes, etc.; Satpute et al., 2015). Subcortical structures are presented on the right. Inconsistent with the idea there is a core set of brain regions supporting affective experience, there were no brain regions showing reliable activation during affective v. neutral conditions across all sensory modalities. For example, the amygdala were reliably engaged during visual stimuli, and also during auditory and olfactory stimuli (albeit with some lateralization), it was not reliably engaged during studies using gustatory and somatosensory affect inductions. The findings are consistent with degeneracy in functional neural organization: different neural pathways may support affective experience depending on the situation.

underlie over a century of research in psychology and neuroscience. Predictive processing models suggest that psychological states are generated by degenerate patterns of neural activity. How then should researchers go about inferring mental states from neural activity (i.e., reverse inference; Poldrack, 2015, 2011, 2006).

In the early days of fMRI, psychologists were excited by the prospect of using fMRI to decode mental states that occur during a task (Haynes and Rees, 2006; Poldrack et al., 2009). If fMRI could be used to reveal how a person feels about something using neural activity alone it would side-step thorny issues in using self-report measurements such as detrimental abilities to introspect on feelings (e.g., as in Autism spectrum disorder; Frith and Happé, 1999) or biases reporting what a person thinks they should feel rather rather than what they do feel (e.g., from social desirability or demand characteristics).

The initial enthusiasm was quickly dampened when it was pointed out that researchers were making a logical fallacy (Poldrack, 2015, 2011, 2006). Functional neuroimaging studies are generally designed to make "forward inferences" by inferring the probability of activation in a brain region given a certain psychological state induced by a task. They are not well-suited to make "reverse inferences," or inferring the probability of a psychological state from brain activity. A study might manipulate unpleasantness using a task and make the forward inference that certain brain regions are engaged during negative affect (e.g., the amygdala). However, observing amygdala activity during another task does not mean that a participant is feeling unpleasantness during the task (i.e., the reverse inference). To make a reverse inference, a researcher must show that activation in a brain region, or a pattern of activation across brain tissue, is reliably *and selectively* associated with a specific psychological state. Individual experiments are unable to test selectivity because it requires examining neural activity during all possible psychological states, and showing that the pattern of activation only occurs during negative affect, for example, but not other states.

One solution to the reverse inference problem is to estimate selectivity by using results from thousands of brain imaging studies (Yarkoni et al., 2011). Another solution is to forego strong claims of selectivity and instead to test for selectivity given the much smaller subset of psychological states or experimental conditions that are introduced in a study. MVPA studies reviewed in Section 2.2 have taken this latter

approach (albeit this limitation is rarely mentioned). Both solutions are based on the assumptions of a simple feature detection model; the analysis implies a search for the presence of a fixed neural activation pattern that is reliable and selective for a specific psychological state. As reviewed in Section 2, the evidence for that idea in affective neuroscience is mixed at best.

Predictive processing models are incompatible with this formulation of the reverse inference problem. A psychological state may involve different brain regions in different moments (degeneracy) depending on how a stimulus perturbs (introduces prediction error) in the brain's ongoing activity. Ongoing activity is itself non-stationary, and thus these perturbations and observed functional activities will vary across time (or trials). Further, the full set of predictions constitutes a person's internal model and may be idiosyncratic to a person's experience. For these reasons, it is unlikely that a fixed pattern of neural activity will be reliably and selectively associated with a certain psychological state.

To make a forward or reverse inference while adopting a predictive processing model requires conditioning these inferences on the prior brain state. That is, a forward inference involves inferring the probability of activation in a brain region *given the previous brain state* in addition to the psychological state induced by a stimulus. A reverse inference, then, involves inferring the probability of a specific psychological state *given the previous brain state* as well as the present brain state (that is, the combination of the two reflects a perturbation in the ongoing activity of the brain). Likewise, reliability and selectivity are also conditioned on the previous brain state. Modeling the previous brain state will necessarily be much more complicated than most contemporary analytical approaches permit. However, if predictive processing models are the way forward in affective neuroscience, then this may be the critical step to achieve high levels of reliability and sensitivity in future work.

5. Reevaluating internal and external validity

Predictive processing models also suggest reconsidering the traditional practice of emphasizing internal validity at the cost of external validity. This implication echos previous debates on this topic several decades ago (Brunswik, 1949, 1955; Dhami et al., 2004; Tolman and Brunswik, 1935). Traditional approaches in cognitive neuroscience have prioritized internal over external validity (as critiqued in Shamay-Tsoory and Mendelsohn, 2019). Experimental paradigms often bear little resemblance to tasks conducted in everyday life. Using the typical affect induction task from Fig. 1 as our running example, participants are presented with a series of evocative images for a few seconds at a time. The images are often assumed to be unrelated to one other (independent), evoke affective feelings generically (i.e., impersonally), and are presented as isolated and temporally jittered events that disrupt the ordinary flow of information processing in space and time. Further, negative, positive, and neutral events occur in equal proportion, which violates the natural occurrence of evocative events in everyday life. For example, it has been reported that healthy affective functioning typically involves experiencing pleasant affect at a ratio of 3–4 times as frequently as unpleasant affect (Schwartz, 1997; Schwartz et al., 2002). This base rate is commonly violated in affective neuroscience studies which tend to present equal numbers of positive and negative stimuli.

From a traditional approach, these design choices are actually considered a strength not a weakness. They maximize internal validity under the assumptions of a simple feature detection model. If there does exists a neural circuit that responds reliably and selectively to negative valence, it makes sense to study it in isolation from the ostensibly confounding influences of prior experiences (and thus, predictions) related to the semantic and perceptual content of the stimulus, timing, base rates, etc. The traditional view thus proceeds by first establishing causal relationships with high internal validity (assuming a simple feature detection model), and then attempting to translate the findings to more ecologically valid settings. In other words, they anchor on internal validity and then adjust for external validity.

In contrast, a predictive processing approach suggests anchoring on external validity instead (for similar views see Hintzman, 2011; Nabel, 2009; Shamay-Tsoory and Mendelsohn, 2019). Predictive processes from prior experience are not viewed as confounds, rather they are constitutive of psychological states. The brain develops a generative model of its neural activity in the context of the state of the body and current environment. In theory, even the prenatal brain is generating predictions, receiving sensory inputs (including from the internal milieu), processing prediction errors, and updating its predictions (Ciaunica et al., 91AD; Köster et al., 2020; Pereira et al., 2019). The accumulation of these predictions is an infant's internal model which updates and adapts to the environment throughout the lifespan. The everyday contexts in which affective feelings and emotions occur contribute to the predictions that the brain generates and the basis from which there is prediction error. To understand the neural basis of affect and emotion then, it is important to first develop paradigms that anchor on external validity to capture the phenomena with integrity (i.e., paradigms that utilize the internal model that a person has developed across the lifespan), and then introduce manipulations within that paradigm.

