The theory of constructed emotion:

An active inference account of interoception and categorization

Lisa Feldman Barrett

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

Corresponding author: 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 paper, 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.
Ancient philosophers and physicians believed that a human mind was 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 have morphed over the millennia, but generally speaking, they have become 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. 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, Infeld, & Hoffmann, 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 categories). So in this paper, 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, 2006, 2011, 2012; Barrett, 2013; Barrett, 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 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 (Bargmann, 2012; Doya, 2008). 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 & Jardri, 2016). However, a given neuron (or group of neurons) represents different features from moment to moment (e.g., (Spillmann, Dresp-Langley, & Tseng, 2015; Stokes et al., 2013)) because many different neurons synapse onto one (many-to-one connectivity), and a neuron’s receptive field depends on the information it receives at a given moment in time (i.e., depending 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 & Laughlin, 2015)) to help implement instances of different psychological categories; i.e., neurons are multipurpose (for evidence and discussion, see (Anderson, 2014; Anderson & Finlay, 2014; Barrett & Satpute, 2013)), even in subcortical regions like the amygdala (Cerf, personal communication, 7/30/15).

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

The brain achieves complexity through degeneracy (Edelman, J. A., 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 different contexts (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 (for examples, see ( Edelman, J. A., 2001; Marder, 2011; Tononi, Sporns, & Edelman, 1999). Natural selection favors systems with degeneracy because they are high in complexity and robust to damage (J. Whitacre & Bender, 2010). Degeneracy explains why Roger, the patient
who lost his limbic circuitry to herpes simplex type I encephalitis, still experiences emotions (Justin S Feinstein et al., 2010) and why monozygotic twins with fully calcified basolateral sectors of the amygdala (due to Urbach-Wiethe disease) 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, Pettersson, & Horn, 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, in press #9727). This observation is an example of population thinking, pioneered by Darwin’s On the Origin of Species (Mayr, 2007). 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, 2007)). 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, Wager, Satpute, & Barrett, in press) e.g., ( Wilson-Mendenhall, Barrett, & Barsalou, 2015; Wilson-Mendenhall, Barrett, Simmons, & Barsalou, 2011)). 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, 2007); as applied to emotion concepts and categories, see (Barrett, 2017), and as applied to concepts and categories more generally, see ( Barsalou, 1983; Voorspoels, Vanpaemel, & Storms, 2011)). The fact that human brains effortlessly and automatically construct such representations (e.g., ( Murphy, 2004; Posner & Keele, 1968)) helps to explain why scientists continue to believe in the classical view and even propose it as an innovation (e.g., (D. J. Anderson, & Adolphs, R., 2014)), even as evidence continues to call it into doubt (e.g., (Barrett, 2006; Barrett, 2011; Barrett, 2012; Barrett et al., 2007); 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 & 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 \textit{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 make 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; \textit{efficiency} requires the ability to \textit{anticipate} the body’s needs and satisfy them before they arise (Sterling, 2012; Sterling & Laughlin, 2015). An animal thrives when it has sufficient resources to explore the world, and to consolidate the details of 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, Heijnen, Kavelaars, Laye, & Capuron, 2014)) is suboptimal. Prolonged imbalances can lead to illness (e.g., (Hunter & McEwen, 2013; McEwen et al., 2015) that remolds the brain (Crossley et al., 2014; Goodkind et al., 2015) and the sympathetic nervous system, with corresponding behavior changes ((Capitanio & Cole, 2015; Sloan et al., 2007); for a review, see (Sloan, Capitanio, & Cole, 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 & Laughlin, 2015), p. 91). This regulation helps explain why, in mammals, the regions that are responsible for performing allostasis (the amygdala, ventral striatum, insula, orbitofrontal cortex, anterior cingulate cortex, medial prefrontal cortex, collectively called \textit{visceromotor regions}) are usually assumed to contain the circuits for emotion. In fact, many of these visceromotor regions happen to be 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, and with the systems that control skeletomotor movements. Therefore these regions are clearly multipurpose when it comes to 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. In psychology, we refer to this modeling as *embodied simulations* (Barsalou, 2008; Lawrence W Barsalou, Simmons, Barbey, & Wilson, 2003) (for example, see Figure 3). An internal model is metabolic investment, implemented by intrinsic activity (e.g., (Berkes, Orbán, Lengyel, & Fiser, 2011) that, in a human, occupies 20% of the total energy consumed (Raichle, 2010). 