OXFORD

Social Cognitive and Affective Neuroscience, 2017, 1–23

doi: 10.1093/scan/nsw154
Advance Access Publication Date: 19 October 2016
Original article

The theory of constructed emotion: an active inference account of interoception and categorization

Lisa Feldman Barrett[1,2,3]

[1]Department of Psychology, Northeastern University, Boston, MA, USA, [2]Athinoula, A. Martinos Center for Biomedical Imaging and [3]Psychiatric Neuroimaging Division, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, USA.

Correspondence should be addressed to Lisa Feldman Barrett. Department of Psychology, Northeastern University, Boston, MA, USA. E-mail: l.barrett@neu.edu

Abstract

The science of emotion has been using folk psychology categories derived from philosophy to search for the brain basis of emotion. The last two decades of neuroscience research have brought us to the brink of a paradigm shift in understanding the workings of the brain, however, setting the stage to revolutionize our understanding of what emotions are and how they work. In this article, we begin with the structure and function of the brain, and from there deduce what the biological basis of emotions might be. The answer is a brain-based, computational account called the theory of constructed emotion.

Key words: emotion; predictive coding; construction; interoception; categorization; concepts; affect

Ancient philosophers and physicians believed a human mind to be a collection of mental faculties. They divided the mind, not with an understanding of biology or the brain, but to capture the essence of human nature according to their concerns about truth, beauty and ethics. The faculties in question have morphed over the millennia, but generally speaking, they encompass mental categories for thinking (cognitions), feeling (emotions) and volition (actions, and in more modern versions, perceptions). These mental categories symbolize a cherished narrative about human nature in Western civilization: that emotions (our inner beast) and cognitions (evolution's crowning achievement) battle or cooperate to control behavior.[1] The classical view of emotion (Figure 1) was forged in these ancient ideas. Affective neuroscience takes its inspiration from this faculty-based approach. Scientists begin with emotion concepts that are most recognizably English (Pavlenko, 2014; Wierzbicka, 2014), such as anger, sadness, fear, and disgust, and search for their elusive biological essences (i.e. their neural signatures or fingerprints), usually in subcortical regions. This inductive approach assumes that the emotion categories we experience and perceive as distinct must also be distinct in nature.

If the history of science has taught us anything, however, it is that human experiences rarely reveal the way that the natural world works: 'Physical concepts are free creations of the human mind, and are not; however, it may seem, uniquely determined by the external world' (Einstein *et al.*, 1938, p. 33). The last two decades of neuroscience research have brought us to the brink of a paradigm shift in understanding the workings of the brain, setting the stage to revolutionize our study of emotions (or any mental category). So in this article, we turn the typical inductive approach on its head. We begin not with mental categories but with the structure and function of the brain, and from there deduce what the biological basis of emotions might be. The answer, I suggest, will look something like the theory of constructed emotion (Barrett, 2017), formerly, the conceptual act theory of emotion (Barrett, 2006b, 2011a, 2012, 2013, 2014).

To begin this discussion, I first outline enough background on brain structure and function to start asking informed questions about the biological basis of emotion. I then introduce the theory of constructed emotion, contrast it briefly with the classical view when instructive to do so, and consider selected

1 Throughout the millennia, with few exceptions, cognitions were thought to reside in the brain, emotions in the body, and then later, emotions were relocated to the parts of the brain that control the body. For example, Aristotle placed both thinking and feeling in organs of the body; Descartes kept emotions in the body and placed cognition in the pineal gland of the brain).

Received: 13 August 2016; **Revised:** 13 August 2016; **Accepted:** 11 October 2016

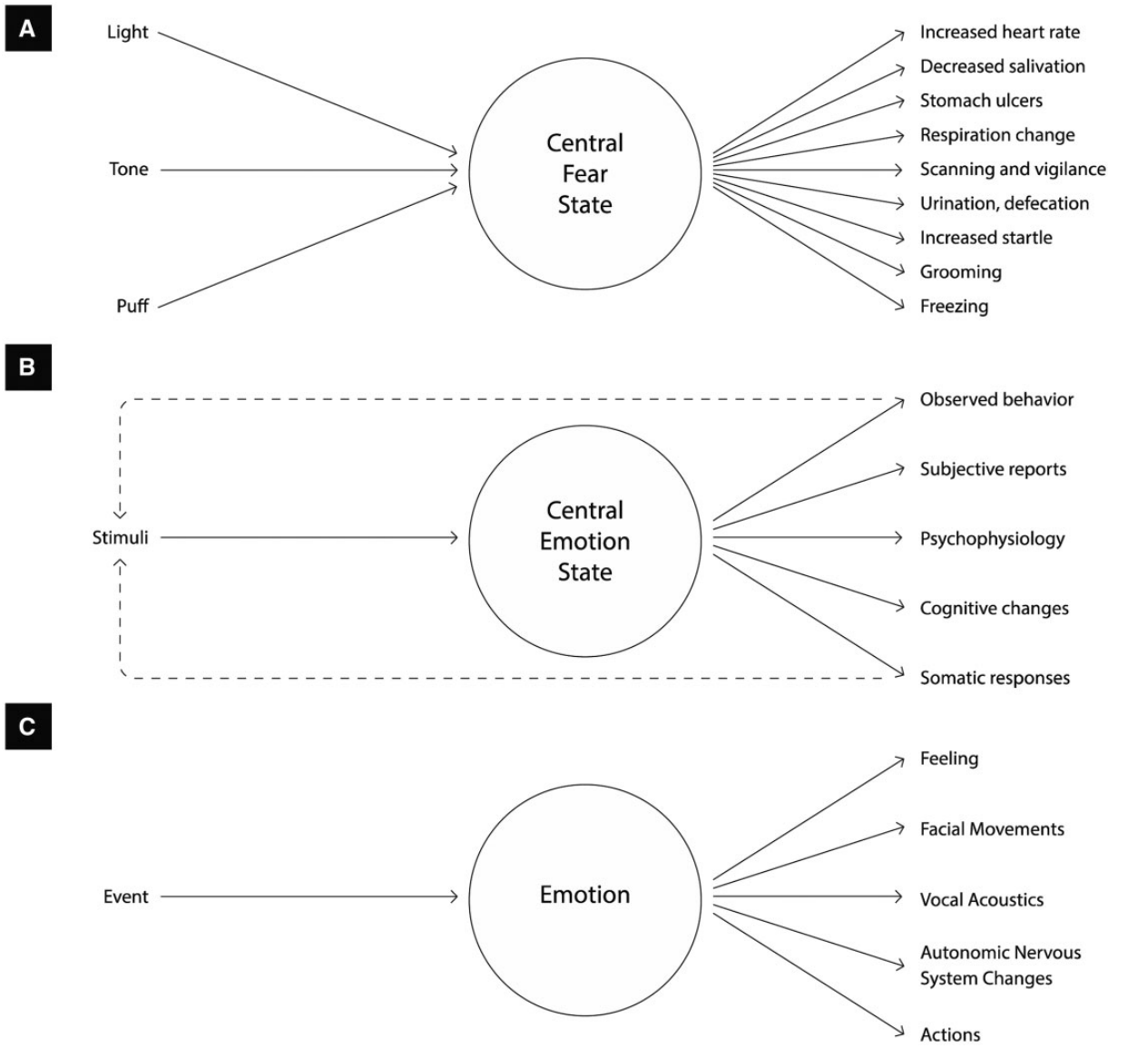

Fig. 1. The classical view of emotion. The classical view of emotion includes basic emotion theories (e.g. for a review, see Tracy and Randles, 2011), causal appraisal theories (e.g. Scherer, 2009; Roseman, 2011), and theories of emotion that rely on black-box functionalism (Davis, 1992; Anderson and Adolphs, 2014). Each emotion faculty is assumed to have its own innate 'essence' that distinguishes it from all other emotions. This might be a Lockean essence (an underlying causal mechanism that all instances of an emotion category share, making them that kind of emotion and not some other kind of emotion, depicted by the circles in the figure). Lockean essences might be a biological, such as a set of dedicated neurons, or psychological, such as a set of evaluative mechanisms called 'appraisals'. An emotion category is usually assumed to have a Platonic essence [a physical fingerprint that instances of that emotion share, but that other emotions do not, such a set of facial movements (an 'expression'), a pattern of autonomic nervous system activity, and/or a pattern of appraisals]. Of course, no one is expecting complete invariance, but it is assumed that instances of a category are similar enough to be easily diagnosed as the same emotion using objective (perceiver-independent) measures alone. **(A)** is adapted from Davis (1992). **(B)** is adapted Anderson and Adolphs (2014). **(C)** is adapted from Barrett (2006a), which reviews the growing evidence that contracts the classical view of emotion.

findings on the brain basis of emotion through the theory's lens. In the process, I offer a series of novel hypotheses about what emotions are and how they work.

The biological background

What is a brain?

A brain is a network of billions of communicating neurons, bathed in chemicals called neurotransmitters, which permit neurons to pass information to one another (Doya, 2008; Bargmann, 2012). The firing of a single neuron (or a small population of neurons) represents the presence or absence of some feature at a moment in time (Deneve, 2008; Deneve and Jardri, 2016). However, a given neuron (or group of neurons) represents different features from moment to moment (e.g. Stokes *et al.*, 2013; Spillmann *et al.*, 2015) because many neurons synapse onto one (many-to-one connectivity), and a neuron's receptive field depends on the information it receives (i.e. depends on its neural context in the moment; McIntosh, 2004). Conversely, one neuron also synapses on many

other neurons [one-to-many connectivity\ (Sporns, 2011; Sterling and Laughlin, 2015)] to help implement instances of different psychological categories. As a consequence, neurons are multipurpose [for evidence and discussion, see (Barrett and Satpute, 2013; Anderson, 2014; Anderson and Finlay, 2014)], even in subcortical regions like the amygdala (Cerf, personal communication, 30 July 2015).[2]

When the brain is viewed as a massive network, rather than a single organ or a collection of 'mental modules', it becomes apparent that this one anatomic structure of neurons can create an astounding number of spatiotemporal patterns, making the brain a network of high complexity (Sporns, 2011; Bullmore and Sporns, 2012; Rigotti *et al.*, 2013). Natural selection prefers high complexity systems as they can reconfigure themselves into a multitude of different states (Whitacre, 2010; Whitacre and Bender, 2010; Sterling and Laughlin, 2015).

The brain achieves complexity through 'degeneracy' (Edelman and Gally, 2001), the capacity for dissimilar representations (e.g. different sets of neurons) to give rise to instances of the same category (e.g. anger) in the same context (i.e. many-to-one mappings of structure to function). Degeneracy is ubiquitous in biology, from the workings inside a single cell to distributed brain networks (e.g. see Tononi *et al.*, 1999; Edelman and Gally, 2001; Marder and Taylor, 2011). Natural selection favors systems with degeneracy because they are high in complexity and robust to damage (Whitacre and Bender, 2010). Degeneracy explains why Roger, the patient who lost his limbic circuitry to herpes simplex type I encephalitis, still experiences emotions (Feinstein *et al.*, 2010) and why monozygotic twins with fully calcified basolateral sectors of the amygdala [due to Urbach-Wiethe disease (UWD)] have markedly different emotional capacity, despite genetic and environmental similarity (Becker *et al.*, 2012; Mihov *et al.*, 2013). Degeneracy also explains how a characteristic can be highly heritable even without a single set of necessary and sufficient genes (e.g. Turkheimer *et al.*, 2014).

In emotion research, degeneracy means that instances of an emotion (e.g. fear) are created by multiple spatiotemporal patterns in varying populations of neurons. Therefore, it is unlikely that all instances of an emotion category share a set of core features (i.e. a single facial expression, autonomic pattern or set of neurons; see Clark-Polner *et al.*, 2016). This observation is an example of population thinking, pioneered in Darwin's *On the Origin of Species* (Mayr, 2004).[3] By observing the natural world, Darwin realized that biological categories, such as a species, are conceptual categories (highly variable instances, grouped together by 'a goal' rather than by similar features or a single, shared underlying cause; (Mayr, 2004). My hypothesis, following Darwin's insight, is that fear (or any other emotion) is a 'category' that is populated with highly variable instances (Clark-Polner *et al.*, 2016; Clark-Polner, Johnson & Barrett, 2016; e.g. Wilson-Mendenhall *et al.*, 2011, 2015). The summary representation of any emotion category is an abstraction that need not exist in nature (as is true for any biological category; for a discussion of population thinking, see Mayr, 2004; as applied to emotion concepts and categories, see Barrett, 2017; and, as applied to concepts and categories more generally see Barsalou, 1983; Voorspoels *et al.*, 2011). The fact that human brains effortlessly and automatically construct such representations (e.g. Murphy, 2002; Posner and Keele, 1968) helps to explain why scientists continue to believe in the classical view and even propose it as an innovation (e.g. Anderson and Adolphs, 2014), even as evidence continues to call it into doubt (e.g. Barrett, 2006a, 2016b, 2012; Barrett *et al.*, 2007a; see Table 1 for specific neuroscience examples, with a particular focus on fear as the category that has garnered the most support for the classical view).

What is a brain for?

A brain did not evolve for rationality, happiness or accurate perception. All brains accomplish the same core task (Sterling and Laughlin, 2015): to efficiently ensure resources for physiological systems within an animal's body (i.e. its internal milieu) so that an animal can grow, survive and reproduce. This balancing act is called 'allostasis' (Sterling, 2012). Growth, survival and reproduction (and therefore gene transmission) require a continual intake of metabolic and other biological resources. Metabolic and other expenditures are required to plan and execute the physical movements necessary to acquire those resources in the first place (and to protect against threats and dangers). Allostasis is not a condition of the body, but a process for how the brain regulates the body according to costs and benefits; 'efficiency' requires the ability to anticipate the body's needs and satisfy them before they arise (Sterling, 2012; Sterling and Laughlin, 2015).[4] An animal thrives when it has sufficient resources to explore the world, and to consolidate the details of

2 Cells in the medial temporal lobe (including the amygdala) appear to act as a memory cache for important things (e.g. photos of friends, family, famous people, the patients themselves, landscapes, directions; some cells don't respond to anything for a few days, and then begin to respond when the experimenters walk into the room); at some other point, the cells might adopt and code for something entirely different that becomes important (Cerf, personal communication, 30 July 2015). Even primary sensory neurons are not coding for single sensory features but for associations between one feature (like the presence or absence of a line) with other sensory features; e.g. V1 neurons have receptive fields that include auditory and sensorimotor changes (e.g. Liang *et al.*, 2013).

3 Before *On the Origin of Species*, a 'species' was defined as biological type (i.e. with a set of unchanging physical characteristics or features that are passed down through the generations). This typological characterization fundamentally underestimates within-category variation (in its phenotypic features and in it's gene pool) and over-estimates between category variation (and borderline cases are often encountered; Mayr, 2004; Gelman and Rhodes, 2012). One of Darwin's greatest conceptual innovations in *Origin* was to revolutionize the concept of a species as a biopopulation of highly variable individuals (instead of a group of highly similar creatures who share a set of co-occurring biological features) (Mayr, 2004). Since then, the concept of a 'species' has been characterized on the basis of what category members do (i.e. functionally), not on the basis of a shared gene pool or a set of physical features: A species is a reproductive community (sometimes, members of different species are reproductively incompatible; sometimes they don't, such as when they are geographically isolated). Fundamentally, this translates into the insight that a biological category (a 'species') is a conceptual category, rather than a typological one: a species is a population of physically unique individuals who similarities are defined functionally, not physically.

4 Sometimes allostasis involves dynamically regulating resource allocation (i.e. diverting glucose, electrolytes, water, etc. from one system to another) to meet the body's spending needs; e.g. in advance of standing up, the heart beats stronger and faster, blood vessels constrict, and blood pressure raises to ensure that the brain continues to receive the blood (and oxygen). Sometimes allostasis involves signaling the need for resources before the body runs out (e.g. drinking before dehydration occurs) or preparing for the intake of resources in advance of their ingestion; e.g. saliva example in humans and some other mammals, saliva is made of alpha-amylase which is an enzyme that breaks down glucose. When the body is in need of glucose, saliva is pre-emptively secreted (even before anything is ingested). Even just imaging food causes glucose secretion.