It is interesting to consider how a predictive neural architecture would engage with a traditional experimental task in cognitive neuroscience. Arguably, the brain will learn to predict various idiosyncratic features of the artificial task environment (e.g., generating predictions pertaining to the stimulus contents, motor demands, and even the timing and uncertainty of timings in jitters). As a result, the laboratory context may lead to findings that might not otherwise be observed in daily life in the broader world. These findings may even be "reliable" given the boundaries of the task, but may not generalize to a similar study with different parameters for stimulus contents, timing, etc. Alternatively, by emphasizing external validity, researchers will start with an approximation of functional organization for mental phenomena as they occur in everyday life. In turn, these insights may be more likely to generalize across contexts and lead to meaningful applications such as treatments for mood disorders. To be sure, emphasizing external validity does not mean abandoning internal validity. Instead, it means designing studies to be as externally valid for the phenomena in question as possible while still being able to establish causal conclusions.

As an illustrative example in affective science of how researchers might prioritize external validity but also maintain sufficient internal validity to draw causal conclusions, we outline a study examining manipulations of mood in everyday life (Kramer et al., 2014). Researchers examined the effects of positive and negative Newsfeed content on social behavior in Facebook users. They first modeled the base rates of positive and negative content for each individual. Then, they manipulated the amount of positive or negative content by 10 % relative to each participant's unique base rates. This means that different participants received different numbers of positive and negative messages. Indeed, on average there was about twice as much positive content as negative content. Newsfeed content is tailored to each participant, which means different participants received different, idiographic semantic content in the Newsfeed (e.g., positive content may concern uplifting messages for one participant, cute animals for another participant). Although more traditional experimental psychology might view these differences might be viewed as flaws, predictive processing models would view them as strengths. This is because the stimulus situation clearly represents the domain of interest, ecological validity is high, and the researchers can still infer causality.

Notwithstanding the unique challenges of a scanning environment, certain emerging paradigms are adopting more naturalistic stimulus situations that preserve the rich contextual details and spatiotemporal continuities of information processing of everyday life. For example, researchers have presented participants in fMRI studies with lengthy clips from movies (Baldassano et al., 2017; Chen et al., 2017; Hanke et al., 2014), music (Koelsch et al., 2005), and narrative story tellings

(Huth et al., 2016). These studies pose unique analytical challenges that are also being addressed with new analytical approaches (e.g., inter-subject functional connectivity; Chen et al., 2017), and creatively using analytical approaches to address hypotheses from predictive processing models (e.g., Richardson and Saxe, 2020).

Ultimately, to assess the usefulness of the predictive processing approach in affective neuroscience, there is a need for studies that directly compare findings from studies using these paradigms (e.g., Kim et al., 2020) with those using more traditional paradigms using computational models that are consistent with a predictive processing account (Smith et al., 2021, 2019a; Spratling, 2019). Notably, prior work in general may benefit from testing for external validity even if taking a traditional approach. As illustrated in Fig. 2, prior MVPA work has implicated heterogeneous sets of brain regions as carrying information for classifying valence. External validity may provide a yardstick against which to separate which brain regions are ultimately of greatest relevance if one is inclined to assume a simple feature detection model of the mind and brain.

6. Conclusion

Wilhelm Wundt (1897) popularized the scientific usage of affect over a century ago. Since then, researchers have devoted significant effort to uncover its basis in the brain (Barrett and Bliss-Moreau, 2009; Baucom et al., 2012; Berridge, 2019; Bush et al., 2017; Colibazzi et al., 2010; Lewis et al., 2007; Mather et al., 2016; Miskovic and Anderson, 2018; Phan et al., 2002; Tye, 2018). Despite this effort, a neural signature of affect has proven to be elusive. This may be because prior work implicitly or explicitly assumes that the brain operates like a simple feature detector for affect. In contrast, there is increasing evidence that the brain tries to actively predict its inputs rather than passively await them (Doya et al., 2007; Hutchinson and Barrett, 2019; Rao and Ballard, 1999).

As outlined in this review, predictive processing models suggest new theoretical, experimental, and analytical directions in understanding the neural basis of affective experience (and other phenomena in cognitive neuroscience). In this review we have not focused on specific predictive processing models of affect (for work developing in this direction see, e.g., Allen and Friston, 2018; Barrett, 2017; Hesp et al., 2019; Hutchinson and Barrett, 2019; Smith et al., 2019) but rather on more general consequences of theories of predictive processing for the field of affective neuroscience. Specifically, we considered three major aspects of predictive processing, the importance of the interplay between predictions and prediction errors, the temporal dependence of neural activity, and the hierarchical organization of neural structures. These aspects have significant inferential implications for affective neuroscience. In this regard, predictive processing seems to converge with other positions that are critical of traditional approaches (e.g., Shamay-Tsoory and Mendelsohn, 2019) but offers added value by providing a consilient yet parsimonious view of brain function. While there remains much to do in terms of developing new experimental paradigms and computational approaches that are informed by a predictive processing account, these advances may ultimately deliver a transformative new model for how the brain creates subjective feelings of pleasure, arousal, and emotion.

Acknowledgements

The authors thank Eliza Bliss-Moreau, Mariska Kret, and Jorg Massen for organizing the interdisciplinary Workshop on "Comparative Affective Science: The intersection of Biology and Psychology" (2017), kindly supported by the Lorentz Center of the University of Leiden (The Netherlands), where several ideas present in this paper were originally discussed. The authors also thank members of the Affective and Brain Sciences Lab, Maureen Ritchey, and Zulqarnain Khan for comments on earlier drafts. Research reported in this publication was supported by the Department of Graduate Education (NCS 1835309) and the Brain and Cognitive Sciences Division of the National Science Foundation (1947972), by the BIAL Foundation under grant number 242/14, and the National Institute of Mental Health of the National Institutes of Health under award number F32MH122062-01A1.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.neubiorev.2021.09.009.