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. 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, 2017; Barrett & Simmons, 2015b; Chanes & Barrett, 2016); for related discussions, see (Pezzulo, Rigoli, & Friston, 2015; A. K. Seth, 2013; Anil K. Seth, Suzuki, & Critchley, 2012). Interoceptive sensations are usually experienced as lower dimensional feelings of affect (Barrett, 2017; Barrett & Bliss-Moreau, 2009). As such, the properties of affect—valence and arousal (Barrett & Russell, 1999; Kuppens, Tuerlinckx, Russell, & Barrett, 2013)—are basic features of consciousness (Antonio R Damasio, 1999; Dreyfus & Thompson, 2007; G. M. Edelman & 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, for example, learn, remember, make predictions, and forage in service to allostasis (Freddolino & Tavazoie, 2012; Sterling & 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 ecological 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 center on 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. Ample evidence shows that ongoing brain activity influences how the brain processes incoming sensory information (e.g., (Sayers, Beagley, & Henshall, 1974)), and that neurons fire intrinsically within large networks without any need for external stimuli (Larry W 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 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. This hypothesis is variously called predictive coding, active inference, or belief propagation (e.g., (Barrett & Simmons, 2015b; Chanes & Barrett, 2016; Andy Clark, 2013; Deneve & Jardri, 2016; Friston, 2010; Hohwy, 2013; Rao & Ballard, 1999; A. K. Seth, 2013; A. K. Seth, Suzuki, & Critchley, 2011)). 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. They are embodied, whole brain representations that anticipate (1) upcoming sensory events both inside the body and out as well as (2) the best action to deal with the impending sensory events. Their consequence for allostasis is made available to 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?” (M. Bar, 2009a, 2009b). Similarity is based on 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 (Bastos et al., 2012; Deneve, 2008), a brain uses pattern completion to decide among simulations and implement one of them (Gallivan, Logan, Wolpert, & Flanagan, 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) is 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 (Hohwy, 2013; Lochmann & Deneve, 2011). 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 visceromotor predictions (to control your autonomic nervous system, your neuroendocrine system, and your immune system) and voluntary motor predictions, anticipating 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. 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 depicted in Figure 4, 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; Roseman, 2011; Scherer, 2009); for a discussion, see (Barrett, Mesquita, Ochsner, & Gross, 2007; James J. Gross & Barrett, 2011). Appraisals as
descriptions of the world (e.g., (Clore & Ortony, 2008), however, are produced by categorization with concepts (e.g., (Si, 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 hypothesize that in assembling populations of predictions, each one having some probability (computed with Bayesian priors) of being the best fit to the current circumstances, the brain is constructing concepts (Barrett, 2017) (or what Barsalou and Lupyan refer to as an “ad hoc” concept; (L. W. Barsalou, 1983; Lawrence W. Barsalou, 2003; L. W. Barsalou, Kyle Simmons, Barbey, & Wilson, 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, 2006; Barrett, 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 (1) actionable in a situated way and therefore (2) 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, 2015; Barbas & Rempel-Clower, 1997) 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, 2017; Barrett & Simmons, 2015a; Chanes & Barrett, 2016)). The first key observation is that prediction signals are carried via “feedback” connections that originate in cortical regions
that have 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, allostatically control physiology by relaying descending prediction signals to the internal milieu via a system of subcortical regions (K.-J. Bar, de la Cruz, F., Schumann, A., Koehler, S., Sauer, H., Critchley, H., & Wagner, G., 2016; Kleckner, under review), including the central nucleus of the amygdala (Ghashghaei, Hilgetag, & Barbas, 2007), the ventral and dorsal striatum, and the central pattern generators (Larry W 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., the mid cingulate cortex, medial prefrontal cortex, ventrolateral prefrontal cortex, premotor cortex, etc., see Figure 5A and B). My hypothesis is that the visceromotor regions of the brain that are responsible for performing 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 & DuBrow, 2015; Hasson, Chen, & Honey, 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 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 (Adams, Shipp, & Friston, 2013; Barbas, 2015; Barrett & Simmons, 2015; Bastos et al., 2012; Chanes & Barrett, 2016). Following the evidence for how the cytoarchitectural gradients in the cortical sheet predict information flow across cortical regions (Barbas, 1997 #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 motor cortex, primary interoceptive cortex, and the primary visual, auditory, and somatosensory regions.16

