OXFORD

Cerebral Cortex, March 2017;27:1944–1948

doi:10.1093/cercor/bhw028
Advance Access Publication Date: 29 February 2016
Commentary

COMMENTARY

Multivoxel Pattern Analysis Does Not Provide Evidence to Support the Existence of Basic Emotions

Elizabeth Clark-Polner[1], Timothy D. Johnson[2] and Lisa Feldman Barrett[3,4,5]

[1]Department of Psychology, University of Chicago, Chicago, IL, USA, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA, [3]Department of Psychology, Northeastern University, Boston, MA, USA, [4]Department of Psychiatry and [5]the Martinos Center for Biomedical Imaging, Massachusetts General Hospital/ Harvard Medical School, Boston, MA, USA

Address correspondence to Lisa Feldman Barrett, Department of Psychology, Northeastern University, Boston, MA 02115, USA. Email: l.barrett@neu.edu

Abstract

Saarimaki et al. (2015) published a paper claiming to find the neural “fingerprints” for anger, fear, disgust, happiness, sadness, and surprise using multivariate pattern analysis. There are 2 ways in which Saarimaki et al.’s interpretation mischaracterizes their actual findings. The first is statistical: a pattern that successfully distinguishes the members of one category from the members of another (with an accuracy greater than that which might be expected by chance) is not a “fingerprint” (i.e., an essence); it is an abstract, statistical summary of a variable population of instances. The second way in which Saarimaki et al.’s interpretation mischaracterizes their results is conceptual: their findings do not actually meet the specific criteria for basic emotion theory. Instead, their findings are more consistent with a theory of constructed emotion. In our view, Saarimaki et al. is elegant in method and important in that it demonstrates empirical support for a theory of emotion that relies on population thinking; it is also an example of how essentialism—the belief that all instances of a category possesses necessary features that define what is, and what is not, a category member—contributes to a fundamental misunderstanding of the neural basis of emotion.

Key words: basic emotion theory, conceptual act theory, construction, emotion, essentialism, multivoxel pattern analysis, pattern classification

Introduction

Over the past decade, multivariate pattern and machine learning analyses have become increasingly popular tools for use in human functional neuroimaging (Kamitani and Tong 2005; Haynes and Rees 2006; Norman et al. 2006). These approaches are now in widespread use across cognitive, affective, and social neuroscience, receiving particular attention when applied to questions for which older techniques have failed to produce the expected results. One example comes from affective neuroscience, where it has long been hypothesized that a small set of emotion categories (e.g., anger, fear, disgust, happiness, sadness, and surprise) have distinct neural essences that support different survival functions (for a recent discussion, see Tracy and Randles 2011, who wrote, the “agreed-upon gold standard is the presence of neurons *dedicated* to the emotion’s activation” p. 398; emphasis added). This hypothesis belongs to the theory of basic emotions, which hypothesizes that certain categories of emotion are biologically primitive (meaning that they are biological building blocks that cannot be further reduced to more fundamental mechanisms at the biological level). The theory of basic emotions is an example of a broader faculty psychology approach to the brain, within which it is assumed that psychological categories are natural kinds with an essence in the Lockean sense—for example, a shared neural circuit that is the underlying cause of instances of anger, making them anger and not some other emotion, such as fear (Barrett 2006; Barrett et al. 2007; Lindquist and Barrett 2012). Recently, Saarimäki et al. (2015) published a paper claiming to find the neural “fingerprints” for anger, fear, disgust, happiness, sadness, and surprise using multivariate pattern analysis (p. 8). They successfully created a multivoxel pattern for each category that diagnosed new contrast maps with an accuracy above what would be expected by chance. Saarimaki et al. interpreted their findings as empirical support for basic emotion theory (as the title of their paper indicates).