References

Adams, R.A., Shipp, S., Friston, K.J., 2013. Predictions not commands: active inference in the motor system. Brain Struct. Funct. 218, 611–643.
Addante, R.J., de Chastelaine, M., Rugg, M.D., 2015. Pre-stimulus neural activity predicts successful encoding of inter-item associations. Neuroimage 105, 21–31.
Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., Muckli, L., 2010. Stimulus predictability reduces responses in primary visual cortex. J. Neurosci. 30, 2960–2966.
Allen, M., Friston, K.J., 2018. From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese 195, 2459–2482.
Apps, M.A.J., Tsakiris, M., 2013. Predictive codes of familiarity and context during the perceptual learning of facial identities. Nat. Commun. 4 https://doi.org/10.1038/ncomms3698.
Arnal, L.H., Giraud, A.-L., 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci. 16, 390–398.
Atal, B.S., 2006. The history of linear prediction. IEEE Signal Process. Mag. 23, 154–161.
Atlas, L.Y., Wager, T.D., 2014. A meta-analysis of brain mechanisms of placebo analgesia: consistent findings and unanswered questions. Handb. Exp. Pharmacol. 225, 37.
Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721.
Baldassano, C., Hasson, U., Norman, K.A., 2018. Representation of real-world event schemas during narrative perception. J. Neurosci. 38, 9689–9699.
Bar, M., 2003. A cortical mechanism for triggering top-down facilitation in visual object recognition. J. Cogn. Neurosci. 15, 600–609.
Bar, M., 2004. Visual objects in context. Nat. Rev. Neurosci. 5, 617–629.
Bar, M., 2007. The proactive brain: using analogies and associations to generate predictions. Trends Cogn. Sci. 11, 280–289.
Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E., 2006. Top-down facilitation of visual recognition. Proc. Natl. Acad. Sci. 103, 449–454.
Barbas, H., 2015. General cortical and special prefrontal connections: principles from structure to function. Annu. Rev. Neurosci. 38, 269–289.
Barbas, H., Rempel-Clower, N., 1997. Cortical structure predicts the pattern of corticocortical connections. Cereb. Cortex 7, 635–646. https://doi.org/10.1093/cercor/7.7.635.
Barrett, L.F., 2006. Solving the emotion paradox: categorization and the experience of emotion. Pers. Soc. Psychol. Rev. 10, 20–46.
Barrett, L.F., 2017. The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci. 12, 1–23. https://doi.org/10.1093/scan/nsw154.
Barrett, L.F., Bliss-Moreau, E., 2009. Affect as a psychological primitive. Adv. Exp. Soc. Psychol. 41, 167–218.
Barrett, L.F., Simmons, W.K., 2015. Interoceptive predictions in the brain. Nat. Rev. Neurosci. 16, 419–429.
Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711.
Baucom, L.B., Wedell, D.H., Wang, J., Blitzer, D.N., Shinkareva, S.V., 2012. Decoding the neural representation of affective states. Neuroimage 59, 718–727.
Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141. https://doi.org/10.1016/j.neuron.2005.05.020.
Belova, M.A., Paton, J.J., Morrison, S.E., Salzman, C.D., 2007. Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron 55, 970–984.
Benedetti, F., Amanzio, M., Rosato, R., Blanchard, C., 2011. Nonopioid placebo analgesia is mediated by CB1 cannabinoid receptors. Nat. Med. 17, 1228–1230. https://doi.org/10.1038/nm.2435.
Berridge, K.C., 2019. Affective valence in the brain: modules or modes? Nat. Rev. Neurosci. 20, 225–234.
Berridge, K.C., Kringelbach, M.L., 2015. Pleasure systems in the brain. Neuron 86, 646–664. https://doi.org/10.1016/j.neuron.2015.02.018.
Bingel, U., Wanigasekera, V., Wiech, K., Mhuircheartaigh, R.N., Lee, M.C., Ploner, M., Tracey, I., 2011. The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. Sci. Transl. Med. 3, 70ra14. https://doi.org/10.1126/scitranslmed.3001244.
Blake, D.T., Heiser, M.A., Caywood, M., Merzenich, M.M., 2006. Experience-dependent adult cortical plasticity requires cognitive association between sensation and reward. Neuron 52, 371–381.