Because motor cortex has a laminar organization that is less well developed than visual, auditory, somatosensory and interoceptive primary sensory regions (Barbas & García-Cabezas, 2015), we hypothesize that motor cortex sends efferent copies to those sensory regions as
sensory predictions (see Figure 5D, red line). Furthermore, because of their differential laminar development, we hypothesize that primary interoceptive cortex in mid-to-posterior dorsal insula, forward 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 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 & 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, Bains, van Mourik, Norris, & de Lange, 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, such as 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 eventually make sense of the upcoming sensory 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. 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 moment (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 signals arise from physiologic changes within the internal milieu and ascend via vagal afferents 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 (Craig, 2015; A. Damasio & Carvalho, 2013). Notice that, in the context of this framework, perception (i.e., the “meaning” of sensory inputs) is constructed in the service of 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, Johansen-Berg, & Rushworth, 2009; Buckner, Krienen, Castellanos, Diaz, & Yeo, 2011; Haber & Behrens, 2014; Kleckner, under review)). I hypothesize that information from the amygdala to the cortex are not “emotional” per se, but signal uncertainty (Whalen, 1998) about the predicted sensory input (via the basolateral complex) and help 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, Barrett, & Barsalou, 2013)) can be considered as learning signal (Li & 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 its in the event of disruption), rather than playing a necessary or sufficient role in rewards themselves (Guitart-Masip, Duzel, Dolan, & Dayan, 2014; Salamone & Correa, 2012)). Other neuromodulators, such as opioids, seem to be more intrinsic to reward in that regard (e.g., (H. L. Fields & Margolis, 2015)). 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 cingulate cortex, hypothalamus, and the amygdala (Buckner et al., 2011; Schmahmann, 2010; Schmahmann & Pandya, 1997; Strick, Dum, & Fiez, 2009).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, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Wager, 2015)).

A brain implements an internal model of the world with concepts because it is metabolically efficient. Even before birth, a brain begins to build its internal model by processing prediction error from the body and the world (see Barrett, 2017 #24551). Prediction errors (i.e., the unanticipated sensory inputs) cascade in a feedforward cortical sweep, originating in the upper layers of cortices with more developed lamination 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 & Uchiyama, 2015). This dimension reduction efficiently represents 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 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 (such that 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.

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 7 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 have refined these hypotheses (presented in Figure 6). I hypothesize, as others do (Buckner, 2012; Hassabis & Maguire, 2009; Mesulam, 2002), that the default mode network is necessary for the brain’s internal model. Regardless of the other mental categories mapped to this activity, the simulations initiated by the default mode network eventually create concepts that eventually categorize sensory inputs and guide movements in the service of allostasis. This hypothesis is consistent with the hypothesis that the default mode network represents semantic concepts (Binder & Desai, 2011; J. R. Binder, Desai, Graves, & Conant, 2009) (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 (Barsalou, 2016; Fernandino et al., 2016; Fernandino et al., 2015; Pulvermüller, 2013)).

I further hypothesize that the salience network tunes the internal model by predicting which prediction errors to pay attention to (i.e., that are likely to be allostatically relevant and therefore worth the cost of encoding and consolidation; called precision signals (A. Clark, 2013; Feldman & Friston, 2010; Moran et al., 2013; Shipp et al., 2013)). 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 & Barrett, 2016)). They directly alter the gain on neurons that compute prediction error from incoming sensory input (i.e., 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., 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 & 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 (John, Zikopoulos, Bullock, & Barbas,
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, Hollenbeck, Dickerson, & Barrett, 2012; Uddin, 2015; Ullsperger, Danielmeier, & Jocham, 2014).

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., it is integrating information across the few moments that constitute an event, but also across longer time frames at various scales; for similar ideas, see (Hasson et al., 2015; Wilson, Gaffan, Browning, & Baxter, 2010)). 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.

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. A brain can be thought of as running an internal model that controls the central pattern generators (for more on pattern generators, see (Burrows, 1996; Sterling & 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 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 (which 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 categorization (a situated conceptualization; (Barrett, 2006; Barrett, Wilson-Mendenhall, & Barsalou, 2015; Barsalou, 1999; Lawrence W Barsalou et al., 2003)) 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 result is an instance of emotion.

This species-general 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 concepts are species-general, e.g., allostasis, interoception, affect, concept, while others that require the capacity to construct certain types of concepts (Barrett, 2012) are more likely to be species-specific (e.g., emotion concepts). It is necessary to understand which concepts 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 (D. J. Anderson, & Adolphs, R., 2014; Kragel & Labar, 2016)). With these observations in mind, here is a partial list of claims I am *not* making, to avoid further confusion:

- *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).

- *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.