Table 1. Examples of neuroscience evidence that disconfirm the classical view of emotion

Observation	Method	Example Citations
Different emotion categories cannot be specifically and consistently localized to distinct populations of neurons within a single region of the human brain.	Human neuroimaging: task-related data	(Vytal and Hamann, 2010; Lindquist et al., 2012)
Different emotion categories cannot be consistently localized to specific intrinsic networks in the human brain.	Human neuroimaging: intrinsic connectivity data	(Barrett and Satpute, 2013; Touroutoglou et al., 2015)
The instances of an emotion category need not share a population of neurons that are necessary or sufficient to implement them.	Human neuroimaging: multi-voxel pattern analysis	(Clark-Polner, Johnson, & Barrett, 2016)
Individual neurons, when stimulated in studies of experience, expression, and perception, do not have emotion-specific receptive fields.	Intracranial stimulation in humans	(Guillory and Bujarski, 2014)
Lesions to the amygdala produce variable functional consequences. Monozygotic twins, whose basolateral nuclei of both amygdalae are calcified due to UWD, do not show equivalent deficits in experiencing and perceiving fear; patient BG has deficits similar to patient SM (who has complete loss of both amygdalae due to UWD), whereas her sister, AM, is able to experience and perceive fear when BG cannot. Other people with basolateral lesions from UWD show different problems in fear perception (they are vigilant to rather than neglectful of posed fear faces). Patient SM can experience intense fear in the real world under certain circumstances, and her impairments in fear perception appear to be limited to experiments where she is asked to view stereotyped, fear poses and explicitly categorize them as fearful. There is ample evidence that she is able to perceive fear in various circumstances in real life (see Box 1).	Behavioral observations in humans with amygdala lesions	(Bechara et al., 1995, 1999; Adolphs et al., 1999; Adolphs and Tranel, 1999, 2003; Atkinson et al., 2007; de Gelder et al., 2014; Hampton et al., 2007; Tsuchiya et al., 2009; Hurlemann et al., 2007; De Martino et al., 2010; Boes et al., 2011; Feinstein et al., 2011, 2013, 2016; Becker et al., 2012; Terburg et al., 2012; Feinstein, 2013; Mihov et al., 2013)
Various circuits acting in parallel or collaboratively support learning when to perform behaviors that are typically referred to as 'fear behaviors'. Some have argued that these circuits represent distinct pathways from the amygdala to the periaqueductal gray through the hypothalamus to control different situation-specific fear behaviors, but others find that there are many (circuits) to one (behavior) mappings. Others find one (circuit) to many (behavior) mappings. Still others find that the amygdala, or specific parts (e.g. the basolateral nuclei), are not necessary for the expression of previously learned aversive responses (i.e. the way that learning is expressed depends on the context and the available options for behavior). Also, cortical regions (e.g. dmPFC and vmPFC) appear necessary for aversive learning when the context is more ecologically valid and less artificially simple. One thing is certain: scientists routinely engage in mental inference and refer to circuits as controlling different types of fear when in fact they are studying context-dependent behaviors that may not bear a one-to-one correspondence to fear.	Optogenetic research and some lesion research in rodents	(Furlong et al., 2010; Gross and Canteras, 2012; Herry and Johansen, 2014; Sharpe and Killcross, 2015; Tovote et al., 2015; McGaugh, 2016)
The expression of aversive learning depends on the state of the animal. For example, when the conditioned stimulus (a tone) is presented alone after having been paired with the unconditioned stimulus (an electric shock), the animal typically freezes, its heart rate increases, and its skin conductance goes up, which is usually taken as evidence that the animal has learned fear. Yet when an animal is restrained in position as it hears the tone, its heart rate decreases.	Classical conditioning in rats	(Iwata and LeDoux, 1988)
Adult monkeys with amygdala lesions are more likely to explore novel objects right away (which is usually interpreted as a 'lack of fear') but an alternative explanation is that the amygdala helps to regulate exploratory behavior in novel situations, which in turn will increase sensory processing when there is substantial prediction error. 'Fear' is not necessary. An amygdala might be required for aversive learning, but not the behavioral response learned (paralleling observations in rodent experiments).	Lesions in non-human primates	(Mason et al., 2006; Antoniadis et al., 2009)

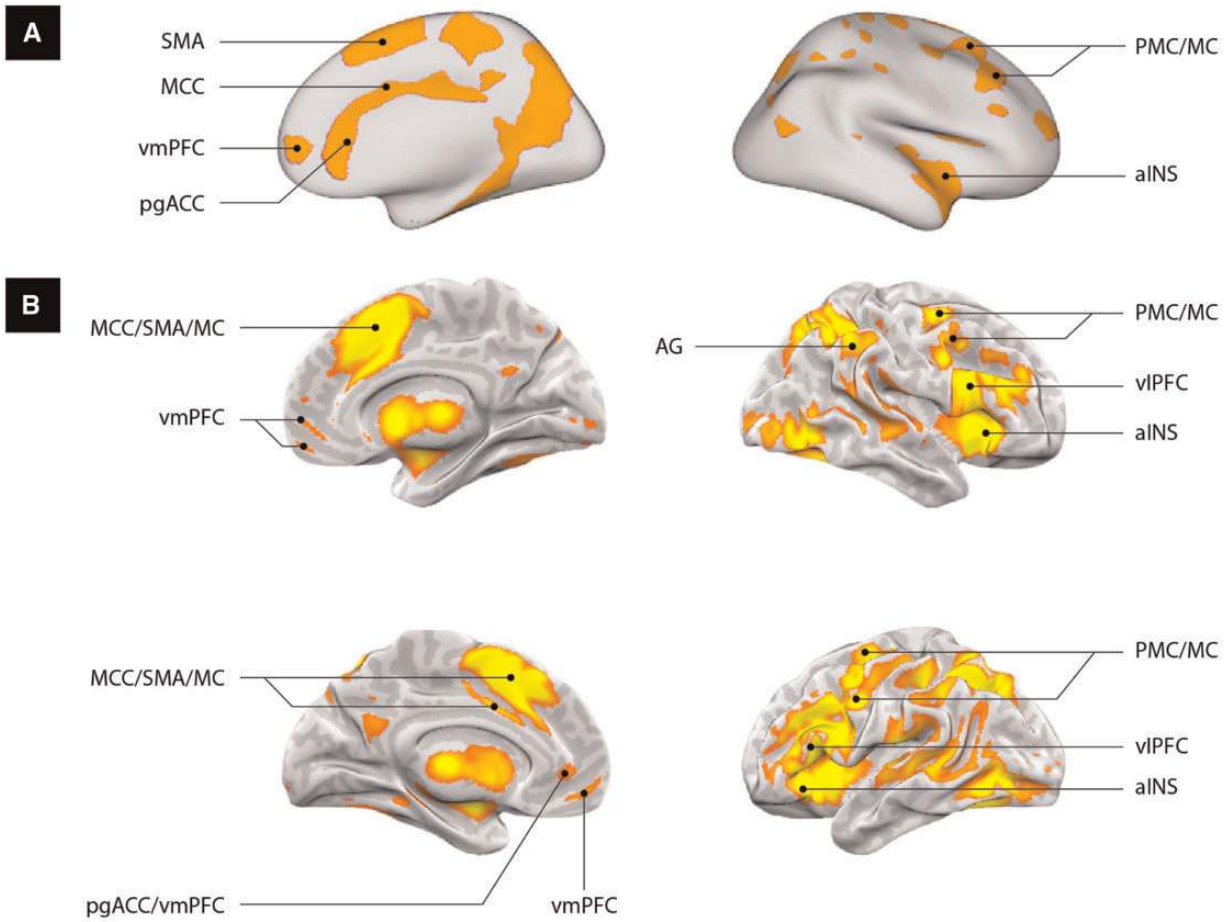

Fig. 2. Hubs in the human brain. **(A)** Hubs of the rich club, adapted from van den Heuvel and Sporns (2013). These regions are strongly interconnected with one another and it is proposed that they integrate information across the brain to create large-scale patterns of information flow (i.e. synchronized activity; van den Heuvel and Sporns, 2013). They are sometimes referred to as convergence or confluence zones (e.g. Damasio, 1989; Meyer and Damasio, 2009). **(B)** Results of a forward inference analysis, revealing 'hot spots' in the brain that show a better than chance increase in BOLD signal across 5633 studies from the Neurosynth database. Activations are thresholded at FWE $P < 0.05$. Limbic regions (i.e. agranular/dysgranular with descending projections to visceromotor control nuclei) include the cingulate cortex [midcingulate cortex (MCC), pregenual anterior cingulate cortex (pgACC)], ventromedial prefrontal cortex (vmPFC), supplementary motor and premotor areas (SMA and PMC), medial temporal lobe, the anterior insula (aINS) and ventrolateral prefrontal cortex (vlPFC) (e.g. Carrive and Morgan, 2012; Bar *et al.*, 2016); for a discussion and additional references, see (Kleckner *et al.*, in press). AG, angular gyrus; MC, motor cortex.

experience within the brain's synaptic connections, making those experiences available to guide later decisions about future expenditures and deposits. Too much of a resource (e.g. obesity in mammals) or not enough (e.g. fatigue, Dantzer *et al.*, 2014) is suboptimal. Prolonged imbalances can lead to illness (e.g. Hunter and McEwen, 2013; McEwen *et al.*, 2015) that remodels the brain (Crossley *et al.*, 2014; Goodkind *et al.*, 2015) and the sympathetic nervous system, with corresponding behavior changes (Sloan *et al.*, 2007; Capitanio and Cole, 2015; for a review, see Sloan *et al.*, 2008).

Whatever else your brain is doing—thinking, feeling, perceiving, emoting—it is also regulating your autonomic nervous system, your immune system and your endocrine system as resources are spent in seeking and securing more resources. All animal brains operate in the same manner (i.e. even insect brains coordinate visceral, immune and motor changes; Sterling and Laughlin, 2015, p. 91). This regulation helps explain why, in mammals, the regions that are responsible for implementing allostasis (the amygdala, ventral striatum, insula, orbitofrontal cortex, anterior cingulate cortex, medial prefrontal cortex (mPFC), collectively called 'visceromotor regions') are usually assumed to contain the circuits for emotion. In fact, many of these visceromotor regions are some of the most highly connected regions in the brain, and they exchange information with midbrain, brainstem, and spinal cord nuclei that coordinate autonomic, immune, and endocrine systems with one another, as well as with the systems that control skeletomotor movements and that process sensory inputs. Therefore these regions are clearly multipurpose when it comes to constructing the mental events that we group into mental categories (see Figure 2).

How does a brain perform allostasis?

For a brain to effectively regulate its body in the world, it runs an internal model of that body in the world.[5] In psychology, we refer to this modeling as 'embodied simulation' (Barsalou, 2008; Barsalou *et al.*, 2003) (e.g. see Figure 3). An internal model is metabolic investment, implemented by intrinsic activity (e.g.

5 There is a well-known principle of cybernetics: anything that regulates (i.e. acts on) a system must contain an 'internal model' of that system (Conant and Ross Ashby, 1970). From a brain's perspective, the 'system' in question includes it's body and it's ecological niche; a body must be watered, fed and cared for, so that a creature can grow, thrive, and ultimately, reproduce and care for it's young so as to pass its genes to the subsequent generation.

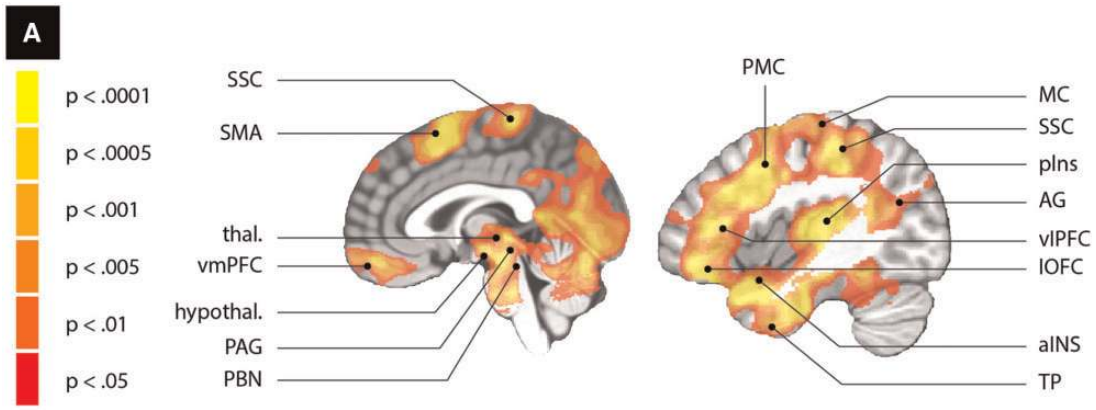

Fig. 3. Neural activity during simulation. $N = 16$ (data from Wilson-Mendenhall *et al.*, 2013). Participants listened with eyes closed to multimodal descriptions rich in sensory details and imagined each real-world scenario as if it was actually happening to them (i.e. the experiences were high in subjective realism). Contrast presented is scenario immersion > resting baseline; maps are FDR corrected $P < 0.05$. Left image, $x = 1$; right image, $x = -42$. Heightened neural activity in primary visual cortex (not labeled), somatosensory cortex (SSC), and MC during scenario immersion replicated prior simulation research (McNorgan, 2012) and established the validity of the paradigm. Notice that simulation was associated with an increase in BOLD response within primary interoceptive cortex (i.e. the pINS), in the sensory integration network of lateral orbitofrontal cortex (lOFC) (Ongur *et al.*, 2003) and in the thalamus; increased BOLD responses were also seen, as expected, in limbic and paralimbic regions such as the vmPFC, the aINS, the temporal pole (TP), SMA and vlPFC, as well as in the hypothalamus and the subcortical nuclei that control the internal milieu. PAG, periacquiductal gray; PBN, parabrachial nucleus.

Berkes *et al.*, 2011) that, in humans, occupies 20% of our total energy consumed (Raichle, 2010).[6] Given these considerations, modeling the world 'accurately' in some detached, disembodied manner would be metabolically reckless. Instead, the brain models the world from the perspective of its body's physiological needs.[7] As a consequence, a brain's internal model includes not only the relevant statistical regularities in the extrapersonal world, but also the statistical regularities of the internal milieu. Collectively, the representation and utilization of these internal sensations is called 'interoception' (Craig, 2015). Recent research suggests that interoception is at the core of the brain's internal model and arises from the process of allostasis (Barrett and Simmons, 2015; Chanes and Barrett, 2016; Barrett, 2017; for related discussions, see Seth *et al.*, 2012; Seth, 2013; Pezzulo *et al.*, 2015; Seth & Friston, 2016). Interoceptive sensations are usually experienced as lower dimensional feelings of affect (Barrett and Bliss-Moreau, 2009; Barrett, 2017). As such, the properties of affect—valence and arousal (Barrett and Russell, 1999; Kuppens *et al.*, 2013)—are basic features of consciousness (Damasio, 1999; Dreyfus and Thompson, 2007; Edelman and Tononi, 2000; James, 1890/2007; Searle, 1992, 2004; Wundt, 1897) that, importantly, are not unique to instances of emotion.

All animals run an internal model of their world for the purpose of allostasis (i.e. the notion of an internal model is species-general). Even single-celled organisms that lack a brain learn, remember, make predictions, and forage in service to allostasis (Freddolino and Tavazoie, 2012; Sterling and Laughlin, 2015). The content of any internal model is species-specific, however, including only the parts of the animal's physical surroundings that its brain has judged relevant for growth, survival and reproduction (i.e. a brain creates its affective niche in the present based on what has been relevant for allostasis in the past). Everything else is an extravagance that puts energy regulation at risk.