Blank, H., Davis, M.H., 2016. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biol. 14, e1002577.
Bonnet, L., Comte, A., Tatu, L., Millot, J.-L., Moulin, T., Medeiros de Bustos, E., 2015. The role of the amygdala in the perception of positive emotions: an "intensity detector". Front. Behav. Neurosci. 9, 178.
Brodski-Guerniero, A., Paasch, G.F., Wollstadt, P., Ozdemir, I., 2017. Information-theoretic evidence for predictive coding in the face-processing system. J. Neurosci. 37, 8273–8283. https://doi.org/10.1523/JNEUROSCI.0614-17.2017.
Brunswik, E., 1949. Systematic and representative design of psychological experiments. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability 143–202.
Brunswik, E., 1955. Representative design and probabilistic theory in a functional psychology. Psychol. Rev. 62, 193.
Büchel, C., Geuter, S., Sprenger, C., Eippert, F., 2014. Placebo analgesia: a predictive coding perspective. Neuron 81, 1223–1239.
Bush, R.R., Mosteller, F., 1951a. A mathematical model for simple learning. Psychol. Rev. 58, 313–323.
Bush, R.R., Mosteller, F., 1951b. A model for stimulus generalization and discrimination. Psychol. Rev. 58, 413–423.
Bush, K.A., Inman, C.S., Hamann, S.B., Kilts, C.D., James, G.A., 2017. Distributed neural processing predictors of multi-dimensional properties of affect. Front. Hum. Neurosci. 11, 459.
Bush, K.A., Gardner, J., Privratsky, A.A., Chung, M.-H., James, G.A., Kilts, C.D., 2018. Brain states that encode perceived emotion are reproducible but their classification accuracy is stimulus-dependent. Front. Hum. Neurosci. 12, 262.
Bruni, L.E., Giorgi, F., 2015. Towards a heterarchical approach to biology and cognition. Prog. Biophys. Mol. Biol. 119, 481–492.
Canli, T., Desmond, J.E., Zhao, Z., Glover, G., Gabrieli, J.D., 1998. Hemispheric asymmetry for emotional stimuli detected with fMRI. Neuroreport 9, 3233–3239.
Chanes, L., Barrett, L.F., 2016. Redefining the role of limbic areas in cortical processing. Trends Cogn. Sci. 20, 96–106.
Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., Wager, T.D., 2015. A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol. 13, e1002180. https://doi.org/10.1371/journal.pbio.1002180.
Chen, J., Leong, Y.C., Honey, C.J., Yong, C.H., Norman, K.A., Hasson, U., 2017. Shared memories reveal shared structure in neural activity across individuals. Nat. Neurosci. 20, 115–125.
Chikazoe, J., Lee, D.H., Kriegeskorte, N., Anderson, A.K., 2014. Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. 17, 1114.
Ciaunica, A., Constant, A., Preissl, H., Fotopoulou, K., 1991. The first prior: from co-embodiment to co-homeostasis in early life. Conscious. Cogn., 103117
Ciric, R., Nomi, J.S., Uddin, L.Q., Satpute, A.B., 2017. Contextual connectivity: intrinsic dynamic architecture of large-scale functional brain networks. Sci. Rep.
Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36, 181–204.
Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Zhu, H., Kangarlu, A., Duan, Y., Russell, J.A., 2010. Neural systems subserving valence and arousal during the experience of induced emotions. Emotion 10, 377.
Costa, V.D., Lang, P.J., Sabatinelli, D., Versace, F., Bradley, M.M., 2010. Emotional imagery: assessing pleasure and arousal in the brain's reward circuitry. Hum. Brain Mapp. 31, 1446–1457. https://doi.org/10.1002/hbm.20948.
Cuthbert, B.N., Lang, P.J., Strauss, C., Drobes, D., Patrick, C.J., Bradley, M.M., 2003. The psychophysiology of anxiety disorder: fear memory imagery. Psychophysiology 40, 407–422.
Dalgleish, T., Dunn, B.D., Mobbs, D., 2009. Affective neuroscience: past, present, and future. Emot. Rev. 1, 355–368. https://doi.org/10.1177/1754073909338307.
Damasio, A., Carvalho, G.B., 2013. The nature of feelings: evolutionary and neurobiological origins. Nat. Rev. Neurosci. 14, 143–152.
Davey, C.G., Pujol, J., Harrison, B.J., 2016. Mapping the self in the brain's default mode network. NeuroImage 132, 390–397.
David, S.V., Fritz, J.B., Shamma, S.A., 2012. Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc. Natl. Acad. Sci. 109, 2144–2149.
Dayan, P., Niv, Y., 2008. Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 18, 185–196.
de la Fuente-Fernández, R., Phillips, A.G., Zamburlini, M., Sossi, V., Calne, D.B., Ruth, T. J., Stoessl, A.J., 2002. Dopamine release in human ventral striatum and expectation of reward. Behav. Brain Res. 136, 359–363.
den Ouden, H.E., Daunizeau, J., Roiser, J., Friston, K.J., Stephan, K.E., 2010. Striatal prediction error modulates cortical coupling. J. Neurosci. 30, 3210–3219. https://doi.org/10.1523/JNEUROSCI.4458-09.2010.
Dhami, M.K., Hertwig, R., Hoffrage, U., 2004. The role of representative design in an ecological approach to cognition. Psychol. Bull. 130, 959–988. https://doi.org/10.1037/0033-2909.130.6.959.
Doya, K., Ishii, S., Pouget, A., Rao, R.P., 2007. Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press.
Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. 20, 425–443.
Edelman, G.M., Gally, J.A., 2001. Degeneracy and complexity in biological systems. Proc. Natl. Acad. Sci. 98, 13763–13768.