There are two ways in which Saarimaki et al.’s interpretation mischaracterizes their actual findings. The first is statistical: A pattern that successfully distinguishes the members of one category from the members of another (with an accuracy greater than that which might be expected by chance) is not a “fingerprint” or an essence. Just as we use “fingerprint” colloquially to suggest something that is both unique to 1 person, and unchanging, the implication is that the specified configuration of voxels is both unique to the specified emotion and unchanging—in other words, it should appear in every instance of that emotion (human fingerprints are used to distinguish one person from another. The fingerprints for a given individual do not look identical each time, due to variation in the surface as well as other variables such as grip strength, skin temperature, etc., but they are similar enough to function as a unique identifier, because they share an underlying essence [the skin ridges on the finger pad that leave marks on surfaces]). This is, however, mathematically incorrect; the voxels that make up a “pattern” for a category do not represent a configuration that is seen in every (or even any) single instance of that category. We demonstrate this with a simulation in Figure 1. First, we created a simple “brain”, consisting of 1000 voxels. We then specified 5 distinct patterns of voxels—each of which corresponds to a different “category,” consistent with the basic emotion theory hypothesis that anger, fear, disgust, happiness, sadness, and surprise are each represented by a unique, fixed pattern (a neural essence). We then generated 200 individual instances from each pattern (1000 in total), adding random noise to the specified “essence” to mirror the variability due to induction method, subject, and conditions within a true imaging experiment (for details, see Fig. 1). Finally, we trained a classifier on the original “essences” and tested these classifiers on the generated individual instances, using the same type of machine learning algorithm as reported in Saarimaki et al. Our simulation achieved accuracy rates similar to (and even higher than) those reported by Saarimaki et al., but the pattern of voxels (indeed, even a single voxel) was not present in every instance of its associated category (see Fig. 2).

A pattern classifier that successfully diagnoses the members of an emotion category does not produce the brain state for that emotion category. Multivoxel pattern analysis and other forms of pattern classification work by the logic of population thinking. According to the evolutionary biologist Ernst Mayr (2004), one of Darwin’s greatest innovations was to vanquish essentialism