As an animal's integrated physiological state changes constantly throughout the day, its immediate past determines the aspects of the sensory world that concern the animal in the present, which in turn influences what its niche will contain in the immediate future. This observation prompts an important insight: neurons do not lie dormant until stimulated by the outside world, denoted as stimulus→response.[8] Ample evidence shows that ongoing brain activity influences how the brain processes incoming sensory information (e.g. Sayers *et al.*, 1974),[9] and that neurons fire intrinsically within large networks without any need for external stimuli (Swanson, 2012). The implications of these insights are profound: namely, it is very unlikely that perception, cognition, and emotion are localized in dedicated brain systems, with perception triggering emotions that battle with cognition to control behavior (Barrett, 2009). This means classical accounts of emotion, which rely on this S→R narrative, are highly doubtful.

An internal model is predictive, not reactive

An increasingly popular hypothesis is that the brain's simulations function as Bayesian filters for incoming sensory input, driving action and constructing perception and other psychological phenomena, including emotion. Simulations are thought to function as prediction signals (also known as 'top-down' or 'feedback' signals, and more recently as 'forward' models) that continuously anticipate events in the sensory environment.[10]

6 Long-range neural connections, like those that form the human brain's broadly distributed intrinsic networks, are particularly expensive (Bullmore and Sporns, 2012; Sterling and Laughlin, 2015), with most of the energy costs going to signaling between neurons, particularly in post-synaptic processes (Attwell and Laughlin, 2001; Attwell and Iadecola, 2002; Alle *et al.*, 2009; Harris *et al.*, 2012).

7 A trivial example, of course, is that infrared light is not normally something a human can see and so your brain (and mine) does not normally represent it. In this regard, we humans have been able to expand our ecological niche (and therefore our internal models) with technology.

8 This mistaken belief is an artifact of studying neurons in isolation (e.g. Hodgkin and Huxley, 1952), which creates a misleading picture of how the nervous system functions (Marder, 2011). For a similar view, see (Dewey, 1896).

9 Also see Makeig *et al.*, 2002, 2004; Mazaheri and Jensen, 2010; Laxminarayan *et al.*, 2011; Qian and Di, 2011; Scheeringa *et al.*, 2011.

10 The term 'feedback' derives from a time when the brain was thought to be largely stimulus driven (Sartorius *et al.*, 1993). Nonetheless, the

This hypothesis is variously called predictive coding, active inference, or belief propagation (e.g. Rao and Ballard, 1999; Friston, 2010; Seth *et al*., 2012; Clark, 2013a,b; Hohwy, 2013; Seth, 2013; Barrett and Simmons, 2015; Chanes and Barrett, 2016; Deneve and Jardri, 2016).[11] Without an internal model, the brain cannot transform flashes of light into sights, chemicals into smells and variable air pressure into music. You'd be experientially blind (Barrett, 2017). Thus, simulations are a vital ingredient to guide action and construct perceptions in the present.[12] They are embodied, whole brain representations that anticipate (i) upcoming sensory events both inside the body and out as well as (ii) the best action to deal with the impending sensory events. Their consequence for allostasis is made available in consciousness as affect (Barrett, 2017).

I hypothesize that, using past experience as a guide, the brain prepares multiple competing simulations that answer the question, 'what is this new sensory input most similar to?' (see Bar, 2009a,b). Similarity is computed with reference to the current sensory array and the associated energy costs and potential rewards for the body. That is, simulation is a partially completed pattern that can classify (categorize) sensory signals to guide action in the service of allostasis. Each simulation has an associated action plan. Using Bayesian logic (Deneve, 2008; Bastos *et al*., 2012), a brain uses pattern completion to decide among simulations and implement one of them (Gallivan *et al*., 2016), based on predicted maintenance of physiological efficiency across multiple body systems (e.g. need for glucose, oxygen, salt etc.).

From this perspective, unanticipated information from the world (prediction error) functions as feedback for embodied simulations (also known as 'bottom-up' or, confusingly, 'feedforward' signals). Error signals track the difference between the predicted sensations and those that are incoming from the sensory world (including the body's internal milieu). Once these errors are minimized, simulations also serve as inferences about the causes of sensory events and plans for how to move the body (or not) to deal with them (Lochmann and Deneve, 2011; Hohwy, 2013). By modulating ongoing motor and visceromotor actions to deal with upcoming sensory events, a brain infers their likely causes.

In predictive coding, as we will see, sensory predictions arise from motor predictions; simulations arise as a function of visceromotor predictions (to control your autonomic nervous system, your neuroendocrine system, and your immune system) and voluntary motor predictions, which together anticipate and prepare for the actions that will be required in a moment from now. These observations reinforce the idea that the stimulus→response model of the mind is incorrect.[13] For a given event, perception follows (and is dependent on) action, not the other way around. Therefore, all classical theories of emotion are called into question, even those that explain emotion as iterative stimulus →response sequences.

The computational architecture of the brain is a conceptual system plus pattern generators

The mechanistic details of predictive coding provide yet another deep insight: a brain implements its internal model with 'concepts' that 'categorize' sensations to give them meaning (Barrett, 2017).[14] Predictions are concepts (see Figure 4). Completed predictions are categorizations that maintain physiological regulation, guide action and construct perception. The meaning of a sensory event includes visceromotor and motor action plans to deal with that event. As detailed in Figure 5, meaning does not trigger action, but results from it. This makes classical appraisal theories highly doubtful, because they assume that a response derives from a stimulus that is evaluated for its meaning (e.g. Lazarus, 1991; Scherer, 2009; Roseman, 2011; for a discussion, see (Barrett *et al*., 2007b; Gross and Barrett, 2011). Appraisals as descriptions of the world (e.g. Clore and Ortony, 2008), however, are produced by categorization with concepts (e.g. Si *et al*., 2010).

Traditionally, a 'category' is a population of events or objects that are treated as similar because they all serve a particular goal in some context; a 'concept' is the population of representations that correspond to those events or objects.[15] I

history of science is laced with the idea that the mind drives perception [e.g. in the 11th century by Ibn al-Haytham who helped to invent the scientific method, in the 18th century by Kant (1781), and in the 19th century by Helmholtz]. In more modern times, see Craik's concept of internal models (1943), Tolman's cognitive maps (1948), Johnson-Laird's internal models, and for more recent references, Neisser (1967) and Gregory (1980). The novelty in recent formulations can be found in (i) the hypothesis that predictions are 'embodied' simulations of sensory-motor experiences, (ii) they are ultimately in the service of allostasis and therefore interoception is at their core and, of course (iii) the breadth of behavioral, functional, and anatomic evidence supporting the hypothesis that the brain's internal model implements active inference as prediction signals, including (iv) the specific computational hypotheses implementing a predictive coding account.

11 Notably, Buzsáki (2006) wrote that 'Brains are foretelling devices'. There is accumulating evidence that prediction and prediction error signals oscillate at different frequencies within the brain (e.g. Arnal and Giraud, 2012; Bressler and Richter, 2015; Brodski *et al*., 2015).

12 Simulations also constitute representations of the past (i.e. memories) and the future (i.e. prospections), and implement imagination, mind wandering, and daydreams (Schacter *et al*., 2007; Buckner *et al*., 2008).

13 For an excellent treatment of the conceptual baggage in words like 'organism', 'stimulus' and 'response', see Danziger (1997); in particular, see the thoughtful historical discussion of how motives and emotions became conceptualized as sources of 'internal' stimulation and how physiological changes became 'responses' to that stimulation.

14 For those who are unfamiliar with predictive coding, consider the problem of allostasis from the perspective of your own brain. For your entire life, your brain is entombed in a dark, silent box (i.e. a skull). It has to figure out the causes of sensory events outside your skull to guide action in the service of allostasis, but all it has access to their consequences in the form of sights, sounds, smells, touches, tastes and interoceptive sensations (i.e. sensations from your heart pumping, your lungs expanding, from inflammation, from metabolic processes and so on). So, your brain is faced with a problem of reverse inference: any given sensation—a flash of light or a sound or an ache or cramp—can have many different causes. In addition, the sensory information is dynamically changing, noisy, and ambiguous. Your brain solves this puzzle by using the only other source of information available to it—past experiences—to create simulations that predict incoming sensory events before their consequences arrive to the brain. In this way, your brain efficiently uses the statistical regularities from its past to anticipate future events that must be dealt with.

15 Traditionally, categories are supposed to exist in the world, whereas concepts are supposed to exist in the brain. This distinction makes sense for natural kind categories (where the boundaries exist independent of perceivers) or when the instances of a category share physical similarities – some set of statistical regularities in their sensory aspects or perceptual features (e.g., most human faces have a certain set of visual features – two eyes, two ears, a nose, some hair, and a mouth – in approximately the same orientation). In many cases, however, the boundary between a category and its concept is blurred. For example, consider a category whose instances share a similar function, but do not share any physical features (e.g., currency throughout the course of human history). These conceptual

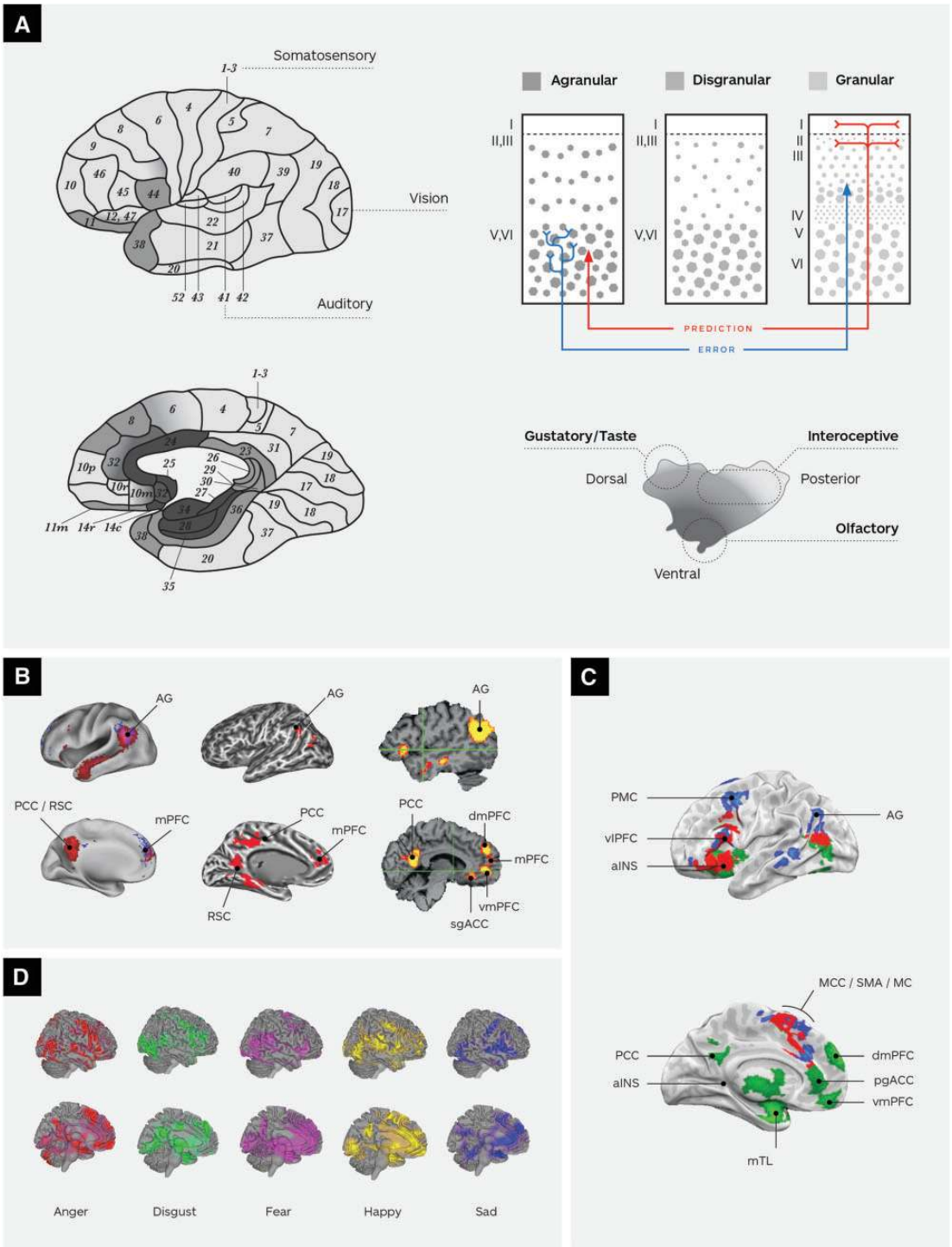

Fig. 4. The brain is a concept generator. **(A)** Brodmann areas are shaded to depict their degree of laminar organization, including the insula (bottom right). The brain's computational architecture is depicted (adapted from Barbas, 2015), where prediction signals flow from the deep layers of less granular regions (cell bodies depicted with triangles) to the upper layers of more granular regions; this, can also be thought of concept construction [as described in Barrett (2017)]. I hypothesize that agranular (i.e. limbic) cortices generatively combine past experiences to initiate the construction of embodied concepts; multimodal summaries cascade to sensory and motor systems to create the simulations that will become motor plans and perceptions. Prediction error processing, in turn, is akin to concept learning. The upper layers of cortex compress prediction errors and reduce error dimensionality, eventually creating multimodal summaries, by virtue of a cytoarchitectural gradient: prediction error flows from the upper layers of primary sensory and motor regions (highly granular cortex) populated with many small pyramidal cells with few connections towards less granular heteromodal regions (including limbic cortices) with fewer but larger pyramidal cells having many connections (Finlay and Uchiyama, 2015). **(B)** Evidence of conceptual processing in the default mode network: Multimodal summaries for emotion concepts [adapted from Skerry and Saxe (2015), Figure 1B]; summary representations of sensory-motor properties (color, shape, visual motion, sound and physical manipulation [Fernandino *et al.* (2016), Figure 5]; and, semantic processing [adapted from Binder and Desai (2011), Figure 2]. **(C)** Regions that consistently increase activity during emotional experience (green), emotion regulation (blue), and their overlap (red) [as appears in Clark-Polner *et al.* (2016); adapted from Buhle *et al.* (2014) and Satpute *et al.* (2015)]. Overlaps are observed in the aIns, vlPFC, the MCC, SMA and posterior superior temporal sulcus. Studies of emotional experience show consistent increase in activity that is consistent with manipulating predictions (i.e. the default mode and salience networks), whereas reappraisal instructions appear to manipulate the modification of those predictions (i.e. the frontoparietal and salience networks). **(D)** Intensity maps for five emotion categories examined by Wager *et al.* (2015). Maps represent the expected activations or population centers, given a specific emotion category. Maps also reflect expected co-activation patterns. Notice that population centers for all emotion categories can be found within the default mode and salience networks. These are probabilistic summaries, not brain states for emotion Adapted from Wager *et al.* (2015).

hypothesize that in assembling populations of predictions, each one having some probability of being the best fit to the current circumstances (i.e., Bayesian priors), the brain is constructing concepts (Barrett, 2017) or what Barsalou refers to as '*ad hoc*' concepts (Barsalou, 1983, 2003; Barsalou *et al*., 2003). In the language of the brain, a concept is a group of distributed 'patterns' of activity across some population of neurons. Incoming sensory evidence, as prediction error, helps to select from or modify this distribution of predictions, because certain simulations will better fit the sensory array (i.e. they will have stronger priors), with the end result that incoming sensory events are categorized as similar to some set of past experiences. This, in effect, is the original formulation of the conceptual act theory of emotion (see Barrett, 2006b, 2017): the brain uses emotion concepts to categorize sensations to construct an instance of emotion. That is, the brain constructs meaning by correctly anticipating (predicting and adjusting to) incoming sensations. Sensations are categorized so that they are (i) actionable in a situated way and therefore (ii) meaningful, based on past experience. When past experiences of emotion (e.g. happiness) are used to categorize the predicted sensory array and guide action, then one experiences or perceives that emotion (happiness).