Egner, T., Monti, J.M., Summerfield, C., 2010. Expectation and surprise determine neural population responses in the ventral visual stream. J. Neurosci. 30, 16601–16608.
Elias, P., 1955. Predictive coding–I. IRE Trans. Inf. Theory 1, 16–24.
Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 0956797620916786.
Emberson, L.L., Richards, J.E., Aslin, R.N., 2015. Top-down modulation in the infant brain: learning-induced expectations rapidly affect the sensory cortex at 6 months. Proc. Natl. Acad. Sci. 112, 9585–9590.
Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47. https://doi.org/10.1093/cercor/1.1.1.
Ferreira-Santos, F., 2016. The role of arousal in predictive coding. Behav. Brain Sci. 39 https://doi.org/10.1017/S0140525X15001788.
Finlay, B.L., Uchiyama, R., 2015. Developmental mechanisms channeling cortical evolution. Trends Neurosci. 38, 69–76.
Foster, B.L., He, B.J., Honey, C.J., Jerbi, K., Maier, A., Saalmann, Y.B., 2016. Spontaneous neural dynamics and multi-scale network organization. Front. Syst. Neurosci. 10, 7.
Freeman, S., Yu, R., Egorova, N., Chen, X., Kirsch, I., Claggett, B., Kaptchuk, T.J., Gollub, R.L., Kong, J., 2015. Distinct neural representations of placebo and nocebo effects. Neuroimage 112, 197–207.
Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. B Biol. Sci. 360, 815–836. https://doi.org/10.1098/rstb.2005.1622.
Friston, K., 2008. Hierarchical models in the brain. PLoS Comput. Biol. 4 https://doi.org/10.1371/journal.pcbi.1000211.
Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127–138.
Friston, K.J., Daunizeau, J., Kiebel, S.J., 2009. Reinforcement learning or active inference? PLoS One 4, e6421.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G., 2015. Active inference and epistemic value. Cogn. Neurosci. 6, 187–214.
Frith, U., Happé, F., 1999. Theory of mind and self-consciousness: What is it like to be autistic? Mind Lang. 14, 82–89.
Garrett, A.S., Maddock, R.J., 2001. Time course of the subjective emotional response to aversive pictures: relevance to fMRI studies. Psychiatry Res. Neuroimaging 108, 39–48.
Garrido-Vásquez, P., Pell, M.D., Paulmann, S., Kotz, S.A., 2018. Dynamic facial expressions prime the processing of emotional prosody. Front. Hum. Neurosci. 12, 244.
Gavornik, J.P., Shuler, M.G., Loewenstein, Y., Bear, M.F., Shouval, H.Z., 2009. Learning reward timing in cortex through reward dependent expression of synaptic plasticity. Proc. Natl. Acad. Sci. U. S. A. 106, 6826–6831. https://doi.org/10.1073/pnas.0901835106.
Gilbert, C.D., Li, W., 2013. Top-down influences on visual processing. Nat. Rev. Neurosci. 14, 350.
Gładziejewsk, P., 2016. Predictive coding and representationalism. Synthese 193, 559–582.
Goulas, A., Zilles, K., Hilgetag, C.C., 2018. Cortical gradients and laminar projections in mammals. Trends Neurosci. 41, 775–788.
Hampton, A.N., Adolphs, R., Tyszka, J.M., O'Doherty, J.P., 2007. Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. Neuron 55, 545–555.
Hamzani, O., Mazar, T., Itkes, O., Petranker, R., Kron, A., 2020. Semantic and affective representations of valence: Prediction of autonomic and facial responses from feelings-focused and knowledge-focused self-reports. Emotion 20 (3), 486–500.
Hanke, M., Baumgartner, F.J., Ibe, P., Kaule, F.R., Pollmann, S., Speck, O., Zinke, W., Stadler, J., 2014. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Sci. Data 1, 1–18.
Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci. (Regul. Ed.) 19, 304–313.
Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings. Neuroimage 62, 852–855.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Hechler, T., Endres, D., Thorwart, A., 2016. Why harmless sensations might hurt in individuals with chronic pain: about heightened prediction and perception of pain in the mind. Front. Psychol. 7, 1638.
Hertel, P.T., 2000. The cognitive-initiative account of depression-related impairments in memory. Psychol. Learn. Motiv. 39, 47–71.
Hesp, C., Smith, R., Allen, M., Friston, K., Ramstead, M., 2019. Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. Unpubl. Manuscr..
Hintzman, D.L., 2011. Research strategy in the study of memory: fads, fallacies, and the search for the "coordinates of truth". Perspect. Psychol. Sci. 6, 253–271.
Honey, C.J., Kötter, R., Breakspear, M., Sporns, O., 2007. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proc. Natl. Acad. Sci. 104, 10240–10245.
Honey, C.J., Newman, E.L., Schapiro, A.C., 2017. Switching between internal and external modes: a multiscale learning principle. Netw. Neurosci. 1, 339–356.
Huang, Y., Rao, R.P., 2011. Predictive coding. Wiley Interdiscip. Rev. Cogn. Sci. 2, 580–593.
Hutchinson, J.B., Barrett, L.F., 2019. The power of predictions: an emerging paradigm for psychological research. Curr. Dir. Psychol. Sci. 0963721419831992.
Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.
Iacoboni, M., Lieberman, M.D., Knowlton, B.J., Molnar-Szakacs, I., Moritz, M., Throop, C.J., Fiske, A.P., 2004. Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. Neuroimage 21, 1167–1173.
Itkes, O., Kimchi, R., Haj-Ali, H., Shapiro, A., Kron, A., 2017. Dissociating affective and semantic valence. J. Exp. Psychol. Gen. 146, 924.