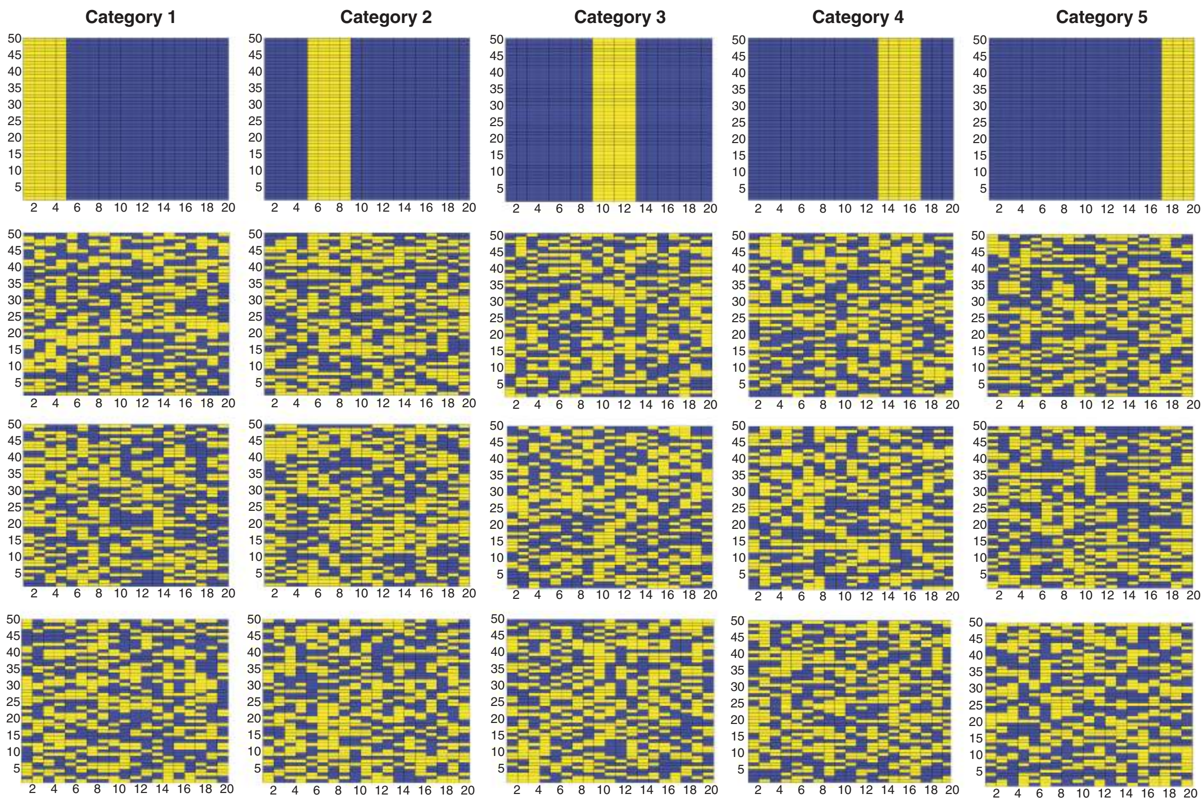

Figure 1. Multivoxel pattern simulation. We constructed a simple simulation using a “brain” consisting of 1000 voxels. We specified 5 distinct patterns of voxels, each representing a hypothetical “category” and then trained a naive Bayesian classifier on these patterns. Next, we generated 1000 individual instances (“trials”), based on these patterns, adding random noise to remodel the noise that would occur in a true imaging experiment. We added the amount of noise that would produce accuracy rates similar to those reported in Saarimaki et al. (see Fig. 2d; “movies across subjects”): for each category (i.e., each “emotion”), there was a 0.52–0.58 probability of activation for voxels that were part of the intended pattern, and a 0.5 probability of activation for all other voxels. Here we picture the “neural essence” for each category and examples of correctly classified instances. With this level of noise, we were able to achieve accuracy rates similar to (and even higher than) those reported by Saarimaki et al. (0.455, 0.75, 0.88, 0.61, 0.715). Even with these above chance levels of classification accuracy, none of the patterns are seen in every correctly classified instance, demonstrating that the patterns produced by the classification analysis do not reflect “fingerprints” that are seen in every incidence of a category.