In other words, an instance of emotion is constructed the same way that all other perceptions are constructed, using the same well-validated neuroanatomical principles for information flow within the brain. Barbas and colleagues' structural model of corticocortical connections (Barbas and Rempel-Clower, 1997; Barbas, 2015) provides specific hypotheses about how concepts categorize incoming sensory inputs to guide action and create perception, and in doing so fills the computational and neural gaps in my initial theoretical formulation of the theory, providing novel hypotheses about how a brain constructs emotional events [outlined in Barrett and Simmons (2015) and Chanes and Barrett (2016), Barrett (2017)]. The first key observation is that prediction signals are carried via 'feedback' connections that originate in cortical regions with the least well-developed laminar structure, referred to as 'agranular'. Agranular regions are cytoarchitecturally arranged to send but not receive prediction signals within the cerebral cortex. Another name for agranular cortices is 'limbic'. Limbic cortices, such as the anterior cingulate cortex and the ventral portion of the anterior insula (aINS), allostatically control physiology by relaying descending prediction signals to the internal milieu via a system of subcortical regions (Bar *et al*., 2016; Kleckner *et al*., in press), including the central nucleus of the amygdala (Ghashghaei *et al*., 2007), the ventral and dorsal striatum, and the central pattern generators (Swanson, 2012) across hypothalamus, the parabrachial nucleus, periaqueductal grey, and the solitary nucleus (see Figure 5A and B). Cortical regions with a dysgranular structure, which are referred to as limbic (Barbas, 2015) or paralimbic (Mesulam, 1998), also issue descending prediction signals to the body's internal milieu [e.g. midcingulate cortex, mPFC (MCC), ventrolateral prefrontal cortex (vlPFC), premotor cortex (PMC), etc., see Figure 5A and B]. My hypothesis is that these 'visceromotor' regions of the brain that are responsible for implementing allostasis, and that are usually assigned an emotional function, are 'driving' the perception signals, i.e. the 'concepts', that constitute the brain's internal model, in conjunction with the hippocampus (e.g. see Davachi and DuBrow, 2015; Hasson *et al*., 2015).

A concept is not only the descending prediction signals that control the viscera. It also includes the efferent copies of those signals that cascade to primary motor cortex (MC) as skeletomotor prediction signals, as well as to all primary sensory cortices as sensory prediction signals (see Figure 5C and D, respectively; for a discussion, see Bastos *et al*., 2012; Adams *et al*., 2013; Barbas, 2015; Barrett and Simmons, 2015; Chanes and Barrett, 2016). Following the evidence for how the cytoarchitectural gradients in the cortical sheet predict information flow across cortical regions (Barbas *et al*., 1997, p. 6608), prediction signals flow from deep layers of limbic cortices and terminate in the upper layers of cortical regions with more developed (i.e. more granular) structure, such as gustatory and olfactory cortex, primary MC, primary interoceptive cortex, and the primary visual, auditory and somatosensory regions.[16]

Because MC has a laminar organization that is less well developed than primary visual, auditory, somatosensory and interoceptive sensory regions (Barbas and García-Cabezas, 2015), I hypothesize that MC sends efferent copies to those sensory regions as sensory predictions (see Figure 5D, red lines). A similar arrangement between motor and somatosensory cortex has been proposed by Friston and colleagues (Adams *et al*., 2013). Furthermore, because of their differential laminar development, I hypothesize that primary interoceptive cortex in mid-to-posterior dorsal insula forwards sensory predictions to visual, auditory and somatosensory cortices (propagating across either a single or multiple synapses; Figure 5D, gold lines). The skeletomotor prediction signals prepare the body for movement, the interoceptive prediction signals initiate a change in affect (i.e. the expected sensory consequences of allostatic changes within the body's internal milieu), and the extrapersonal sensory prediction signals prepare upcoming perceptions. This hypothesis is consistent not only with over three decades of tract tracing studies in non-human animals, but also with engineering design principles (i.e. compute locally, and relay only the information that is needed to assemble a larger pattern; Sterling and Laughlin, 2015). Predictions literally change the firing of primary sensory and motor neurons, even though the incoming sensory input has not yet arrived (and may never arrive; e.g. Kok *et al*., 2016). Accordingly, all action and perception are created with concepts. All concepts contribute to allostasis and represent changes in affect, not just those that construct the events that feel affectively intense or are created with emotion concepts.

To consider how this works, try this thought experiment: in the past, you have experienced diverse instances of happiness, may be lying outdoors on a sunny day, finishing a strenuous workout, hugging a close friend, eating a piece of delectable chocolate or winning a competition. Each instance is different from every other, and when the brain creates a concept of happiness to categorize and make sense of the upcoming sensory

categories have also been called abstract or nominal categories. Biological categories are conceptual categories, as we learned in *On the Origin of Species*. The categories of social reality, such as flowers and weeds, or emotion categories, are conceptual because functions are imposed on physically disparate instances by virtual of collective agreement (Barrett, 2012).

16 Primary visual, auditory, and somatosensory cortices have the most developed laminar structure within the cerebral cortex and therefore receive prediction signals but are unable to send them. Primary interoceptive cortex is relatively less developed than these regions, and therefore sends multimodal sensory predictions to these exteroceptive regions. Primary motor cortex (with a small granular layer; (Barbas and García-Cabezas, 2015) has an even less well-developed laminar structure and therefore sends predictions to somatosensory cortex (Adams *et al*., 2013; Shipp *et al*., 2013) as well as all the primary sensory regions mentioned so far. Agranular limbic cortices send, but do not receive prediction signals, because they have the least well-developed laminar structure of the entire cortical mantle.

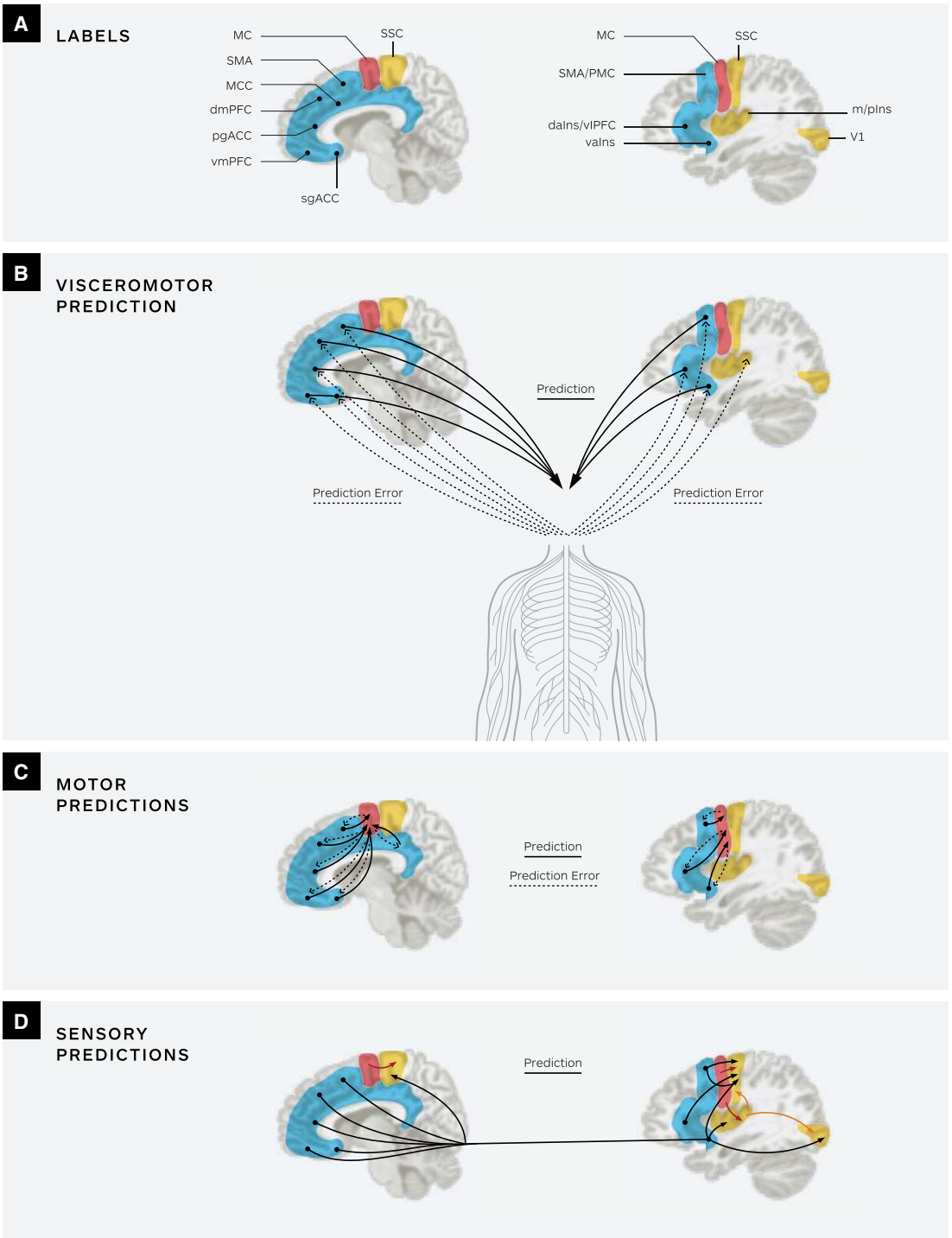

Fig. 5. A depiction of predictive coding in the human brain. **(A)** Key limbic and paralimbic cortices (in blue) provide cortical control the body's internal milieu. Primary MC is depicted in red, and primary sensory regions are in yellow. For simplicity, only primary visual, interoceptive and somatosensory cortices are shown; subcortical regions are not shown. **(B)** Limbic cortices initiate visceromotor predictions to the hypothalamus and brainstem nuclei (e.g. PAG, PBN, nucleus of the solitary tract) to regulate the autonomic, neuroendocrine, and immune systems (solid lines). The incoming sensory inputs from the internal milieu of the body are carried along the vagus nerve and small diameter C and Aδ fibers to limbic regions (dotted lines). Comparisons between prediction signals and ascending sensory input results in prediction error that is available to update the brain's internal model. In this way, prediction errors are learning signals and therefore adjust subsequent predictions. **(C)** Efferent copies of visceromotor predictions are sent to MC as motor predictions (solid lines) and prediction errors are sent from MC to limbic cortices (dotted lines). **(D)** Sensory cortices receive sensory predictions from several sources. They receive efferent copies of visceromotor predictions (black lines) and efferent copies of motor predictions (red lines). Sensory cortices with less well developed lamination (e.g. primary interoceptive cortex) also send sensory predictions to cortices with more well-developed granular architecture (e.g. in this figure, somatosensory and primary visual cortices, gold lines). For simplicity's sake, prediction errors are not depicted in panel D. sgACC, subgenual anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; pgACC, pregenual anterior cingulate cortex; dmPFC, dorsomedial prefrontal cortex; MCC, midcingulate cortex; vaIns, ventral anterior insula; daIns, dorsal anterior insula and includes ventrolateral prefrontal cortex; SMA, supplementary motor area; PMC, premotor cortex m/pIns, mid/posterior insula (primary interoceptive cortex); SSC, somatosensory cortex; V1, primary visual cortex; and MC, motor cortex (for relevant neuroanatomy references, see Kleckner *et al.*, in press).

events, it constructs a population of simulations (as potential actions and perceptions) according to the rules of Bayes' Theorem, whose priors reflect their similarity to the current situation (before the evidence is taken into account). The similarity need not be perceptual—it can be goal-based. So the brain constructs an on-line concept of happiness, not in absolute terms, but with reference to a particular goal in the situation (to be with friends, to enjoy a meal, to accomplish a task), all in the service of allostasis. This implies that 'happiness' has a specific meaning, but its specific meaning changes from one instance to the next.

As prediction signals cascade across the synapses of a brain, incoming sensory signals arriving to the brain (i.e. from the external environment and the internal periphery) simultaneously allow for computations of prediction error that are encoded to update the internal model (correcting visceromotor and motor action plans, as well as sensory representations; see Figure 5, dotted lines). Viscerosensory prediction errors arise from physiologic changes within the internal milieu and ascend via vagal and small diameter afferents in the dorsal horn of the spinal cord, through the nucleus of the solitary tract, the parabrachial nucleus, the periaqueductal gray and finally to the ventral posterior thalamus, before arriving in granular layer IV of the primary interoceptive insular cortex (Damasio and Carvalho, 2013; Craig, 2015). Notice that, in the context of this framework, perception (i.e. the 'meaning' of sensory inputs) is constructed with reference to allostasis, and sensory prediction errors are treated, at a very basic level, as information that guides a 'predicted' visceromotor and motor action plan.

Prediction errors also arise within the amygdala, the basal ganglia, and the cerebellum and are forwarded to the cortex to correct its internal model (see (Beckmann *et al.*, 2009; Buckner *et al.*, 2011; Haber and Behrens, 2014; Kleckner *et al.*, in press). I hypothesize that information from the amygdala to the cortex is not 'emotional' *per se*, but signals uncertainty (Whalen, 1998) about the predicted sensory input (via the basolateral complex) and helps to adjust allostasis (via the central nucleus) as a result.[17] The arousal signals that are associated with increases in amygdala activity (e.g. Wilson-Mendenhall *et al.*, 2013) can be considered a learning signal (Li and McNally, 2014). Similarly, prediction errors from the ventral striatum to the cortex (referred to as 'reward prediction errors; Schultz, 2016) convey information about sensory inputs that impact allostasis more than expected (i.e. that this information should be encoded and consolidated in the cortex, and acted upon in the moment). Dopamine is associated with engaging in vigorous action and learning that is necessary to achieve the rewards that maintain efficient allostasis (or restore it in the event of disruption), rather than playing a necessary or sufficient role in rewards themselves (Salamone and Correa, 2012; Guitart-Masip *et al.*, 2014). Other neuromodulators, such as opioids, seem to be more intrinsic to reward in that regard (e.g. Fields and Margolis, 2015).

17 Even more interestingly, there is some evidence to suggest that the cortical regions projecting to the brainstem nuclei which originate these neuromodulators (such as the locus coeruleus for norepinephrine) are largely entrained by limbic cortical regions via descending visceromotor predictions that project directly from the anterior cingulate cortex and dorsomedial prefrontal cortex, as well as indirectly via projections from the central nucleus of the amygdala and the hypothalamus (see Counts and Mufson, 2012). The locus coeruleus also receives ascending interoceptive and nociceptive prediction errors (see Counts and Mufson, 2012). This is yet another way that allostasis is altered by modulating the gain or excitability of neurons that represent sensory and motor prediction errors.

The cerebellum models prediction errors from the periphery and relays them to cortex to modify motor predictions [i.e. it predicts the sensory consequences of a motor command much faster than actual sensory prediction errors can manage, (Shadmehr *et al.*, 2010), and helps the cortex reduce the sensory consequences caused by one's own movements]. The same may be true for visceromotor predictions, given the connectivity between the cerebellum and the cingulate cortex, hypothalamus and amygdala (Schmahmann and Pandya, 1997; Strick *et al.*, 2009; Schmahmann, 2010; Buckner *et al.*, 2011).[18] This would give the cerebellum a major role in allostasis, concept generation, and the construction of emotion (e.g. see meta-analytic evidence in Kober *et al.*, 2008; Lindquist *et al.*, 2012; Wager *et al.*, 2015).

A brain implements an internal model of the world with concepts because it is metabolically efficient to do so. Even before birth, a brain begins to build its internal model by processing prediction error from the body and the world (see Barrett, 2017 for discussion). Prediction errors (i.e. unanticipated sensory inputs) cascade in a feedforward cortical sweep, originating in the upper layers of cortices with more developed laminar organization and terminating in the deep layers of cortices with less well-developed lamination. As information flows from sensory regions (whose upper layers contain many smaller pyramidal neurons with fewer connections) to limbic and other heteromodal regions in frontal cortex (whose upper layers contain fewer but larger pyramidal neurons with many more connections, see Figure 4A), it is compressed and reduced in dimensionality (Finlay and Uchiyama, 2015). This dimension reduction allows the brain to represent a lot of information with a smaller population of neurons, reducing redundancy and increasing efficiency, because smaller populations of neurons are summarizing statistical regularities in the spiking patterns in larger populations with in the sensory and motor regions. Additional efficiency is achieved because conceptually similar representations reuse neural populations during simulation (e.g. Rigotti *et al.*, 2013). As a result, different predictions are separable, but are not spatially separate (i.e. multimodal summaries are organized in a continuous neural territory that reflects their similarity to one another). Therefore, the hypothesis is that all new learning (e.g. the processing of prediction error) is concept learning, because the brain is condensing redundant firing patterns into more efficient (and cost-effective) multimodal summaries. This information is available for later use by limbic cortices as they generatively initiate prediction signals, constructed as low-dimensional, multimodal summaries (i.e. 'abstractions'); these summaries, consolidated from prior encoding of prediction errors, become more detailed and particular as they propagate out to more architecturally granular sensory and motor regions to complete embodied concept generation.