Johansen, J.P., Tarpley, J.W., LeDoux, J.E., Blair, H.T., 2010. Neural substrates for expectation-modulated fear learning in the amygdala and periaqueductal gray. Nat. Neurosci. 13, 979.
Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in the human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.
Kassam, K.S., Markey, A.R., Cherkassky, V.L., Loewenstein, G., Just, M.A., 2013. Identifying emotions on the basis of neural activation. PLoS One 8, e66032.
Keller, G.B., Mrsic-Flogel, T.D., 2018. Predictive processing: A canonical cortical computation. Neuron 100, 424–435.
Khan, Z., Wang, Y., Sennesh, E.Z., Dy, J., Ostadabbas, S., van de Meent, J.-W., Hutchinson, J.B., Satpute, A.B., 2020. A computational neural model for mapping degenerate neural architectures. bioRxiv.
Kiebel, S.J., Daunizeau, J., Friston, K.J., 2008. A hierarchy of time-scales and the brain. PLoS Comput. Biol. 4, e1000209.
Kim, J., Wang, J., Wedell, D.H., Shinkareva, S.V., 2016. Identifying core affect in individuals from fMRI responses to dynamic naturalistic audiovisual stimuli. PLoS One 11, e0161589.
Kim, J., Shinkareva, S.V., Wedell, D.H., 2017. Representations of modality-general valence for videos and music derived from fMRI data. Neuroimage 148, 42–54.
Kim, J., Weber, C.E., Gao, C., Schulteis, S., Wedell, D.H., Shinkareva, S.V., 2020. A study in affect: predicting valence from fMRI data. Neuropsychologia 107473.
Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K.A., Wager, T.D., 2008. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage 42, 998–1031.
Koelsch, S., Fritz, T., Schulze, K., Alsop, D., Schlaug, G., 2005. Adults and children processing music: an fMRI study. Neuroimage 25, 1068–1076.
Kok, P., Rahnev, D., Jehee, J.F., Lau, H.C., De Lange, F.P., 2011. Attention reverses the effect of prediction in silencing sensory signals. Cereb. Cortex 22, 2197–2206.
Kok, P., Jehee, J.F., De Lange, F.P., 2012. Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75, 265–270.
Kok, P., Mostert, P., De Lange, F.P., 2017. Prior expectations induce prestimulus sensory templates. Proc. Natl. Acad. Sci. 114, 10473–10478.
Kong, J., Kaptchuk, T.J., Polich, G., Kirsch, I., Vangel, M., Zyloney, C., Rosen, B., Gollub, R., 2009. Expectancy and treatment interactions: a dissociation between acupuncture analgesia and expectancy evoked placebo analgesia. Neuroimage 45, 940–949.
Köster, M., Kayhan, E., Langeloh, M., Hoehl, S., 2020. Making sense of the world: infant learning from a predictive processing perspective. Perspect. Psychol. Sci. 1745691619895071.
Kramer, A.D., Guillory, J.E., Hancock, J.T., 2014. Experimental evidence of massive-scale emotional contagion through social networks. Proc. Natl. Acad. Sci. 111, 8788–8790.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.
Kringelbach, M.L., Berridge, K.C., 2009. Towards a functional neuroanatomy of pleasure and happiness. Trends Cogn. Sci. 13, 479–487.
Kveraga, K., Ghuman, A.S., Bar, M., 2007. Top-down predictions in the cognitive brain. Brain Cogn. 65, 145–168.
Lang, P.J., Bradley, M.M., Fitzsimmons, J.R., Cuthbert, B.N., Scott, J.D., Moulder, B., Nangia, V., 1998. Emotional arousal and activation of the visual cortex: an fMRI analysis. Psychophysiology 35, 199–210.
Lapate, R.C., Heller, A.S., 2020. Context matters for affective chronometry. Nat. Hum. Behav. 1–2.
LeDoux, J.E., 2014. Coming to terms with fear. Proc. Natl. Acad. Sci. 111, 2871–2878.
Lewis, P.A., Critchley, H.D., Rotshtein, P., Dolan, R.J., 2007. Neural correlates of processing valence and arousal in affective words. Cereb. Cortex 17, 742–748.
Lindquist, K.A., Satpute, A.B., Wager, T.D., Weber, J., Barrett, L.F., 2016. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. Cereb. Cortex 26, 1910–1922.
Lui, F., Colloca, L., Duzzi, D., Anchisi, D., Benedetti, F., Porro, C.A., 2010. Neural bases of conditioned placebo analgesia. Pain 151, 816–824. https://doi.org/10.1016/j.pain.2010.09.021.
Lupyan, G., Clark, A., 2015. Words and the world predictive coding and the language-perception-cognition interface. Curr. Dir. Psychol. Sci. 24, 279–284.
MacLean, P.D., 1990. The Triune Brain in Evolution: Role in Paleocerebral Functions. Plenum Press, New York, NY.
Marder, E., Taylor, A.L., 2011. Multiple models to capture the variability in biological neurons and networks. Nat. Neurosci. 14, 133–138.
Markov, N.T., Vezoli, J., Chameau, P., Falchier, A., 2014. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J. Comp. Neurol. 522, 225–259. https://doi.org/10.1002/cne.23458.
Mason, P.H., Winter, B., Grignolio, A., 2015. Hidden in plain view: degeneracy in complex systems. Biosystems 128, 1–8.
Mather, M., Clewett, D., Sakaki, M., Harley, C.W., 2016. Norepinephrine ignites local hotspots of neuronal excitation: how arousal amplifies selectivity in perception and memory. Behav. Brain Sci. 39 https://doi.org/10.1017/S0140525X15000667.
Maunsell, J.H.R., Van Essen, D.C., 1983. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. J. Neurosci. 3, 2563–2586.
McCulloch, W.S., 1945. A heterarchy of values determined by the topology of nervous nets. Bull. Math. Biophys. 7, 89–93.