- *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).

- *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 behaviors are functional in the context, given the state of the animal.

- *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.*

- *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.

- *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, Tugade, & Engle, 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.

• *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”).

**Selected Implications of the Theory**

Scientific revolutions in any field are difficult. At the beginning, new paradigms raise more questions than they answer. They may explain existing anomalies or resolve lingering questions within the old paradigm, but they also introduce a new set of questions that can be answered only with new experimental paradigms and computational techniques. This is a feature, not a bug, because it allows 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 cannot be found in answering the questions of the old approach. 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).

Ironically, perhaps the strongest evidence to date for the theory comes from studies that use pattern classification to distinguish categories of emotion. Several recent papers 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, Markey, Cherkassky, Loewenstein, & Just, 2013; Kragel & 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 (Philip A. Kragel & LaBar, 2013; Stephens, Christie, & Friedman, 2010).

Generally speaking, pattern classification results in the field 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. Indeed, the voxels that make up a “pattern” for a category are not observed in every (or even any) single instance of that category. A classic study by Posner & 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 & Barrett, in press).

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 (Woodworth & 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 & Britton’s "sham rage" (Cannon & Britton, 1925) where a decorticated cat, upon waking from anesthesia, spontaneously spit, clawed, and arched its back. Importantly, Cannon referred to these actions as "fury."

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, which I call the mental inference fallacy (Barrett, 2017), has wreaked havoc with the scientific accumulation of knowledge about emotion (also see (Barrett, 2012; LeDoux, 2012)). Behavior does 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 & Rumsey, 2008).

When viewed in this light, one cannot claim that studies of the human brain with functional magnetic resonance imaging (fMRI) and studies of non-human animal brains with
lesions or optogenetics yield different findings (R. Adolphs, 2013). In reality, both are making physical measurements and mapping emotion constructs to them, and neither set of findings is described appropriately by classical emotion constructs. 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., (C. T. Gross & Canteras, 2012)). 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 even for comparing humans of different cultures (Crivelli, Jarillo, Russell, & Fernandez-Dols, 2016; M. Gendron, Roberson, van der Vyver, & Barrett, 2014a, 2014b; M. Gendron, Roberson, D., & Barrett, , 2015; Russell, 1991; Wierzbicka, 2014).

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. We must also model how a brain makes meaning of those physical changes.

This paper is a mere sketch of a much larger scientific landscape. The theory also proposes that emotions should be modeled holistically, as whole brain phenomena. 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 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, Kahn, Goni, & Sporns, 2012; van den Heuvel & Sporns, 2011; van den Heuvel & Olaf Sporns, 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—for example, that reappraisal and other regulation processes (Etkin, Buchel, & Gross, 2015; James J Gross, 2015) are accomplished with predictions that categorize sensory inputs and control action with concepts (see Figure 4).

The theory of constructed emotion also views the distinction between the central and peripheral nervous systems as a historical distinction, rather than one that is scientifically important. For example, ascending interoceptive signals bringing sensory prediction errors from the internal milieu to the brain via lamina I and vagal afferent pathways that are anatomically positioned to modulate descending visceromotor predictions that control the internal milieu (e.g., (H. Fields, 2004)). This suggests the hypothesis that concepts (i.e., prediction signals) act like a volume dial to influence prediction errors before they reach the brain. This provides new hypotheses about the chronification of pain (see Barrett, 2017) that treat 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. It provides a common framework for understanding mental, physical, and neurodegenerative disorders, and collapses the artificial boundaries between cognitive, affective, and social neurosciences. Ultimately, it equips the science of emotion to solve the age-old mysteries of how a human brain creates a human mind.
Endnotes

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. E.g., 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.

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, 7/30/15). 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, Mouraux, Hu, & Iannetti, 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; (Gelman & Rhodes, 2012; Mayr, 2007). 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, 2007). 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 means 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 means 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 when the body is in need of glucose (even before anything is ingested). Even just imaging food causes glucose secretion).

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 & Ross Ashby, 1970). From the brain’s perspective, the “system” in question as a body that constructs an ecological niche; it must be watered, fed, and cared for, so that it can grow, thrive, and ultimately, reproduce and
care for its young so as to pass its genes to the next generation. So, the brain runs an internal
model of the world it inhabits so as to act on that world (i.e., to grow, survive and reproduce).