Taking a network perspective

In a keynote address in 2006, I first proposed that several of the brain's intrinsic networks (what would come to be called the default mode, salience, and frontoparietal control networks) are domain-general or multi-use networks that are involved in constructing emotional episodes (e.g. see Figure 6 in Kober *et al.*, 2008). Building on the findings so far, as well as the anatomical distribution of limbic cortices within the brain (see Figure 5), I

18 In a fly brain, the mushroom bodies may play an analogous role in predictive coding (Sterling and Laughlin, 2015)

have refined these hypotheses (see Figure 6). I hypothesize, as others do (Mesulam, 2002; Hassabis and Maguire, 2009; Buckner, 2012), that the default mode network is necessary for the brain's internal model. Regardless of the other mental categories mapped to default mode network activity, the simulations initiated within this network cascade to create concepts that eventually categorize sensory inputs and guide movements in the service of allostasis. This hypothesis is partially consistent with the hypothesis that the default mode network represents semantic concepts (Binder *et al.*, 2009; Binder and Desai, 2011) (see Figure 4B). I hypothesize that the default mode network hosts 'part' of their patterns, but simulations are more than just multimodal sensorimotor summaries; they are fully embodied brain states. They emerge as default mode summaries cascade out to primary sensory and motor regions to become detailed and particularized [i.e. to modulate the spiking patterns of sensory and motor neurons (Barrett, 2017); for supporting evidence on embodied representations of concepts, see (Pulvermüller, 2013; Fernandino *et al.*, 2015, 2016; Barsalou, 2016).[19]

I further hypothesize that the salience network tunes the internal model by predicting which prediction errors to pay attention to [i.e. those errors that are likely to be allostatically relevant and therefore worth the cost of encoding and consolidation; called precision signals (Feldman and Friston, 2010; Clark, 2013a,b; Moran *et al.*, 2013; Shipp *et al.*, 2013)].[20] Specifically, I hypothesize that precision signals optimize the sampling of the sensory periphery for allostasis, and they are sent to every sensory system in the brain (for anatomical and functional justifications, see (Chanes and Barrett, 2016)]. They directly alter the gain on neurons that compute prediction error from incoming sensory input (i.e. they apply attention) to signal the degree of confidence in the predictions (i.e. the priors), confidence in the reliability or quality of incoming sensory signals, and/or predicted relevance for allostasis. Unexpected sensory inputs that are anticipated to have resource implications (i.e. are likely to impact survival, offering reward or threat, or are of uncertain value) will be treated as 'signal' and learned (i.e. encoded) to better predict energy needs in the future, with all other prediction error treated as 'noise' and safely ignored (Li and McNally, 2014; for discussion, see Barrett, 2017). Limbic regions within the salience network may also indirectly signal the precision of incoming sensory inputs via their modulation of the reticular nucleus that encircles that thalamus and controls the sensory input that reaches the cortex via thalamocortical pathways [for relevant anatomy, see (Zikopoulos and Barbas, 2006, 2012; John *et al.*, 2016)].[21] My hypothesis, then, is that cortical limbic regions within the salience network are at the core of the brain's ability to adjust its internal model to the conditions of the sensory periphery, again in the service of allostasis (e.g. see Figure 6) This is consistent with the salience network's role in attention regulation; e.g. (Power *et al.*, 2011; Touroutoglou *et al.*, 2012; Ullsperger *et al.*, 2014; Uddin, 2015).

In addition, I hypothesize that neurons with the frontoparietal control network sculpt and maintain simulations for longer than the several hundred milliseconds it takes to process imminent prediction errors), and they may also help to suppress or inhibit simulations whose priors are very low (because those priors are influenced not only by the current sensory array, but also by what the brain predicts for the future). It pays to be flexible, to be able to construct and use patterns that extend over longer periods of time (different animals have different timescales that are relevant to their behavioral repertoire and ecological niche). It's also valuable to learn on a single trial, without being guided by recurring statistical regularities in the world, particularly if you reside in a quickly changing environment. As a prediction generator, the brain is constructing simulations (as concepts) across many different timescales (i.e. integrating information across the few moments that constitute an event, but also across longer time frames at various scales; for similar ideas, see Wilson *et al.*, 2010; Hasson *et al.*, 2015). Therefore, a brain may be pattern matching to categorize not only on short processing timescales of milliseconds but also on much longer timescales (seconds to minutes to hours or even longer). The lesson here, for the science of emotion, is that the brain does not process individual stimuli—it processes events across temporal windows. Emotion perception is event perception, not object perception.[22]

The theory of constructed emotion

Now we can see how a multi-level, constructionist view like the theory of constructed emotion offers an approach to understanding the brain basis of emotion that is consistent with emerging computational and evolutionary biological views of the nervous system.[23] A brain can be thought of as running an internal model that controls central pattern generators in the service of allostasis (for more on pattern generators, see Burrows, 1996; Sterling and Laughlin, 2015; Swanson, 2000). An internal model runs on past experiences, implemented as concepts. A concept is a collection of embodied, whole brain representations that predict what is about to happen in the sensory environment, what the best action is to deal with impending events, and their consequences for allostasis (the latter is made available to consciousness as affect). Unpredicted information (i.e. prediction error) is encoded and consolidated whenever it is predicted to result in a physiological change in state of perceiver (i.e. whenever it impacts allostasis). Once prediction error is minimized, a prediction becomes a perception or an experience. In doing so, the prediction explains the cause of sensory events and directs action; i.e. it categorizes the sensory event. In this way, the brain uses past experience to construct a

19 Notice that this hypothesis is species-general: rats have a default mode network and are not able to engage in mental time travel as far as we can tell (Hsu *et al.*, 2016), but this in no way disconfirms the hypothesis that the network is running an internal model of the animal's world in the service of allostasis.

20 This allows for the encoding of statistical patterns of uncertain value that can later be reconstructed when they are of use (Dunsmoor *et al.*, 2015).

21 Salience regions may also play a role in maintaining the internal model that is unconstrained by the sensory world (such as when constructing memories, imaginings, dreams, reveries, mind-wandering and so on). Salience regions also help accomplish multimodal integration [compare, e.g. the topography of the salience network and the multimodal integration network found in Sepulcre *et al.* (2012)].

22 The frontopartial control network (which contains key limbic rich club hubs in the midcingulate cortex and anterior insula) may also have a role to play in managing sensory prediction errors, by applying attention to select those body movements that will generate the expected sensory inputs, presumably with help from cerebellar and striatal prediction errors.

23 The theory of constructed emotion integrates ideas and empirical findings from neuroconstruction (Mareschal *et al.*, 2007; Westermann *et al.*, 2007; Karmiloff-Smith, 2009), rational constructivism (Xu and Kushnir, 2013), psychological construction (Russell, 2003; Barrett, 2006b, 2012, 2013; Barrett *et al.*, 2015) and social construction (Boiger and Mesquita, 2012), as well as descriptive appraisal theories (e.g., Clore and Ortony, 2008).

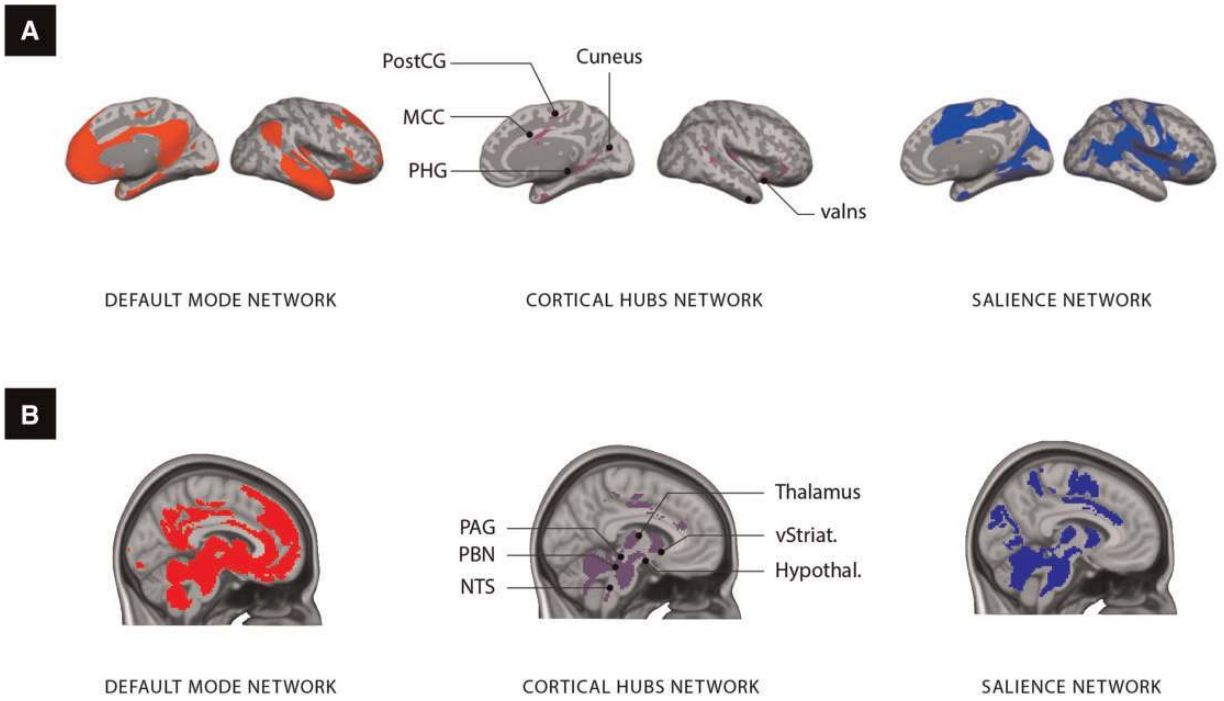

Fig. 6. A large-scale system for allostasis and interoception in the human brain. **(A)** The system implementing allostasis and interoception is composed of two large-scale intrinsic networks (shown in red and blue) that are interconnected by several hubs (shown in purple; for coordinates, see Kleckner *et al.*, in press). Hubs belonging to the 'rich club' are labeled. These maps were constructed with resting state BOLD data from 280 participants, binarized at $p < 10^{-5}$, and then replicated on a second sample of 270 participants. vaIns, ventral anterior insula; MCC, midcingulate cortex; PHG, parahippocampal gyrus; PostCG, postcentral gyrus; PAG, periaqueductal gray; PBN, parabrachial nucleus; NTS, the nucleus of the solitary tract; vStriat., ventral striatum; Hypothal, hypothalamus. **(B)** Reliable subcortical connections, thresholded $P < 0.05$ uncorrected, replicated in 270 participants.

categorization [a situated conceptualization; (Barsalou, 1999; Barsalou *et al.*, 2003; Barrett, 2006b; Barrett *et al.*, 2015)] that best fits the situation to guide action. The brain continually constructs concepts and creates categories to identify what the sensory inputs are, infers a causal explanation for what caused them, and drives action plans for what to do about them. When the internal model creates an emotion concept, the eventual categorization results in an instance of emotion.

This hypothesis is consistent with the conceptual innovations in Darwin's *On the Origin of Species* [(Darwin, 1859/1964), even as it is inconsistent with '*The Expression of the Emotions in Man and Animals*' (Darwin, 1872/2005); for a discussion, see Barrett, 2017]. Some of the psychological constructs used in the theory of constructed emotion are species-general (e.g. allostasis, interoception, affect, and concept), while others require the capacity for certain types of concepts (Barrett, 2012) and are more species-specific (e.g. emotion concepts). It is necessary to understand which constructs are species-general vs. species-specific to solve the puzzle of the biological basis of emotion. Mistaking one for the other is a category error that interferes with scientific progress.

Constructionism, as a scientific paradigm, makes different assumptions than the classical paradigm (Barrett, 2015), asks different questions, and requires different methods and analytic procedures than those of the classical view (whose methods are ill-suited to testing it). As a consequence, constructionism is often profoundly misunderstood [for two recent examples, see (Anderson and Adolphs, 2014; Kragel and LaBar, 2016)]. With these observations in mind, here is a partial list of claims I am not making, to avoid further confusion:

i. I am not saying that emotions are illusions. I'm saying that emotion categories don't have distinct, dedicated neural essences. Emotion categories are as real as any other conceptual categories that require a human perceiver for existence, such as 'money' (i.e. the various objects that have served as currency throughout human history share no physical similarities; Barrett, 2012, 2017).
ii. I am not saying that all neurons do everything (a.k.a. equipotentiality). I am suggesting that a given neuron does more than one thing (has more than one receptive field), and that there are no emotion-specific neurons.
iii. I am not claiming that networks are Lego blocks with a static configuration and an essential function. I am suggesting that, when it comes to understanding the physical basis of psychological categories, it is necessary to focus on ensembles of neurons rather than individual neurons. A neuron does not function on its own, and many neurons are part of more than one network. Moreover, networks function via degeneracy, meaning that a given network has a repertoire of functional configurations (i.e. functional motifs) that is constrained by its anatomical structure (i.e. its structural motif).
iv. I am not claiming that subcortical regions are irrelevant to emotion. I hypothesize that an instance of emotion is a brain state that makes the sensory array meaningful, and in so doing engages the pattern generators for whatever actions are functional in the context, given a person's current state.
v. I am not saying that the default mode and salience networks implement allostasis and therefore should not be mapped to other psychological categories. I am claiming that these (and other) domain-general networks can be mapped to many psychological categories at the same time.

vi. I am not saying that concepts are stored in the default mode network. I'm saying that the default mode network represents efficient, multimodal summaries, from which a cascade of predictions issues through the entire cortical sheet, terminating in primary sensory and motor regions. The whole cascade is an instance of a concept.
vii. I am not saying that emotions are deliberate, nor denying that automaticity exists. I am saying that in humans, actual executive control (e.g. via the frontoparietal control network in primates) and the experience of feeling in control are not synonymous (Barrett *et al.*, 2004). All animal brains create concepts to categorize sensory inputs and guide action in an obligatory and automatic way, outside of awareness. Automaticity and control are different brain modes (each of which can be achieved with a variety of network configurations), not two battling brain systems.
viii. I am not saying that non-human animals are emotionless. I'm saying that emotion is perceiver-dependent, so questions about the nature of emotion must include a perceiver. 'Is the fly fearful?' is not a scientific question, but 'Does a human perceive fear in the fly?' and 'Does the fly feel fear?' can be answered scientifically (and the answers are 'yes' and 'no'). Notice that I am not claiming that a fly feels nothing; it may feel affect (Barrett, 2017).

Selected implications of the theory

Scientific revolutions are difficult. At the beginning, new paradigms raise more questions than they answer. They may explain existing anomalies or redefine lingering questions out of existence, but they also introduce a new set of questions that can be answered only with new experimental and computational techniques. This is a feature, not a bug, because it fosters scientific discovery (Firestein, 2012). A new paradigm barely gets started before it is criticized for not providing all the answers. But progress in science is often not answering old questions but asking better ones. The value of a new approach is never based on answering the questions of the old approach.