McIntosh, A.R., Jirsa, V.K., 2019. The hidden repertoire of brain dynamics and dysfunction. Netw. Neurosci. 3, 994–1008.
Meyer, T., Olson, C.R., 2011. Statistical learning of visual transitions in monkey inferotemporal cortex. Proc. Natl. Acad. Sci. 108, 19401–19406.
Miskovic, V., Anderson, A., 2018. Modality general and modality specific coding of hedonic valence. Curr. Opin. Behav. Sci. Emot.-Cognit. Interact. 19, 91–97. https://doi.org/10.1016/j.cobeha.2017.12.012.
Murray, S.O., Kersten, D., Olshausen, B.A., Schrater, P., Woods, D.L., 2002. Shape perception reduces activity in human primary visual cortex. Proc. Natl. Acad. Sci. 99, 15164–15169.
Nabel, G.J., 2009. The coordinates of truth. Science 326, 53–54.
Najafi, M., Kinnison, J., Pessoa, L., 2017. Dynamics of intersubject brain networks during anxious anticipation. Front. Hum. Neurosci. 11, 552.
Niv, Y., Schoenbaum, G., 2008. Dialogues on prediction errors. Trends Cogn. Sci. 12, 265–272.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430.
Norman, G.J., Norris, C.J., Gollan, J., Ito, T.A., Hawkley, L.C., Larsen, J.T., Cacioppo, J. T., Berntson, G.G., 2011. Current emotion research in psychophysiology: The neurobiology of evaluative bivalence. Emot. Rev. 3, 349–359. https://doi.org/10.1177/1754073911402403.
O'Doherty, J.P., Kringelbach, M.L., Rolls, E.T., Hornak, J., Andrews, C., 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. Nat. Neurosci. 4, 95–102.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304, 452–454.
Olds, J., 1956. Pleasure centers in the brain. Sci. Am. 195, 105–117.
Otten, L.J., Quayle, A.H., Akram, S., Ditewig, T.A., Rugg, M.D., 2006. Brain activity before an event predicts later recollection. Nat. Neurosci. 9, 489–491.
Owens, A.P., Allen, M., Ondobaka, S., Friston, K.J., 2018. Interoceptive inference: from computational neuroscience to clinic. Neurosci. Biobehav. Rev. 90, 174–183.
Pagnoni, G., Zink, C.F., Montague, P.R., Berns, G.S., 2002. Activity in human ventral striatum locked to errors of reward prediction. Nat. Neurosci. 5, 97–98.
Panksepp, J., 1982. Toward a general psychobiological theory of emotions. Behav. Brain Sci. 5, 407–422. https://doi.org/10.1017/S0140525X00012759.
Papez, J.W., 1937. A proposed mechanism of emotion. Arch. NeurPsychol. 38, 725–743.
Pearce, J.M., Hall, G., 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. 87, 532.
Pereira, M.R., Barbosa, F., de Haan, M., Ferreira-Santos, F., 2019. Understanding the development of face and emotion processing under a predictive processing framework. Dev. Psychol. 55, 1868–1881. https://doi.org/10.1037/dev0000706.
Pessoa, L., 2019. Neural dynamics of emotion and cognition: from trajectories to underlying neural geometry. Neural Networks 120, 158–166.
Pessoa, L., Kastner, S., Ungerleider, L.G., 2002. Attentional control of the processing of neutral and emotional stimuli. Cogn. Brain Res. 15, 31–45.
Petro, N.M., Tong, T.T., Henley, D.J., Neta, M., 2018. Individual differences in valence bias: fMRI evidence of the initial negativity hypothesis. Soc. Cogn. Affect. Neurosci. 13, 687–698.
Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I., 2002. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage 16, 331–348. https://doi.org/10.1006/nimg.2002.1087.
Ploghaus, A., Tracey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., Rawlins, J.N., 1999. Dissociating pain from its anticipation in the human brain. Science 284, 1979–1981. https://doi.org/10.1126/science.284.5422.1979.
Ploner, M., Lee, M.C., Wiech, K., Bingel, U., Tracey, I., 2010. Prestimulus functional connectivity determines pain perception in humans. Proc. Natl. Acad. Sci. 107, 355–360.
Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10, 59–63.
Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697.
Poldrack, R.A., 2015. Reverse inference. In: Toga, A.W., Bandettini, P., Thompson, P., Friston, K. (Eds.), Brain Mapping: An Encyclopedic Reference. Academic Press, Cambridge, MA, pp. 647–650.
Poldrack, R.A., Halchenko, Y.O., Hanson, S.J., 2009. Decoding the large-scale structure of brain function by classifying mental States across individuals. Psychol. Sci. 20, 1364–1372. https://doi.org/10.1111/j.1467-9280.2009.02460.x.
Polley, D.B., Steinberg, E.E., Merzenich, M.M., 2006. Perceptual learning directs auditory cortical map reorganization through top-down influences. J. Neurosci. 26, 4970–4982.
Price, C.J., Devlin, J.T., 2011. The interactive account of ventral occipitotemporal contributions to reading. Trends Cogn. Sci. 15, 246–253.
Price, C.J., Friston, K.J., 2002. Degeneracy and cognitive anatomy. Trends Cogn. Sci. 6, 416–421.
Raichle, M.E., 2015. The brain's default mode network. Annu. Rev. Neurosci. 38, 433–447.
Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proc. Natl. Acad. Sci. 98, 676–682.
Ran, Q., Yang, J., Yang, W., Wei, D., Qiu, J., Zhang, D., 2017. The association between resting functional connectivity and dispositional optimism. PLoS One 12, e0180334.
Ransom, M., Fazelpour, S., Markovic, J., Kryklywy, J., Thompson, E.T., Todd, R.M., 2020. Affect-biased attention and predictive processing. Cognition 203, 104370.
Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79.