6 Long-range neural connections, like those that form the human brain’s broadly
distributed intrinsic networks, are particularly expensive (Bullmore & Sporns, 2012; Sterling &
Laughlin, 2015), with most of the energy costs going to signaling between neurons, particularly
in postsynaptic processes (Alle, Roth, & Geiger, 2009; Attwell & Iadecola, 2002; Attwell &
Laughlin, 2001; Harris, Jolivet, & Attwell, 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. We humans have been able
to expand our ecological niche (and therefore our internal models)

8 This mistaken belief is an artifact of studying neurons in isolation (e.g. (Hodgkin &
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 (Laxminarayan et al., 2011; Makeig, Debener, Onton, & Delorme, 2004;
Makeig et al., 2002; Mazaheri & Jensen, 2010; Qian & Di, 2011; Scheeringa, Mazaheri, Bojak,
Norris, & Kleinschmidt, 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 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 (1) the hypothesis that
predictions are embodied simulations of sensory motor experiences, (2) they are ultimately in the
service of allostasis and therefore interoception is at their core, and, of course (3) the breadth of
behavioral, functional, and anatomic evidence supporting the hypothesis that the brain’s internal
model implements active inference as prediction signals, including (4) the specific computational
hypotheses implementing a predictive coding account.

11 Notably, Buzsáki wrote that “Brains are foretelling devices” (Buzsáki, 2006) and there
is accumulating evidence that prediction and prediction error signals oscillate at different
frequencies within the brain (e.g., (Arnal & Giraud, 2012; Bressler & Richter, 2015; Brodski,
Paach, Helbling, & Wibral, 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 (R. L. Buckner,

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 are “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 an animal’s past to anticipate future events that must be dealt with (Sterling & Laughlin, 2015).

Traditionally, a category is a collection of instances or events that are treated as similar for some purpose (G. Murphy, 2004). A concept is the internal representations of those instances or events. 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 are physical features (e.g., currency throughout the course of human history). These conceptual categories have also been called abstract or nominal categories. Biological categories can be conceptual categories. According to the evolutionary biologist Ernst Mayr (2007), this was one of Darwin’s conceptual innovations in On the Origin of Species. The categories of social reality, such as a flower and a weed, or emotion categories, are also conceptual because functions are imposed on physically disparate instances by virtual of collective agreement (Barrett, 2012).

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 & García-Cabezas, 2015) has an even less well-developed laminar structure and therefore sends predictions to somatosensory cortex (Adams et al., 2013; Shipp, Adams, & Friston, 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.

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 & Mufson, 2012)). The locus coeruleus also receives ascending interoceptive and nociceptive prediction errors (see Counts & Mufson, 2012). This is yet another way that allostatics is linked to the modulating the gain or excitability of neurons that represent sensory and motor prediction errors.
In a fly brain, the mushroom bodies may play an analogous role in predictive coding (Sterling & Laughlin, 2015). (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 allostatic.)

This allows for the encoding of statistical patterns of uncertain value that can later be reconstructed when they are of use; (Dunsmoor, Murty, Davachi, & Phelps, 2015)).

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, for example, the topography of the salience network and the multimodal integration network found in (Sepulcre, Sabuncu, Yeo, Liu, & Johnson, 2012)).

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.

The theory of constructed emotion integrates ideas and empirical findings from neuroconstruction ((Mareschal et al., 2007); (Karmiloff-Smith, 2009); (Westermann et al., 2007)), rational constructivism (Xu & Kushnir, 2013), psychological construction (Barrett, 2006, 2012, 2013; Barrett et al., 2015) and social construction (Boiger & Mesquita, 2012).

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.
References


10.1196/annals.1440.011


Clark-Polner, E., Johnson, T., & Barrett, L. F. (in press). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cerebral Cortex.*


Swanson, L. W. (2012). *Brain architecture: understanding the basic plan*: Oxford University Press.


doi:10.1093/scan/nsu037


Acknowledgements

Many thanks to Ajay Satpute, Amitai Shenav, Suzanne Oosterwijk, and Eliza Bliss-Moreau who read and/or made invaluable comments on an earlier draft of this manuscript. This paper was prepared with support from the National Institutes on Aging (R01 AG030311), the National Cancer Institute (U01 CA193632), the National Science Foundation (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.
**Table 1**

*Examples of Neuroscience Evidence That Disconfirm the Classical View of Emotion*