Table 2. Selected neuroscience evidence supporting the theory of constructed emotion

Observation	Method	Example Citations
Degeneracy: mapping many neurons, regions, networks or patterns to one emotion category	Human neuroimaging: task-related data	(Vytal and Hamann, 2010; Lindquist et al., 2012; Wilson-Mendenhall et al., 2011, 2015; Oosterwijk et al, 2015)
Degeneracy: mapping many neurons, regions, networks or patterns to one emotion category	Human neuroimaging: multi-voxel pattern analysis	(Clark-Polner, Johnson and barrett, 2016); compare the different patterns for a given emotion category across (Kragel and LaBar, 2015; Wager et al., 2015; Saarimaki et al., 2016)
Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category	Intracranial stimulation in humans	(Guillory and Bujarski, 2014)
Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category	Behavioral observations in humans with amygdala lesions	(Becker et al., 2012; Mihov et al., 2013)
Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category	Optogenetic research showing many to one mappings for behaviors in rodents	(Herry and Johansen, 2014)
Neural reuse: Mapping one neural assembly to many emotion categories	Human neuroimaging: task-related data	(Vytal and Hamann, 2010; Wilson-Mendenhall et al., 2011; Lindquist et al., 2012)
Neural reuse: Mapping one neural assembly to many emotion categories	Human neuroimaging: intrinsic connectivity data	(Wilson-Mendenhall et al., 2011; Barrett and Satpute, 2013; Touroutoglou et al., 2015)
Neural reuse: Mapping one neural assembly to many emotion categories	Optogenetic research and some lesion research in rodents	(Tovote et al., 2015)
Predictive coding explains aversive ('fear') learning	Optogenetic research and some lesion research in rodents	(Furlong et al., 2010; McNally et al., 2011; Li and McNally, 2014)
Emotion concepts are embodied	Human neuroimaging: task-related data	(Oosterwijk et al., 2012, 2015)
Multimodal summaries of emotion concepts are represented in the default mode network	Human neuroimaging: task-related data	(Peelen et al., 2010; Skerry and Saxe, 2015)
Default mode and salience network interconnectivity is associated with the intensity of emotional experiences (as distinct from arousal)	Human neuroimaging: task-related data	(Raz *et al.*, 2016)
Embodied simulations are associated with increased activity in primary interoceptive cortex	Human neuroimaging: task-related data	(Wilson-Mendenhall *et al.*, under review)

Box 1. The curious case of SM

Much of our understanding of the neural basis of fear comes from studying Patient S.M. She has difficulty experiencing fear in many normative circumstances (e.g. horror movies, haunted houses), but she experiences intense fear during experiments where she is asked to breathe air with higher concentrations of CO_2. She is able to mount a normal skin conductance response to an unexpectedly loud sound, but her brain seems not to use arousal as a learning cue in mild situations (e.g. standard 'fear learning' paradigms) and she therefore has difficulty learning from prior errors (e.g. she does not mount an anticipatory skin conductance response to aversive stimuli, during the Iowa Gambling Task, and she does not show loss aversion when gambling). Interestingly, however, there is other evidence that S.M. is capable of what is typically called 'fear learning'. S.M. is averse to breaking the law for fear of getting in trouble. She also spontaneously reports feeling worried. She is able 'learn fear' in the real world (e.g. she is averse to seeking medical treatment or visiting the dentist because of pain she experienced on a previous occasions).

S.M.'s impairments in fear perception appear to be clearest in experiments where she is asked to view stereotyped, fear poses and explicitly categorize them as fearful (although she shows more widespread deficits in explicitly perceiving arousal in posed faces, and she appears to have no difficulty rapidly processing fear faces outside of consciousness). She can perceive fear in bodies and voices (as is evidenced by her efforts to help her friend or call the police for others in danger). S.M. can even perceive fear in posed faces when her attention is directed to the eyes of the stimulus face, consistent with evidence that some amygdala neurons are particularly sensitivity to the sclera of eyes (the faces that depict stereotyped fear poses contain widen eyes). By contrast, other patients with Urbach-Wiethe Disease spend a longer time looking at the widened eyes of stereotyped fear poses and have no difficulty correctly categorizing those faces as fearful.

It should be noted (but rarely is) that S.M.'s brain shows abnormalities that extend beyond the amygdala, including the anterior entorhinal cortex and ventromedial prefrontal cortex, both of which show dense, reciprocal connections to the amygdala and very likely play a role in S.M.'s specific behavioral profile. It is also important to note that S.M. has had difficulties sustaining long-term relationships, including friendships, and is distressed by this. There seems to be one clear exception: S.M. has been able to maintain a relationship with the scientists she has worked with for almost two decades. S.M. calls them for support (e.g. when she is worried or afraid, such as when did not want to return for painful medical treatment). They help her with the details of daily life (financial and otherwise). It would be interesting to examine whether S.M. is aware of their hypothesis that the amygdala contains the circuitry for fear.

For references, see Table 1, and also Feinstein *et al.* (2016).

Such is the case with the theory of constructed emotion. Evidence from various domains of research is consistent with the proposed hypotheses (for select neuroscience examples, see Table 2), even as it casts aside some of the old unanswered questions of the classical view.

Ironically, perhaps the strongest evidence to date for the theory comes from studies that use pattern classification to distinguish categories of emotion. Several recent articles taking this approach have reported success in differentiating one emotion category from another—a finding that is routinely construed as providing the long awaited support for the classical view (Kassam *et al.*, 2013; Kragel and LaBar, 2015; Saarimaki *et al.*, 2016). However, patterns that distinguish among the categories in one study do not replicate in the other studies. The same is true for studies that successfully created different patterns of autonomic physiology, despite using the same stimuli and experimental method, and sampling from the same population (e.g. Stephens *et al.*, 2010; Kragel and LaBar, 2013).

Generally speaking, pattern classification results in the science of emotion are routinely misinterpreted. A pattern that diagnoses sadness is not the brain state for sadness but merely a statistical summary of a highly variable set of instances. To assume otherwise is an essentialist error that mistakes a statistical summary for the norm.[24] Indeed, the voxels that make up a pattern for a category need not observed in every (or even any) single instance of that category. A classic study by Posner and Keele (1968) demonstrated a similar general phenomenon almost half a century ago, and we have confirmed this with a simple mathematical simulation (see Clark-Polner, Johnson and Barrett, 2016).

The theory of constructed emotion is consistent with the older literature on decorticate animals that appears to support the classical view. For example, consider the experiment by Woodworth and Sherrington (1904), who surgically removed the cerebral cortex, thalamus, and hypothalamus of cats, and observed what they referred to as 'pseud-affective' (for pseudo-affective) reflexes—reflexive motor actions were left intact but the 'affective' (i.e. allostatic) driver of these responses was gone. As a consequence, these animals appeared to behave emotionally, but the actions were no longer in service of survival. These findings can be interpreted as demonstrating the existence of pattern generators when the machinery of the conceptual system has been removed. A similar observation can be made for Cannon and Britton's (1925) 'sham rage' where a decorticated cat, upon waking from anesthesia, spontaneously spit, clawed, and arched its back. Importantly, Cannon referred to these actions as 'fury' and in doing so, inferred the presence of a mental state from a set of actions.

Scientists carefully map the circuitry for behaviors (or mere movements in some cases) in non-human animals, but some mistakenly believe that they are mapping the circuitry for emotions. An example is observing freezing behavior in a rat (e.g. in response to electric shock) and calling it 'fear'. Rats don't freeze only in situations that we perceive as threatening (i.e. one behavior maps to many categories), and rats exhibit a variety of behaviors in threatening situations (i.e. many behaviors map to one category). This error, assuming that an action is equivalent to an emotion, which I call the mental inference fallacy (Barrett,

24 The average size of a US family in 2015 was 3.14 persons, but that does not mean that every family (or even any family) contained 3.14 people.

2017), has wreaked havoc with the scientific accumulation of knowledge about emotion (also see Barrett, 2012; LeDoux, 2012). Motor movements do not provide a direct indication of an internal state, be it in a rodent, a monkey or a human [e.g. see decades of research on mental inference processes (Gilbert, 1998) and opacity of mind (Robbins and Rumsey, 2008)].

When viewed in this light, it is an error to claim that studies of the human brain using functional magnetic resonance imaging (fMRI) yield different results than studies of non-human animal brains with lesions or optogenetics (e.g. Adolphs, 2013). In reality, all are making physical measurements and mapping emotion concepts to them, and no set of findings is described appropriately by classical emotion concepts. For example, in an attempt to deal with all the variation within the category 'fear', scientists create finer-grained typologies in an attempt to bring nature under control and make it easier to identify their essences (e.g. Gross and Canteras, 2012). But this does not avoid the problem of the mental state fallacy. Furthermore, it makes no sense to elevate categories for anger, sadness, fear, disgust and happiness to a common ethological framework for comparing humans with other animals, when there is ample evidence from linguistics, anthropology and psychology that these categories do not offer a robust, universal framework for comparing humans of different cultures (Russell, 1991; Gendron *et al.*, 2014a,b, 2015; Wierzbicka, 2014; Crivelli *et al.*, 2016).

Conclusions and future directions

Scientists must abandon essentialism and study emotions in all their variety. We must not merely focus on the few stereotypes that have been stipulated based on a very selective reading of Darwin. We must assume variability to be the norm, rather than a nuisance to be explained after the fact. It will never be possible to measure an emotion by merely measuring facial muscle movements, changes in autonomic nervous system signals, or neural firing within the periaqueductal gray or the amygdala. To understand the nature of emotion, we must also model the brain systems that are necessary for making meaning of physical changes in the body and in the world.

This article is a mere sketch of a much larger scientific landscape. The theory of constructed emotion proposes that emotions should be modeled holistically, as whole brain-body phenomena in context. My key hypothesis is that the dynamics of the default mode, salience and frontoparietal control networks form the computational core of a brain's dynamic internal working model of the body in the world, entraining sensory and motor systems to create multi-sensory representations of the world at various time scales from the perspective of someone who has a body, all in the service of allostasis (for evidence consistent with this view, see van den Heuvel *et al.*, 2012; van den Heuvel and Sporns, 2011, 2013). In other words, allostasis (predictively regulating the internal milieu) and interoception (representing the internal milieu) are at the anatomical and functional core of the nervous system. These insights offer a range of new hypotheses—e.g. that reappraisal and other regulation processes (Etkin *et al.*, 2015; Gross, 2015) are accomplished with predictions that categorize sensory inputs and control action with concepts (see Figure 4C).

The theory of constructed emotion also views the distinction between the central and peripheral nervous systems as historical rather than as scientifically accurate. For example, ascending interoceptive signals bring sensory prediction errors from the internal milieu to the brain via lamina I and vagal afferent pathways, and they are anatomically positioned to be modulated by descending visceromotor predictions that control the internal milieu (e.g. Fields, 2004). This suggests the hypothesis that concepts (i.e. prediction signals) act like a volume dial to influence the processing of prediction errors before they even reach the brain. This provides new hypotheses about the chronification of pain (see Barrett, 2017) that considers pain and emotion as two sides of the same coin, rather than separate phenomena that influence one another.

Emotions are constructions of the world, not reactions to it. This insight is a game changer for the science of emotion. It dissolves many of the debates that remained mired in philosophical confusion, and allows us to better understand the value of non-human animal models, without resorting to the perils of essentialism and anthropomorphism. It provides a common framework for understanding mental, physical, and neurodegenerative disorders (e.g., Barrett and Simmons, 2015; Barrett, Quigley & Hamilton, 2016; Barrett, 2017), and collapses the artificial boundaries between cognitive, affective, and social neurosciences (see Barrett & Satpute, 2013). Ultimately, the theory of constructed emotion equips scientists with new conceptual tools to solve the age-old mysteries of how a human nervous system creates a human mind.

Funding

This article was prepared with support from the National Institute on Aging (R01 AG030311), the National Cancer Institute (U01 CA193632), the National Science Foundation (CMMI 1638234) and the US Army Research Institute for the Behavioral and Social Sciences (W911NF-15-1-0647 and W911NF- 16-1-0191). The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Conflict of interest. None declared.

Glossary

Agranular: Cerebral cortex with the least developed laminar organization involving no definable layer IV, and no clear distinction between the neurons in layers II and III.

Allostasis: Regulating the internal milieu by anticipating physiological needs and preparing to meet them before they arise.

Concept: Traditionally, a category is a group of instances that are similar for some function or purpose; a concept is the mental representations of those category members. In the theory of constructed emotion, a concept is a collection of embodied, whole brain representations that predicts what is about to happen in the sensory environment, what the best action is to deal with these impending events, and their consequences for allostasis.

Degeneracy: Degeneracy refers to the capacity for biologically dissimilar systems or processes to give rise to identical functions. Degeneracy is different from redundancy (which is inefficient and to be avoided).

Dysgranular: Cerebral cortex with a moderately developed laminar organization involving a rudimentary layer IV and better developed layers II and III.

Hub: A group of the brain's most inter-connected neurons. The hubs with the most dense connections are referred to as

'rich club' hubs, and include visceromotor regions, as well as other heteromodal regions. They are thought to function as a high-capacity backbone for synchronizing neural activity, integrating information (and segregating noise) across the entire brain.

Internal Milieu: An integrated sensory representation of the physiological state of the body.

Laminar Organization: The architectural organization of neurons in a cortical column.

Naïve Realism: The belief that one's senses provide an accurate and objective representation of the world.

Pattern Generators: Groups of neurons (i.e. nuclei) that implement the sequences of actions for coordinated behaviors like feeding, running, and mating. An action is a single movement but a behavior is an event. Pattern generators are in the hypothalamus and down in the brainstem near their effector muscles and organs (Sterling and Laughlin, 2015; Swanson, 2005).

Visceromotor: Internal movements involving autonomic, neuroendocrine, and immune systems

Acknowledgements

We thank to Ajay Satpute, Amitai Shenhav, Suzanne Oosterwijk and Eliza Bliss-Moreau who read and/or made invaluable comments on an earlier draft of this article.

References

Adams, R.A., Shipp, S., Friston, K.J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, **218**(3), 611–43.

Adolphs, R. (2013). The biology of fear. *Current Biology*, **23**(2), R79–93.

Adolphs, R., Russell, J.A., Tranel, D. (1999). A role for the human amygdala in recognizing emotional arousal from unpleasant stimuli. *Psychological Science*, **10**(2),167–71.

Adolphs, R., Tranel, D. (1999). Intact recognition of emotional prosody following amygdala damage. *Neuropsychologia*, **37**(11),1285–92.

Adolphs, R., Tranel, D. (2003). Amygdala damage impairs emotion recognition from scenes only when they contain facial expressions. *Neuropsychologia*, **41**(10),1281–9.

Alle, H., Roth, A., Geiger, J.R. (2009). Energy-efficient action potentials in hippocampal mossy fibers. *Science*, **325**(5946),1405–8. doi:10.1126/science.1174331

Anderson, D.J., Adolphs, R. (2014). A framework for studying emotion across species. *Cell*, **157**, 187–200.

Anderson, M.L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge: MIT Press.

Anderson, M.L., Finlay, B.L. (2014). Allocating structure to function: the strong links between neuroplasticity and natural selection. *Frontiers in Human Neuroscience*, **7**, 918.

Antoniadis, E.A., Winslow, J.T., Davis, M., Amaral, D.G. (2009). The nonhuman primate amygdala is necessary for the acquisition but not the retention of fear-potentiated startle. *Biological Psychiatry*, **65**(3), 241–8.

Arnal, L.H., Giraud, A.L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, **16**(7), 390–8.

Atkinson, A.P., Heberlein, A.S., Adolphs, R. (2007). Spared ability to recognise fear from static and moving whole-body cues following bilateral amygdala damage. *Neuropsychologia*, **45**(12), 2772–82.

Attwell, D., Iadecola, C. (2002). The neural basis of functional brain imaging signals. *Trends in Neuroscience*, **25**(12), 621–5.

Attwell, D., Laughlin, S.B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, **21**(10), 1133–45.

Bar, K.J., de la Cruz, F., Schumann, A., *et al.* (2016). Functional connectivity and network analysis of midbrain and brainstem nuclei. *NeuroImage*, **134**, 53–63.

Bar, M. (2009a). A cognitive neuroscience hypothesis of mood and depression. *Trends in Cognitive Sciences*, **13**(11), 456–63.

Bar, M. (2009b). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1521), 1235–43.

Barbas, H. (2015). General cortical and special prefrontal connections: principles from structure to function. *Annual Review of Neuroscience*, **38**, 269–89.

Barbas, H., García-Cabezas, M. (2015). Motor cortex layer 4: less is more. *Trends in Neurosciences*, **38**(5),259–61.

Barbas, H., Rempel-Clower, N. (1997). Cortical structure predicts the pattern of corticocortical connections. *Cerebral Cortex*, **7**(7),635–46.