Rashid, B., Damaraju, E., Pearlson, G.D., Calhoun, V.D., 2014. Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects. Front. Hum. Neurosci. 8, 897.
Reinen, J.M., Chén, O.Y., Hutchison, R.M., Yeo, B.T., Anderson, K.M., Sabuncu, M.R., Öngür, D., Roffman, J.L., Smoller, J.W., Baker, J.T., 2018. The human cortex possesses a reconfigurable dynamic network architecture that is disrupted in psychosis. Nat. Commun. 9, 1157.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York, pp. 64–99.
Ribas-Fernandes, J.J., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y., Botvinick, M.M., 2011. A neural signature of hierarchical reinforcement learning. Neuron 71, 370–379.
Richardson, H., Saxe, R., 2020. Development of predictive responses in theory of mind brain regions. Dev. Sci. 23, e12863.
Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., Fusi, S., 2013. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590.
Robinson, M.D., Clore, G.L., 2002. Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. J. Pers. Soc. Psychol. 83, 198.
Rolls, E.T., 1990. A theory of emotion, and its application to understanding the neural basis of emotion. Cogn. Emot. 4, 161–190.
Rolls, E.T., McCabe, C., Redoute, J., 2008. Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. Cereb. Cortex 18, 652–663.
Royet, J.-P., Zald, D., Versace, R., Costes, N., Lavenne, F., Koenig, O., Gervais, R., 2000. Emotional responses to pleasant and unpleasant olfactory, visual, and auditory stimuli: a positron emission tomography study. J. Neurosci. 20, 7752–7759.
Saab, C.Y., Barrett, L.F., 2017. Thalamic bursts and the EPIC pain model. Front. Comput. Neurosci. 10.
Sabatinelli, D., Bradley, M.M., Lang, P.J., Costa, V.D., Versace, F., 2007. Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex. J. Neurophysiol. 98, 1374–1379.
Sajid, N., Parr, T., Hope, T.M., Price, C.J., Friston, K.J., 2020. Degeneracy and redundancy in active inference. Cereb. Cortex 30, 5750–5766.
Sakoğlu, Ü., Pearlson, G.D., Kiehl, K.A., Wang, Y.M., Michael, A.M., Calhoun, V.D., 2010. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. Magn. Reson. Mater. Phys. Biol. Med. 23, 351–366.
Sambuco, N., Bradley, M.M., Herring, D.R., Lang, P.J., 2020. Common circuit or paradigm shift? The functional brain in emotional scene perception and emotional imagery. Psychophysiology 57, e13522.
Satpute, A.B., Lindquist, K.A., 2019. The default mode network's role in discrete emotion. Trends Cogn. Sci. 23, 851–864.
Satpute, A.B., Kang, J., Bickart, K.C., Yardley, H., Wager, T.D., Barrett, L.F., 2015. Involvement of sensory regions in affective experience: a meta-analysis. Front. Psychol. 6, 1860.
Satpute, A.B., Hanington, L., Barrett, L.F., 2016. Novel response patterns during repeated presentation of affective and neutral stimuli. Soc. Cogn. Affect. Neurosci. 11, 1919–1932.
Satpute, A.B., Kragel, P.A., Barrett, L.F., Wager, T.D., Bianciardi, M., 2019. Deconstructing arousal into wakeful, autonomic and affective varieties. Neurosci. Lett. 693, 19–28.
Schultz, W., Apicella, P., Ljungberg, T., 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J. Neurosci. 13, 900–913.
Schwartz, R.M., 1997. Consider the simple screw: cognitive science, quality improvement, and psychotherapy. J. Consult. Clin. Psychol. 65, 970.
Schwartz, R.M., Reynolds III, C.F., Thase, M.E., Frank, E., Fasiczka, A.L., Haaga, D.A., 2002. Optimal and normal affect balance in psychotherapy of major depression: evaluation of the balanced states of mind model. Behav. Cogn. Psychother. 30, 439.
Seth, A.K., 2013. Interoceptive inference, emotion, and the embodied self. Trends Cogn. Sci. 17, 565–573.
Seth, A.K., Friston, K.J., 2016. Active interoceptive inference and the emotional brain. Philos. Trans. R. Soc. B Biol. Sci. 371, 20160007.
Seth, A.K., Suzuki, K., Critchley, H.D., 2012. An interoceptive predictive coding model of conscious presence. Front. Psychol. 2, 395.
Shamay-Tsoory, S.G., Mendelsohn, A., 2019. Real-life neuroscience: an ecological approach to brain and behavior research. Perspect. Psychol. Sci. 14, 841–859.
Shimaoka, D., Harris, K.D., Carandini, M., 2018. Effects of arousal on mouse sensory cortex depend on modality. Cell Rep. 22, 3160–3167.
Shinkareva, S.V., Gao, C., Wedell, D., 2020. Audiovisual representations of valence: a cross-study perspective. Affect. Sci. 1–10.
Shinkareva, S.V., Wang, J., Kim, J., Facciani, M.J., Baucom, L.B., Wedell, D.H., 2014. Representations of modality-specific affective processing for visual and auditory stimuli derived from functional magnetic resonance imaging data. Human Brain Mapping 35 (7), 3558–3568.
Skerry, A.E., Saxe, R., 2014. A common neural code for perceived and inferred emotion. J. Neurosci. 34, 15997–16008.
Smith, O.A., DeVito, J.L., 1984. Central neural integration for the control of autonomic responses associated with emotion. Annu. Rev. Neurosci. 7, 43–65.
Smith, R., Lane, R.D., Parr, T., Friston, K.J., 2019a. Neurocomputational mechanisms underlying emotional awareness: insights afforded by deep active inference and their potential clinical relevance. Neurosci. Biobehav. Rev. 107, 473–491.
Smith, R., Parr, T., Friston, K.J., 2019b. Simulating emotions: an active inference model of emotional state inference and emotion concept learning. Front. Psychol. 10, 2844.
Smith, R., Badcock, P., Friston, K.J., 2021. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. Psychiatry Clin. Neurosci. 75, 3–13.
Spivey, M., 2008. The Continuity of Mind. Oxford University Press.
Spratling, M.W., 2019. Fitting predictive coding to the neurophysiological data. Brain Res. 1720, 146313.
Spunt, R.P., Satpute, A.B., Lieberman, M.D., 2011. Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. J. Cogn. Neurosci. 23, 63–74. https://doi.org/10.1162/jocn.2010.21446.
Spunt, R.P., Meyer, M.L., Lieberman, M.D., 2015. The default mode of human brain function primes the intentional stance. J. Cogn. Neurosci. 27, 1116–1124.
Spunt, R.P., Kemmerer, D., Adolphs, R., 2016. The neural basis of conceptualizing the same action at different levels of abstraction. Soc. Cogn. Affect. Neurosci. 11, 1141–1151.
Stawarczyk, D., Bezdek, M.A., Zacks, J.M., 2019. Event representations and predictive processing: the role of the midline default network core. Top. Cogn. Sci.
Sterling, P., Laughlin, S., 2015. Principles of Neural Design. MIT Press.
Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.-M., Egner, T., 2008. Neural repetition suppression reflects fulfilled perceptual expectations. Nat. Neurosci. 11, 1004.
Suri, R.E., Schultz, W., 2001. Temporal difference model reproduces anticipatory neural activity. Neural Comput. 13, 841–862.
Sussman, T.J., Weinberg, A., Szekely, A., Hajcak, G., Mohanty, A., 2017. Here comes trouble: prestimulus brain activity predicts enhanced perception of threat. Cereb. Cortex 27, 2695–2707.
Sussman, T.J., Jin, J., Mohanty, A., 2020. The impact of top-down factors on threat perception biases in health and anxiety. In: Aue, T., Okon-Singer, H. (Eds.), Cognitive Biases in Health and Psychiatric Disorders: Neurophysiological Foundations. Elsevier Academic Press, San Diego, CA, pp. 215–241.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. MIT press.
Sweeney-Reed, C.M., Zaehle, T., Voges, J., Schmitt, F.C., Buentjen, L., Kopitzki, K., Richardson-Klavehn, A., Hinrichs, H., Heinze, H.-J., Knight, R.T., 2016. Pre-stimulus thalamic theta power predicts human memory formation. Neuroimage 138, 100–108.
Tolman, E.C., Brunswik, E., 1935. The organism and the causal texture of the environment. Psychol. Rev. 42, 43.
Tye, K.M., 2018. Neural circuit motifs in valence processing. Neuron 100, 436–452.
Wacongne, C., Changeux, J.-P., Dehaene, S., 2012. A neuronal model of predictive coding accounting for the mismatch negativity. J. Neurosci. 32, 3665–3678.
Wager, T.D., Rilling, J.K., Smith, E.E., Sokolik, A., Casey, K.L., Davidson, R.J., Kosslyn, S. M., Rose, R.M., Cohen, J.D., 2004. Placebo-induced changes in FMRI in the anticipation and experience of pain. Science 303, 1162–1167. https://doi.org/10.1126/science.1093065.
Wagner, D.D., Kelley, W.M., Heatherton, T.F., 2011. Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. Cereb. Cortex 21, 2788–2796. https://doi.org/10.1093/cercor/bhr074.
Walter, H., von Kalckreuth, A., Schardt, D., Stephan, A., Goschke, T., Erk, S., 2009. The temporal dynamics of voluntary emotion regulation. PLoS One 4, e6726.
Watson, A., El-Deredy, W., Iannetti, G.D., Lloyd, D., Tracey, I., Vogt, B.A., Nadeau, V., Jones, A.K., 2009. Placebo conditioning and placebo analgesia modulate a common brain network during pain anticipation and perception. Pain 145, 24–30.
Weaverdyck, M.E., Lieberman, M.D., Parkinson, C., 2020. Tools of the Trade Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists. Soc. Cogn. Affect. Neurosci. 15, 487–509.
Wise, R.A., 1980. The dopamine synapse and the notion of 'pleasure centers' in the brain. Trends Neurosci. 3, 91–95.
Woo, C.-W., Koban, L., Kross, E., Lindquist, M.A., Banich, M.T., Ruzic, L., Andrews-Hanna, J.R., Wager, T.D., 2014. Separate neural representations for physical pain and social rejection. Nat. Commun. 5.
Wu, J., Dong, D., Jackson, T., Wang, Y., Huang, J., Chen, H., 2015. The neural correlates of optimistic and depressive tendencies of self-evaluations and resting-state default mode network. Front. Hum. Neurosci. 9, 618.
Wundt, W.M., 1897. Outlines of Psychology. Engelmann, Leipzig, Germany.
Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670. https://doi.org/10.1038/nmeth.1635.
Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165. https://doi.org/10.1152/jn.00338.2011.
Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R., 2007. Event perception: a mind-brain perspective. Psychol. Bull. 133, 273.