<table>
<thead>
<tr>
<th>Observation</th>
<th>Method</th>
<th>Example Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different emotion categories cannot be specifically and consistently localized to distinct populations of neurons within a single region of the human brain.</td>
<td>Human neuroimaging: task-related data</td>
<td>(Lindquist et al., 2012; Vytal &amp; Hamann, 2010)</td>
</tr>
<tr>
<td>The instances of an emotion category need not share a population of neurons that are necessary and sufficient to implement them.</td>
<td>Human neuroimaging: multi-voxel pattern analysis</td>
<td>(Clark-Polner, Johnson, &amp; Barrett, in press)</td>
</tr>
<tr>
<td>Individual neurons, when stimulated, do not have emotion-specific receptive fields in studies of experience, expression, and perception.</td>
<td>Intracranial stimulation in humans</td>
<td>(Guillory &amp; Bujarski, 2014)</td>
</tr>
<tr>
<td>Lesions to the amygdala produce variable functional consequences. Monozygotic twins, whose basolateral nuclei of both amygdalae are calcified due to Urbach-Wiethe Disease (UWD) have calcification of the do not show</td>
<td>Behavioral observations in humans with amygdala lesions</td>
<td>(Ralph Adolphs, Russell, &amp; Tranel, 1999; Ralph Adolphs &amp; Tranel, 1999, 2003; Atkinson, Heberlein, &amp; Adolphs, 2007; A. Bechara,</td>
</tr>
</tbody>
</table>
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. Other people with basolateral lesions from UWD show different problems in fear perception (they are vigilant to rather than neglectful of 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).

Various circuits acting in parallel or collaboratively support the acquisition and performance of 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 appear to control different situation-specific behaviors, but others find that there are many (circuits) to one (behavior) mappings. Others find one (circuits) 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

Optogenetic research and some lesion research in rodents

(i.e., the way that learning is expressed depends on the context and the available options for behavior). Also, cortical regions (e.g., dorsomedial and ventromedial prefrontal cortex) 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 the concept of fear.

| 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 animals are restrained in position as they hear the tone, their heart rates decrease. | Classical conditioning in rats. | (Iwata & 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 prevent exploratory behavior in novel situations to 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 Lesions in non-human primates | (Antoniadis, Winslow, Davis, & Amaral, 2009; Mason, Capitanio, Machado, Mendoza, & Amaral, 2006) |
| rodent experiments |  |  |
Table 2

Selected Neuroscience Evidence Supporting the Theory of Constructed Emotion

<table>
<thead>
<tr>
<th>Observation</th>
<th>Method</th>
<th>Example Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category</td>
<td>Human neuroimaging: task-related data</td>
<td>(Lindquist et al., 2012; Suzanne Oosterwijk, Lindquist, Adebayo, &amp; Barrett, 2015; Vytal &amp; Hamann, 2010; Wilson-Mendenhall et al., 2015; Wilson-Mendenhall et al., 2011)</td>
</tr>
<tr>
<td>Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category</td>
<td>Human neuroimaging: multi-voxel pattern analysis</td>
<td>(Clark-Polner, Johnson, et al., in press); compare the different patterns for a given emotion category across (P. A. Kragel &amp; Labar, 2015; Saarimaki et al., 2016; Wager, 2015)</td>
</tr>
<tr>
<td>Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category</td>
<td>Intracranial stimulation in humans</td>
<td>(Guillory &amp; Bujarski, 2014)</td>
</tr>
<tr>
<td>Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category</td>
<td>Behavioral observations in humans with amygdala lesions</td>
<td>(Becker et al., 2012; Mihov et al., 2013)</td>
</tr>
<tr>
<td>Degeneracy: mapping many neurons, regions, networks, or patterns to one emotion category</td>
<td>Optogenetic research showing many to one mappings for behaviors in rodents</td>
<td>(Herry &amp; Johansen, 2014)</td>
</tr>
<tr>
<td>Neural reuse: Mapping one neural assembly to many emotion categories</td>
<td>Human neuroimaging: task-related data</td>
<td>(Lindquist et al., 2012; Vytal &amp; Hamann, 2010; Wilson-Mendenhall et al., 2011)</td>
</tr>
<tr>
<td>Neural reuse: Mapping one neural assembly to many emotion categories</td>
<td>Human neuroimaging: intrinsic connectivity data</td>
<td>(L F. Barrett &amp; Satpute, 2013; Touroutoglou et al., 2015; Wilson-Mendenhall et al., 2011)</td>
</tr>
<tr>
<td>Neural reuse: Mapping one neural assembly to many emotion categories</td>
<td>Optogenetic research and some lesion research in rodents</td>
<td>(Tovote et al., 2015)</td>
</tr>
<tr>
<td>Predictive coding accounts of aversive (“fear”) learning</td>
<td>Optogenetic research and some lesion research in rodents</td>
<td>(Furlong et al., 2010; Li &amp; McNally, 2014; McNally, Johansen, &amp; Blair, 2011)</td>
</tr>
<tr>
<td>Emotion concepts are embodied</td>
<td>Human neuroimaging: task-related data</td>
<td>(S. Oosterwijk et al., 2012; Suzanne Oosterwijk, Mackey, Wilson-Mendenhall, Winkielman, &amp; Paulus, 2015)</td>
</tr>
<tr>
<td>Multimodal summaries of emotion concepts are represented</td>
<td>Human neuroimaging: task-related data</td>
<td>(Peelen, Atkinson, &amp; Vuilleumier,</td>
</tr>
<tr>
<td>in the default mode network</td>
<td>related data</td>
<td>2010; Skerry &amp; Saxe, 2015)</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Default mode and salience network interconnectivity is associated with the intensity of emotional experiences (as distinct from arousal)</td>
<td>Human neuroimaging: task-related data</td>
<td>(Raz et al., 16)</td>
</tr>
<tr>
<td>Embodied simulations are associated with increased activity in primary interoceptive cortex</td>
<td>Human neuroimaging: task-related data</td>
<td>Wilson-Mendenhall, Henriques, Barsalou, &amp; Barrett, under review</td>
</tr>
</tbody>
</table>
Figure Captions