Bargmann, C.I. (2012). Beyond the connectome: how neuromodulators shape neural circuits. *Bioessays*, **34**(6),458–65. doi:10.1002/bies.201100185

Barrett, L.F. (2006a). Emotions as natural kinds?. *Perspectives on Psychological Science*, **1**, 28–58.

Barrett, L.F. (2006b). Solving the emotion paradox: categorization and the experience of emotion. *Personality and Social Psychology Review*, **10**(1), 20–46.

Barrett, L.F. (2009). The future of psychology: connecting mind to brain. *Perspectives on Psychological Science*, **4**(4), 326–39.

Barrett, L.F. (2011a). Bridging token identity theory and supervenience theory through psychological construction. *Psychological Inquiry*, **22**(2), 115–27.

Barrett, L.F. (2011b). Was Darwin wrong about emotional expressions?. *Current Directions in Psychological Science*, **20**, 400–6.

Barrett, L.F. (2012). Emotions are real. *Emotion*, **12**(3), 413–29.

Barrett, L.F. (2013). Psychological construction: A Darwinian approach to the science of emotion. *Emotion Review*, **5**, 379–89.

Barrett, L.F. (2014). The conceptual act theory: a précis. *Emotion Review*, **6**, 292–7.

Barrett, L.F. (2015). Ten common misconceptions about the psychological construction of emotion. In: Barrett, L.F., Russell, J.A., editors. *The psychological construction of emotion*. p. 45-79, New York: Guilford.

Barrett, L.F. (2017). *How Emotions Are Made: The Secret Life the Brain*. New York, NY: Houghton-Mifflin-Harcourt.

Barrett, L.F., Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, **41**, 167–218.

Barrett, L.F., Lindquist, K.A., Bliss-Moreau, E., *et al.* (2007a). Of mice and men: natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives on Psychological Science*, **2**(3), 297–311.

Barrett, L.F., Mesquita, B., Ochsner, K.N., Gross, J.J. (2007b). The experience of emotion. *Annual Review of Psychology*, **58**, 373–403.

Barrett, L.F., Russell, J.A. (1999). Structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, **8**(1), 10–4.

Barrett, L.F., Satpute, A.B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, **23**(3), 361–72.

Barrett, L.F., Simmons, W.K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, **16**, 419–29.

Barrett, L.F., Tugade, M.M., Engle, R.W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, **130**(4),553–73.

Barrett, L.F., Wilson-Mendenhall, C.D., Barsalou, L.W. (2015). The conceptual act theory: a road map. In: Barrett, L.F., Russell, J.A., editors. *The Psychological Construction of Emotion*. New York: Guilford.

Barsalou, L.W. (1983). Ad hoc categories. *Memory and Cognition*, **11**(3), 211–27.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Science*, **22**(4), 577–609. discussion 610-560.

Barsalou, L.W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, **18**, 513–62.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology*, **59**, 617–45.

Barsalou, L.W. (2016). On staying grounded and avoiding Quixotic dead ends. *Psychonomic Bulletin and Review*, **23**(4), 1122–42.

Barsalou, L.W., Kyle Simmons, W., Barbey, A.K., Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, **7**(2), 84–91.

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**(4), 695–711.

Bechara, A., Damasio, H., Damasio, A.R., Lee, G.P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, **19**(13), 5473–81.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., Damasio, A.R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, **269**(5227), 1115–8.

Becker, B., Mihov, Y., Scheele, D., *et al.* (2012). Fear processing and social networking in the absence of a functional amygdala. *Biological Psychiatry*, **72**(1), 70–7.

Beckmann, M., Johansen-Berg, H., Rushworth, M.F. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. *Journal of Neuroscience*, **29**(4), 1175–90.

Berkes, P., Orbán, G., Lengyel, M., Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, **331**(6013), 83–7.

Binder, J.R., Desai, R.H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, **15**(11), 527–36.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, **19**(12), 2767–96.

Boes, A.D., Mehta, S., Rudrauf, D., *et al.* (2011). Changes in cortical morphology resulting from long-term amygdala damage. *Social Cognitive and Affective Neuroscience*, **7**(5), 588–95.

Boiger, M., Mesquita, B. (2012). The construction of emotion in interactions, relationships, and cultures. *Emotion Review*, **4**

Bressler, S.L., Richter, C.G. (2015). Interareal oscillatory synchronization in top-down neocortical processing. *Current Opinion in Neurobiology*, **31**, 62–6.

Brodski, A., Paach, G.F., Helbling, S., Wibral, M. (2015). The faces of predictive coding. *The Journal of Neuroscience*, **35**, 8997–9006.

Buckner, R.L. (2012). The serendipitous discovery of the brain's default network. *NeuroImage*, **62**(2), 1137–45.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, **1124**, 1–38.

Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, **106**(5), 2322–45. doi:10.1152/jn.00339.2011

Buhle, J.T., Silvers, J.A., Wager, T.D., *et al.* (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cerebral Cortex*

Bullmore, E., Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, **13**(5), 336–49.

Burrows, M. (1996). *The Neurobiology of an Insect Brain*. Oxford; New York: Oxford University Press.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford; New York: Oxford University Press.

Cannon, W.B., Britton, S.W. (1925). Studies on the conditions of activity in endocrine glands. *American Journal of Physiology–Legacy Content*, **72**(2), 283–94.

Capitanio, J.P., Cole, S.W. (2015). Social instability and immunity in rhesus monkeys: the role of the sympathetic nervous system. *Philosophical Transactions of the Royal. Society B: Biological Scicnes*, **370**(1669), 20140104.

Carrive, P., Morgan, M.M. (2012). Periaquiductal gray. In: Mai, J.K., Paxinos, G., editors. *The Human Nervous System*, 3rd edn., pp. 367–400. Boston: Elsevier.

Chanes, L., Barrett, L.F. (2016). Redefining the Role of Limbic Areas in Cortical Processing. *Trends in Cognitive Sciences*, **20**(2), 96–106.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**, 281–53.

Clark, A. (2013b). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology*, **4**, 270.

Clark-Polner, E., Johnson, T., Barrett, L.F. (2016). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cerebral Cortex* doi: 10.1093/cercor/bhw028.

Clark-Polner, E., Wager, T.D., Satpute, A.B., Barrett, L.F. (2016). Neural fingerprinting: Meta-analysis, variation and the search for brain-based essences in the science of emotion. In: Barrett, L.F., Lewis, M., Haviland-Jones, J.M., editors. *The handbook of emotion*, 4th edn. p. 146–165, New York: Guilford.

Clore, G.L., Ortony, A. (2008). Appraisal theories: how cognition shapes affect into emotion. In: Lewis, M., Haviland-Jones, J.M., Barrett, L.F., editors. *Handbook of Emotions*, 3rd edn., pp. 628–42. New York: Guilford Press.

Conant, R.C., Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system†. *International Journal of Systems Science*, **1**(2), 89–97.

Counts, S.E., Mufson, E.J. (2012). The human nervous system. In: Mai, J.K. Paxinos, G., editors. *The Human Nervous System*, pp. 425–38. Boston, MA: Elsevier.

Craig, A.D. (2015). *How Do You Feel?: an Interoceptive Moment with Your Neurobiological Self*. Princeton: Princeton University Press.

Crivelli, C., Jarillo, S., Russell, J.A., Fernandez-Dols, J.M. (2016). Reading emotions from faces in two indigenous societies. *Journal of Experimental Psychology-General*, **145**(7), 830–43.

Crossley, N.A., Mechelli, A., Scott, J., *et al.* (2014). The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain*, **137**(8), 2382–95.

Damasio, A., Carvalho, G.B. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, **14**(2), 143–52.

Damasio, A.R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, **33**(1–2), 25–62.
Damasio, A.R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Houghton: Houghton Mifflin Harcourt.
Dantzer, R., Heijnen, C.J., Kavelaars, A., Laye, S., Capuron, L. (2014). The neuroimmune basis of fatigue. *Trends in Neurosciences*, **37**(1), 39–46.
Danziger, K. (1997). *Naming the Mind: How Psychology Found Its Language*. London, England: Sage.
Darwin, C. (1859/1964). *On the Origin of Species. A Facsimile of the First Edition*. Cambridge, MA: Harvard University Press.
Darwin, C. (1872/2005). *The Expression of the Emotions in Man and Animals Stilwell, KS: Digireads. com*.
Davachi, L., DuBrow, S. (2015). How the hippocampus preserves order: the role of prediction and context. *Trends in Cognitive Sciences*, **19**(2), 92–9.
Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience*, **15**, 353–75.
de Gelder, B., Terburg, D., Morgan, B., Hortensius, R., Stein, D. J., van Honk, J. (2014). The role of human basolateral amygdala in ambiuous social threat perception. *Cortex*, **52**, 28–34.
De Martino, B., Camerer, C.F., Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(8), 3788–92.
Deneve, S. (2008). Bayesian spiking neurons I: inference. *Neural Computation*, **20**, 91–117.
Deneve, S., Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences*, **11**, 40–8.
Dewey, J. (1896). The reflex arc concept in psychology. *The Psychological Review*, **3**, 357–70.
Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, **11**(4), 410–6.
Dreyfus, G., Thompson, E. (2007). Asian perspectives: Indian theories of mind. In: Zelazo, P.D., Moscovitch, M., Thompson, E., editors. *The Cambridge Handbook of Consciousness*, pp. 89–114. Cambridge: Cambridge University Press.
Dunsmoor, J.E., Murty, V.P., Davachi, L., Phelps, E.A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, **520**(7547), 345–8.
Edelman, G.M., Tononi, G. (2000). *A Universe of Consciousness: How Matter Becomes Imagination*. New York: Basic books.
Edelman, G.M., Gally, J.A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13763–8.
Einstein, A., Infeld, L. (1938). Evolution of physics. Cambridge, United Kingdom: Cambridge University Press.
Etkin, A., Buchel, C., Gross, J.J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience*, **16**(11), 693–+.
Feinstein, J.S. (2013). Lesion studies of human emotion and feeling. *Current Opinion in Neurobiology*, **23**(3), 304–9.
Feinstein, J.S., Adolphs, R., Damasio, A., Tranel, D. (2011). The human amygdala and the induction and experience of fear. *Current Biology*, **21**(1), 34–8.
Feinstein, J.S., Adolphs, R., Tranel, D. (2016). A tale of survival from the world of Patient S.M. In: Amaral, D.G. Adolphs, R., editors. *Living without an Amygdala*, pp. 1–38. New York: Guilford.
Feinstein, J.S., Buzza, C., Hurlemann, R., *et al.* (2013). Fear and panic in humans with bilateral amygdala damage. *Nature Neuroscience*, **16**(3), 270–2.
Feinstein, J.S., Rudrauf, D., Khalsa, S.S., *et al.* (2010). Bilateral limbic system destruction in man. *Journal of Clinical and Experimental Neuropsychology*, **32**(1), 88–106.
Feldman, H., Friston, K.J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, **4**, 215.
Fernandino, L., Binder, J.R., Desai, R.H., *et al.* (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, **26**, 2018–34.
Fernandino, L., Humphries, C.J., Seidenberg, M.S., Gross, W.L., Conant, L.L., Binder, J.R. (2015). Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia*, **76**, 17–26.
Fields, H. (2004). State-dependent opioid control of pain. *Nature Reviews Neuroscience*, **5**(7), 565–75.
Fields, H.L., Margolis, E.B. (2015). Understanding opioid reward. *Trends in Neurosciences*, **38**(4), 217–25.
Finlay, B.L., Uchiyama, R. (2015). Developmental mechanisms channeling cortical evolution. *Trends in Neurosciences*, **38**(2), 69–76.
Firestein, S. (2012). *Ignorance: How It Drives Science*. Oxford: Oxford University Press.
Freddolino, P.L., Tavazoie, S. (2012). Beyond homeostasis: a predictive-dynamic framework for understanding cellular behavior. *Annual Review of Cell and Developmental Biology*, **28**, 363–84.
Friston, K. (2010). The free-energy principle: A unified brain theory?. *Nature Reviews Neuroscience*, **11**, 127–38.
Furlong, T.M., Cole, S., Hamlin, A.S., McNally, G.P. (2010). The role of prefrontal cortex in predictive fear learning. *Behavioral Neuroscience* **124**(5), 574.
Gallivan, J.P., Logan, L., Wolpert, D.M., Flanagan, J.R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature Neuroscience*, **19**(2), 320–6.
Gelman, S.A., Rhodes, M. (2012). Two-thousand years of stasis. How psychological essentialism impedes evolutionary understanding. In: Rosengren, K.S., Brem, S., Evans, E.M. Sinatra, G., editors. *Evolution Challenges: Integrating Research and Practice in Teaching and Learning About Evolution*, pp. 3–21.Oxford: Oxford University Press.
Gendron, M., Roberson, D., van der Vyver, J.M., Barrett, L.F. (2014a). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science*, **25**(4), 911–20.
Gendron, M., Roberson, D., van der Vyver, J.M., Barrett, L.F. (2014b). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, **14**(2), 251–62.
Gendron, M., Roberson, D., Barrett, L.F. (2015). Cultural variaton in emotion perception is real: a response to Sauter *et al.* *Psychological Science*, **26**, 357–9.
Ghashghaei, H., Hilgetag, C., Barbas, H. (2007). Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala. *NeuroImage*, **34**(3), 905–23.
Gilbert, D.T. (1998). Ordinary personology. In: Fiske, S.T., Gardner, L., editors. *The Handbook of Social Psychology*, Vol. **1 and 2**, pp. 89–150. New York, NY: McGraw-Hill.
Goodkind, M., Eickhoff, S.B., Oathes, D.J., *et al.* (2015). Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry*, **72**(4), 305–15.
Gross, J.J., Barrett, L.F. (2011). Emotion generation and emotion regulation: one or two depends on your point of view. *Emotion Review*, **3**(1), 8–16.
Gross, C.T., Canteras, N.S. (2012). The many paths to fear. *Nature Reviews Neuroscience*, **13**(9), 651–8.

Gross, J.J. (2015). Emotion regulation: current status and future prospects. *Psychological Inquiry*, **26**(1), 1–26.
Guillory, S.A., Bujarski, K.A. (2014). Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. *Social Cognitive and Affective Neuroscience*, **9**(12), 1880–9.
Guitart-Masip, M., Duzel, E., Dolan, R., Dayan, P. (2014). Action vserus valence in decision making. *Trends in Cognitive Sciences*, **18**, 194–202.
Haber, S.N., Behrens, T.E. (2014). The neural network underlying incentive-based learning: implications for interpreting circuit disruptions in psychiatric disorders. *Neuron*, **83**(5), 1019–39.
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**(4), 545–55.
Harris, J.J., Jolivet, R., Attwell, D. (2012). Synaptic energy use and supply. *Neuron*, **75**(5), 762–77.
Hassabis, D., Maguire, E.A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1521), 1263–71.
Hasson, U., Chen, J., Honey, C.J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*, **19**(6), 304–13.
Herry, C., Johansen, J.P. (2014). Encoding of fear learning and memory in distributed neuronal circuits. *Nature Neuroscience*, **17**(12), 1644–54.
Hodgkin, A.L., Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**(4), 500–44.
Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP Oxford.
Hunter, R.G., McEwen, B.S. (2013). Stress and anxiety across the lifespan: structural plasticity and epigenetic regulation. *Epigenomics*, **5**(2),177–94.
Hurlemann, R., Wagner, M., Hawellek, B., *et al.* (2007). Amygdala control of emotion-induced forgetting and remembering: evidence from Urbach-Wiethe disease. *Neuropsychologia*, **45**(5),877–84.
Iwata, J., LeDoux, J.E. (1988). Dissociation of associative and nonassociative concomitants of classical fear conditioning in the freely behaving rat. *Behavioral Neuroscience*, **102**(1),66–76.
James, W. (1890/2007). *The Principles of Psychology*, Vol. **1**. New York: Dover.
John, Y.J., Zikopoulos, B., Bullock, D., Barbas, H. (2016). The emotional gatekeeper: A computational model of attention selection and suppression through the pathway from the amygdala to the inhibitory thalamic reticular nucleus. *PLOS Computational Biology*, **12**(2),e1004722.
Karmiloff-Smith, A. (2009). Nativism versus neuroconstructivism: rethinking the study of developmental disorders. *Developmental Psychology*, **45**(1),56–63.
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**(6), e66032.
Kleckner, I.R., Zhang, J., Touroutoglou, A., *et al.* (in press). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behavior*. doi: https://doi.org/10.1101/098970.
Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., Wager, T.D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage*, **42**(2), 998–1031.
Kok, P., Bains, L.J., van Mourik, T., Norris, D.G., de Lange, F.P. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, **26**(3), 371–6.
Kragel, P.A., LaBar, K.S. (2013). Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion*, **13**(4), 681–90.
Kragel, P.A., LaBar, K.S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience*, **10**(11), 1437–48.
Kragel, P.A., LaBar, K.S. (2016). Decoding the nature of emotion in the brain. *Trends in Cognitive Sciences*, **20**(6),444–55.
Kuppens, P., Tuerlinckx, F., Russell, J.A., Barrett, L.F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, **139**(4), 917–40.
Laxminarayan, S., Tadmor, G., Diamond, S.G., Miller, E., Franceschini, M.A., Brooks, D.H. (2011). Modeling habituation in rat EEG-evoked responses via a neural mass model with feedback. *Biological Cybernetics*, **105**(5–6), 371–97.
Lazarus, R.S. (1991). *Emotion and Adaptation*. New York, NY: Oxford University Press.
LeDoux, J.E. (2012). Rethinking the emotional brain. *Neuron*, **73**(4),653–76.
Li, S.S.Y., McNally, G.P. (2014). The conditions that promote fear learning: prediction error and Pavlovian fear conditioning. *Neurobiology of Learning and Memory*, **108**, 14–21.
Liang, M., Mouraux, A., Hu, L., Iannetti, G. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature Communications*, **4**
Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, **35**(3),121–43.
Lochmann, T., Deneve, S. (2011). Neural processing as causal inference. *Current Opinion in Neurobiology*, **21**(5),774–81.
Makeig, S., Debener, S., Onton, J., Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, **8**(5),204–10.
Makeig, S., Westerfield, M., Jung, T.P., *et al.* (2002). Dynamic brain sources of visual evoked responses. *Science*, **295**(5555), 690–4.
Marder, E., Taylor, A.L. (2011). Multiple models to capture the variability in biological neurons and networks. *Nature Neuroscience*, **14**, 133–8.
Mareschal, D., Johnson, M.H., Sirois, S., Spratling, M., Thomas, M.S., Westermann, G. (2007). *Neuroconstructivism-I: How the Brain Constructs Cognition*. Oxford: Oxford University Press.
Mason, W.A., Capitanio, J.P., Machado, C.J., Mendoza, S.P., Amaral, D.G. (2006). Amygdalectomy and responsiveness to novelty in rhesus monkeys (Macaca mulatta): generality and individual consistency of effects. *Emotion*, **6**(1), 73.
Mayr, E. (2004). *What Makes Biology Unique?: Considerations on the Autonomy of a Scientific Discipline*. Cambridge: Cambridge University Press.
Mazaheri, A., Jensen, O. (2010). Rhythmic pulsing: linking ongoing brain activity with evoked responses. *Frontiers in Human Neuroscience*, **4**, 177.
McEwen, B.S., Bowles, N.P., Gray, J.D., *et al.* (2015). Mechanisms of stress in the brain. *Nature Neuroscience*, **18**(10), 1353–63.
McGaugh, J.L. (2016). Consolidating memories. *Annual Review of Psychology*, **66**, 1–24.
McIntosh, A.R. (2004). Contexts and catalysts: a resolution of the localization and integration of function in the brain. *Neuroinformatics*, **2**(2), 175–82.
McNally, G.P., Johansen, J.P., Blair, H.T. (2011). Placing prediction into the fear circuit. *Trends in Neurosciences*, **34**(6), 283–92.
McNorgan, C. (2012). A meta-analytic review of multisensory imagery identifies the neural correlates of modality-specific and modality-general imagery. *Frontiers in Human Neuroscience*, **6**, Article 285.

Mesulam, M.M. (1998). From sensation to cognition. *Brain*, **121(Pt 6)**, 1013–52.
Mesulam, M.M. (2002). The human frontal lobes: Transcending the default mode through contingent encoding. In: Stuss, D.T., Knight, R.T., editors. *Principles of Frontal Lobe Function*, pp. 8–30. New York, NY: Oxford University Press.
Meyer, K., Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Cognitive Sciences*, **32**, 376–82.
Mihov, Y., Kendrick, K.M., Becker, B., *et al.* (2013). Mirroring fear in the absence of a functional amygdala. *Biological Psychiatry*, **73**(7),e9–11.
Moran, R.J., Campo, P., Symmonds, M., Stephan, K.E., Dolan, R.J., Friston, K.J. (2013). Free energy, precision and learning: the role of cholinergic neuromodulation. *The Journal of Neuroscience*, **33**(19),8227–36.
Murphy, G. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT press.
Ongur, D., Ferry, A.T., Price, J.L. (2003). Architectonic subdivision of the human orbital and medial prefrontal cortex. *Journal of Comparative Neurology*, **460**(3), 425–49.
Oosterwijk, S., Lindquist, K.A., Adebayo, M., Barrett, L.F. (2015). The neural representation of typical and atypical experiences of negative images: comparing fear, disgust and morbid fascination. *Social Cognitive and Affective Neuroscience*, **11**, 11–22.
Oosterwijk, S., Lindquist, K.A., Anderson, E., Dautoff, R., Moriguchi, Y., Barrett, L.F. (2012). States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, **62**(3), 2110–28.
Oosterwijk, S., Mackey, S., Wilson-Mendenhall, C., Winkielman, P., Paulus, M.P. (2015). Concepts in context: processing mental state concepts with internal or external focus involves different neural systems. *Social Neuroscience*, **10**(3), 294–307.
Pavlenko, A. (2014). *The Bilingual Mind: And What It Tells Us about Language and Thought*. Cambridge: Cambridge University Press.
Peelen, M.V., Atkinson, A.P., Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, **30**(30), 10127–34.
Pezzulo, G., Rigoli, F., Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, **134**, 17–35.
Posner, M.I., Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**(3p1), 353–63.
Power, J.D., Cohen, A.L., Nelson, S.M., *et al.* (2011). Functional network organization of the human brain. *Neuron*, **72**(4),665–78.
Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, **17**(9),458–70.
Qian, C., Di, X. (2011). Phase or amplitude? The relationship between ongoing and evoked neural activity. *Journal of Neuroscience*, **31**(29),10425–6.
Raichle, M.E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, **14**(4),180–90.
Rao, R.P.N., Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, **2**, 79–87.
Raz, G., Touroutoglou, A., Gilam, G., *et al.* (2016). Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. *Cognitive, Affective, and Behavioral Neuroscience*, **16**, 709–23.
Rigotti, M., Barak, O., Warden, M.R., *et al.* (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, **497**(7451),585–90.
Robbins, J., Rumsey, A. (2008). Introduction: Cultural and linguistic anthropology and the opacity of other minds. *Anthropological Quarterly*, **81**(2),407–20.
Roseman, I.J. (2011). Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review*, **3**, 1–10.
Russell, J.A. (1991). Culture and the Categorization of Emotions. *Psychological Bulletin*, **110**(3),426–50.
Saarimaki, H., Gotsopoulos, A., Jaaskelainen, I.P., *et al.* (2016). Discrete Neural Signatures of Basic Emotions. *Cerebral Cortex*, **26**(6),2563–73.
Salamone, J.D., Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron*, **76**, 470–85.
Sartorius, N., Kaelber, C.T., Cooper, J.E., *et al.* (1993). Progress toward achieving a common language in psychiatry: results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in ICD-10. *Archives of General Psychiatry*, **50**(2),115–24.
Satpute, A.B., Wilson-Mendenhall, C.D., Kleckner, I.R., Barrett, L.F. (2015). Emotional experience. In: Toga, A.W., editor. *Brain Mapping: An Encyclopedic Reference*, pp. 65–72. Boston MA: Elsevier.
Sayers, B.M., Beagley, H.A., Henshall, W.R. (1974). The mechanism of auditory evoked EEG responses. *Nature*, **247**, 481–3.
Schacter, D.L., Addis, D.R., Buckner, R.L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience* **8**(9), 657–61.
Scheeringa, R., Mazaheri, A., Bojak, I., Norris, D.G., Kleinschmidt, A. (2011). Modulation of visually evoked cortical FMRI responses by phase of ongoing occipital alpha oscillations. *Journal of Neuroscience*, **31**(10), 3813–20.
Scherer, K.R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364** (1535), 3459–74.
Schmahmann, J.D. (2010). The role of the cerebellum in cognition and emotion: personal reflections since 1982 on the dysmetria of thought hypothesis, and its historical evolution from theory to therapy. *Neuropsychology Review*, **20**(3),236–60.
Schmahmann, J.D., Pandya, D.N. (1997). The cerebrocerebellar system. *International Review of Neurobiology*, **41**, 31–60.
Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, **17**(3), 183–95.
Searle, J.R. (1992). *The Rediscovery of the Mind*. Cambridge: MIT press.
Searle, J.R. (2004). *Mind: A Brief Introduction*. Oxford: Oxford University Press.
Sepulcre, J., Sabuncu, M.R., Yeo, T.B., Liu, H., Johnson, K.A. (2012). Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain. *Journal of Neuroscience*, **32**(31),10649–61.
Seth, A.K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitiive Sciences*, **17**(11), 565–73.
Seth, A.K., Suzuki, K., Critchley, H.D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, **2**, 395.
Seth, A. K., Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B*, doi: 10.1098/rstb.2016.0007.

Sharpe, M.J., Killcross, S. (2015). The prelimbic cortex directs attention toward predictive cues during fear learning. *Learning and Memory*, **22**(6), 289–93.
Shipp, S., Adams, R.A., Friston, K.J. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences*, **36**(12), 706–16.
Si, M., Marsella, S., Pynadath, D. (2010). Modeling appraisal in Theory of Mind Reasoning. *Journal of Autonomous Agents and Multi-Agent Systems*, **20**, 14–31.
Skerry, A.E., Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology*, **25**(15), 1945–54.
Sloan, E.K., Capitanio, J.P., Cole, S.W. (2008). Stress-induced remodeling of lymphoid innervation. *Brain, Behavior, and Immunity*, **22**(1), 15–21.
Sloan, E.K., Capitanio, J.P., Tarara, R.P., Mendoza, S.P., Mason, W.A., Cole, S.W. (2007). Social stress enhances sympathetic innervation of primate lymph nodes: mechanisms and implications for viral pathogenesis. *The Journal of Neuroscience*, **27**(33), 8857–65.
Spillmann, L., Dresp-Langley, B., Tseng, C.H. (2015). Beyond the classical receptive field: The effect of contextual stimuli. *Journal Vision*, **15**(9), 7.
Sporns, O. (2011). *Networks of the Brain*. Cambridge: MIT press.
Stephens, C.L., Christie, I.C., Friedman, B.H. (2010). Autonomic specificity of basic emotions: evidence from pattern classification and cluster analysis. *Biological Psychology*, **84**(3), 463–73.
Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology and Behavior*, **106**(1), 5–15.
Sterling, P., Laughlin, S. (2015). *Principles of Neural Design*. Cambridge: MIT Press.
Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, **78**(2), 364–75.
Strick, P.L., Dum, R.P., Fiez, J.A. (2009). Cerebellum and Nonmotor Function. *Annual Review of Neuroscience*, **32**, 413–34.
Swanson, L.W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research*, **886**(1–2), 113–64.
Swanson, L.W. (2012). *Brain Architecture: Understanding the Basic Plan*. Oxford: Oxford University Press.
Terburg, D., Morgan, B., Montoya, E., *et al.* (2012). Hypervigilance for fear after basolateral amygdala damage in humans. *Translational Psychiatry*, **2**(5), e115.
Tononi, G., Sporns, O., Edelman, G.M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(6), 3257–62.
Touroutoglou, A., Hollenbeck, M., Dickerson, B.C., Barrett, L.F. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage*.
Touroutoglou, A., Lindquist, K.A., Dickerson, B.C., Barrett, L.F. (2015). Intrinsic connectivity in the human brain does not reveal networks for 'basic' emotions. *Social Cognitive and Affective Neuroscience*, **10**(9), 1257–65.
Tovote, P., Fadok, J.P., Lüthi, A. (2015). Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience*, **16**(6), 317–31.
Tracy, J.L., Randles, D. (2011). Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, **3**(4), 397–405.
Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. *Nature Neuroscience*, **12**(10), 1224–5.
Turkheimer, E., Pettersson, E., Horn, E.E. (2014). A phenotypic null hypothesis for the genetics of personality. *Annual Review of Psychology*, **65**, 515–40.
Uddin, L.Q. (2015). Salience procesing and insular cortical function and dysfunction. *Neuron*, **16**, 55–61.
Ullsperger, M., Danielmeier, C., Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, **94**, 35–79.
van den Heuvel, M.P., Kahn, R.S., Goni, J., Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(28), 11372–7.
van den Heuvel, M.P., Sporns, O. (2011). Rich-club organization of the human connectome. *The Journal of Neuroscience*, **31**(44), 15775–86.
van den Heuvel, M.P., Sporns, O. (2013). An anatomical substrate for integration among functional networks in human cortex. *The Journal of Neuroscience*, **33**(36), 14489–500.
van den Heuvel, M.P., Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, **17**, 683–96.
Voorspoels, W., Vanpaemel, W., Storms, G. (2011). A formal ideal-based account of typicality. *Psychonomic Bulletin and Review*, **18**(5), 1006–14.
Vytal, K., Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, **22**(12), 2864–85.
Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., Barrett, L.F. (2015). A Bayesian model of category-specific emotional brain responses. *PLOS Computational Biology*, **11**(4), e1004066.
Westermann, G., Mareschal, D., Johnson, M.H., Sirois, S., Spratling, M.W., Thomas, M.S.C. (2007). Neuroconstructivism. *Developmental Science*, **10**(1), 75–83.
Whalen, P.J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science*, **7**(6), 177–88.
Whitacre, J., Bender, A. (2010). Degeneracy: a design principle for achieving robustness and evolvability. *Journal of Theoretical Biology*, **263**(1), 143–53.
Whitacre, J.M. (2010). Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theoretical Biology and Medical Modelling*, **7**(6), 1–17.
Wierzbicka, A. (2014). *Imprisoned in English. The Hazards of English as a Default Language*. Oxford: Oxford University Press.
Wilson, C.R., Gaffan, D., Browning, P.G., Baxter, M.G. (2010). Functional localization within the prefrontal cortex: missing the forest for the trees?. *Trends in Neurosciences*, **33**(12), 533–40.
Wilson-Mendenhall, C., Barrett, L.F., Barsalou, L.W. (2013). Neural evidence that human emotions share core affective properties. *Psychological Science*, **24**, 947–56.
Wilson-Mendenhall, C.D., Barrett, L.F., Barsalou, L.W. (2015). Variety in emotional life: within-category typicality of emotional experiences is associated with neural activity in large-scale brain networks. *Social Cognitive and Affective Neuroscience*, **10**(1),62–71. doi:10.1093/scan/nsu037
Wilson-Mendenhall, C.D., Barrett, L.F., Simmons, W.K., Barsalou, L.W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, **49**, 1105–27.
Woodworth, R.S., Sherrington, C.S. (1904). A pseudaffective reflex and its spinal path. *Journal of Physiology*, **31**(3–4), 234–43.

Wundt, W. (1897). *Outlines of Psychology* In: Judd, C.H. (Trans.). Leipzig: Wilhelm Engelmann.

Xu, F., Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, **22**(1), 28–32.

Zikopoulos, B., Barbas, H. (2006). Prefrontal projections to the thalamic reticular nucleus form a unique circuit for attentional mechanisms. *The Journal of Neuroscience*, **26**(28), 7348–61.

Zikopoulos, B., Barbas, H. (2012). Pathways for emotions and attention converge on the thalamic reticular nucleus in primates. *Journal of Neuroscience*, **32**(15), 5338–50.