Figure 1. The classical view of emotion. The classical view of emotion includes basic emotion theories (e.g., for a review, see (Tracy & Randles, 2011)), causal appraisal theories (e.g., (Roseman, 2011; Scherer, 2009)), and theories of emotion that rely on black-box functionalism ((D. J. Anderson, & Adolphs, R., 2014; Davis, 1992)). 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 ellipses). 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. Figure 1A is adapted from (Davis, 1992). Figure 1B is adapted ((D. J. Anderson, & Adolphs, R., 2014). Figure 1C is adapted from (Barrett, 2006) which reviews the growing evidence that contracts the classical view of emotion).

Figure 2. Hubs in the human brain. (A) Hubs of the rich club, adapted from (Martijn P van den Heuvel & Olaf 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; (M. P. van den Heuvel & O. Sporns, 2013)). They are sometimes referred to as convergence or confluence zones (e.g., (A. R. Damasio, 1989; Meyer & 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 5,633 studies from the Neurosynth database. Activations are thresholded at FWE p < .05. Limbic regions (i.e., agranular/dysgranular with descending projections to visceromotor control nuclei) include the cingulate cortex (MCC, pgACC), medial prefrontal cortex (vmPFC), supplementary motor and premotor areas (SMA and PMA), medial temporal lobe (mTL), the anterior insula (aINS) and ventrolateral prefrontal cortex (vPFC) (e.g., (K.-J. Bar, de la Cruz, F., Schumann, A., Koehler, S., Sauer, H., Critchley, H., & Wagner, G., 2016; Carrive & Morgan, 2012); for a discussion, see (Kleckner, under review).

Figure 3. Neural activity during simulation. N = 16 (C. Wilson-Mendenhall et al., 2013). Participants listened with eyes closed to multimodal scenarios 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 < .05. Left image, x = 1; right image, x = -42. Heightened neural activity in primary visual cortex, somatosensory cortex (SSC), and motor cortex (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 posterior insula, pINS), in the sensory
integration network of lateral orbitofrontal cortex (lOFC; (Ongur, Ferry, & Price, 2003)) and in the thalamus; increased BOLD responses were also seen, as expected, in limbic regions such as the vmPFC, the aINS, the temporal pole (TP), SMA, and vIPFC, as well as in the hypothalamus and the subcortical nuclei that control the internal milieu.

Figure 4. The brain is a concept generator. (A) Brodmann areas are shaded to depict their degree of laminar development. The brain’s computational architecture for prediction generation and prediction error (laminar gradient adapted from (Barbas, 2015)), where information flows from agranular to granular regions, can also be thought of a concept construction (as described in (Barrett, 2017)). I hypothesize that agranular (i.e., limbic) cortices generatively combine past experiences to construct embodied concepts; multimodal summaries cascade to the 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 their dimensionality, eventually creating multimodal summaries, by virtue of its cytoarchitectural gradient; information flows from the upper layers of primary sensory and motor regions which are populated with many small pyramidal cells with few connections towards heteromodal regions (including limbic cortices) with fewer but larger pyramidal cells having many connections (Finlay & Uchiyama, 2015). (B) Multimodal summaries within the default mode network for emotion concepts (adapted from (Skerry & Saxe, 2015), Figure 1b); for sensory-motor properties (color, shape, visual motion, sound, and physical manipulation, (Fernandino et al., 2016), Figure 5); and, for semantic processing (adapted from (Jeffrey R Binder & 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, Wager, et al., in press); adapted from (Buhle et al., 2013) and (Satpute, Wilson-Mendenhall, Kleckner, & Barrett, 2015). Overlaps are observed in the insula ventrolateral prefrontal cortex (vIPFC), the anterior (aINS), the midcingulate cortex (mACC), supplementary motor area (SMA), and posterior superior temporal sulcus (pSTS). 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 number of activation or population centers, given a specific emotion category. Maps also reflect expectations as to co-activation patterns. Notice that population centers for all emotion categories can be found within the default mode and salience networks. Adapted from (Wager, 2015)

Figure 5. A depiction of predictive coding in the human brain. (A) We identified key limbic cortices (in blue) that provide cortical control the body’s internal milieu. Primary motor cortex 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., periaqueductal gray, parabrachial nucleus, 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 that are carried along the vagus nerve and small diameter C and AΔ fibers (dashed green paths) 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 motor cortex as motor predictions (solid lines) and prediction errors are sent from motor cortex 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 sends sensory predictions to sensory cortices that are more well-developed (e.g., in this figures, somatosensory and primary visual cortices). 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; valns = ventral anterior insula; daIns = dorsal anterior insula and includes ventrolateral prefrontal cortex; SMA = supplementary motor area; PMA = premotor area; 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, under review).

Figure 6. A large-scale system for allostasis in the human brain. (A) The system performing allostasis is composed of two large-scale intrinsic networks (shown in red and blue) that share several hubs (shown in purple; for coordinates, see (Kleckner, under review)). Hubs belonging to the “rich club” are shown in yellow. Rich club hubs figure adapted with permission from van den Heuval & Sporns (2013). 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 260 participants. aMCC = anterior midcingulate cortex; dAmy = dorsal amygdala; dPIns = dorsal posterior insula; dmIns = dorsal mid insula; IFG = inferior frontal gyrus; ITG = inferior temporal gyrus; lvaIns = lateral ventral anterior insula; MCC = midcingulate cortex; mvaIns = medial ventral anterior insula; pACC = pregenual anterior cingulate cortex; PHG = parahippocampal gyrus; pMCC = posterior midcingulate cortex; PostCG = postcentral gyrus; sgACC = subgenual anterior cingulate cortex; STS = superior temporal sulcus. (B) Reliable subcortical connections, thresholded \( p < 0.05 \) uncorrected, replicated in 270 participants.
Central Fear State

- Increased heart rate
- Decreased salivation
- Stomach ulcers
- Respiration change
- Scanning and vigilance
- Urination, defecation
- Increased startle
- Grooming
- Freezing

Central Emotion State

- Observed behavior
- Subjective reports
- Psychophysiology
- Cognitive changes
- Somatic responses

Emotion

- Feeling
- Facial Movements
- Vocal Acoustics
- Autonomic Nervous System Changes
- Actions

A.

Stimuli

B.

C.

Event
Box 1. The Curious Case of SM.

Much of our understanding of the neural basis of fear comes from studying Patient SM. 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., standing “fear learning” paradigms) and therefore has difficulty learning from prior errors (e.g., she does not mount an anticipatory skin conductance response to aversive learning, to the Iowa Gambling Task, and does not show loss aversion when gambling).

Interestingly, however, there is other evidence that SM is capable of what is typically called “fear learning.” SM 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).

SM’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). SM can even perceive fear in 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, and the faces that stereotyped fear poses contain widen eyes (yet, other patients with Urbach-Wiethe disease spend a longer time looking at the widened eyes of stereotyped fear poses and have not difficulty correctly categorizing those faces as fearful).

It should be noted (but rarely is) that SM’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 SM’s specific behavioral profile. It is also important to note that SM has had difficulties sustaining long-term relationships, including friendships, and is distressed by this, but there seems to be one clear exception: the scientists she has worked with for almost two decades. SM 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 SM is aware of their hypothesis that the amygdala contains the circuitry for fear.

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

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

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: 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: Pattern generators are groups of neurons (i.e., nuclei) that implement the sequences of actions that make up coordinated behaviors like feeding, or running, or 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 & Laughlin, 2015; Swanson, 2005).

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