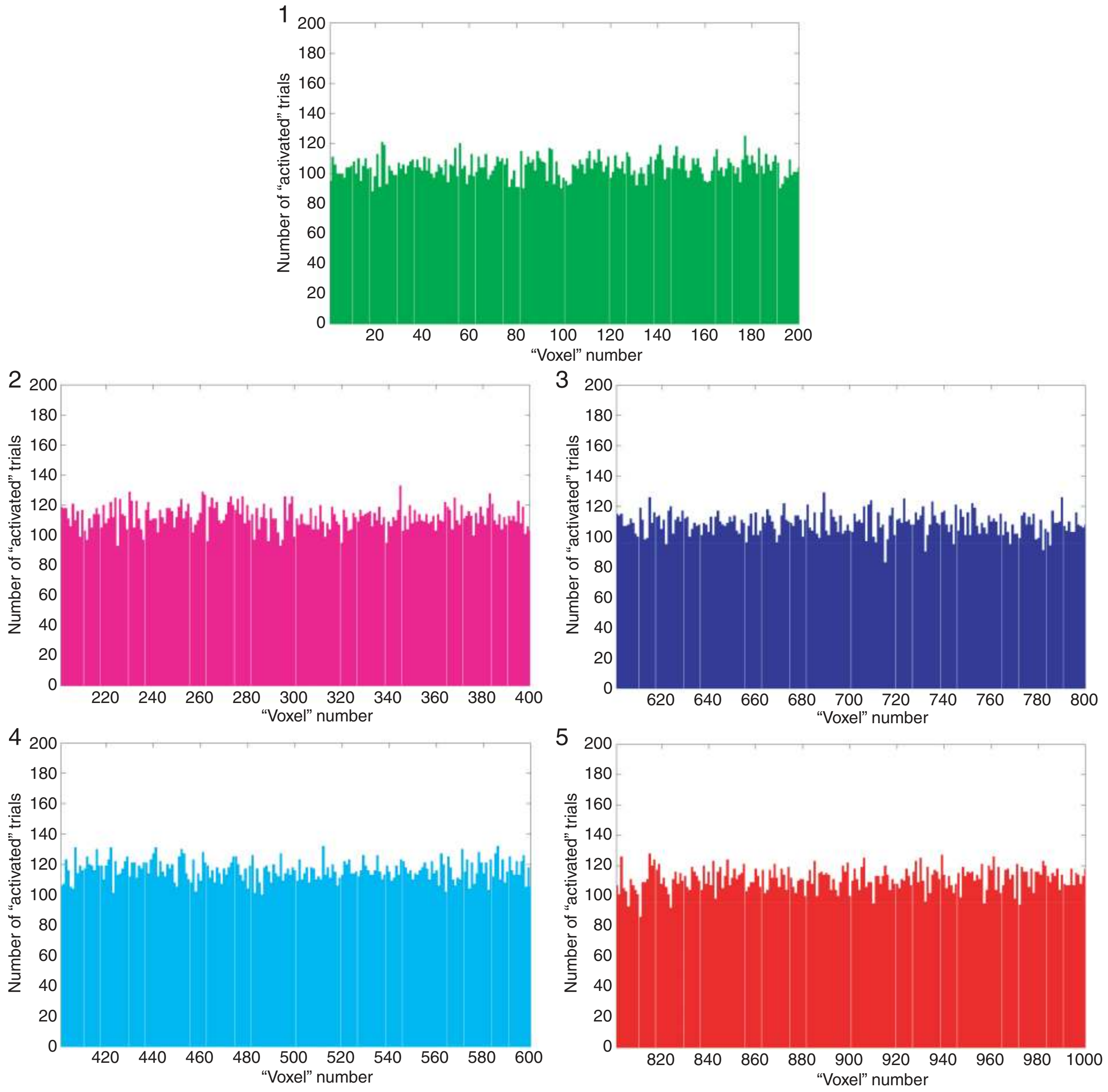

Figure 2. No subsets of "voxels" exist within all instances of any category. There is no single "voxel" for a given pattern that appears in every successfully classified instance of a category. "Voxel" numbers are displayed on the x-axis. Number 1–200 belong to Category 1; no. 201–400 belong to Category 2; no. 401–600 belong to Category 3; no. 601–800 belong to Category 4; no. 801–1000 belong to Category 5. The number of instances within each category in which each "voxel" was "activated" is displayed on the y-axis.

and show that a biological category, such as a species, is not a physical type with a Lockean-type biological essence. Biological categories are conceptual categories populated by unique and highly variable individuals that do not share any necessary features. Any statistical summary of a category is an abstraction that does not necessarily exist in nature. This is also how emotion categories work (Barrett 2013). Although as a group the instances of any emotion category can be diagnosed with a pattern, the pattern itself is an abstraction and does not necessarily describe one feature or set of features that is necessary for every (or even any) single individual instance in the category (see Posner and Keele 1968 for an example of population thinking in abstract categories using random dots). Analogously, the average middle class U.S. family has 3.13 children, but no family actually has 3.13 children (the number is an abstract summary of families with different numbers of children). Thus, in a statistical sense, Saarimaki et al. did not find evidence to support the theory of basic emotion, nor the existence of biologically basic emotion categories.

The second way in which Saarimaki et al.'s interpretation mischaracterizes their results is conceptual: Several versions of basic emotion theory exist (Ekman and Cordaro 2011; Izard 2011; Levenson 2011; Panksepp and Watt 2011; for summary and comparison, see Tracy and Randles 2011), but all agree that several criteria must be met for an emotion to be considered "basic," only two of which are relevant to our discussion here. First and foremost, each emotion category is thought to have a neural essence (dedicated and distinct neural circuitry) shared by all instances. For example, in their summary of basic emotion theories, Tracy and Randles (2011, 398) wrote, "while individual and cultural learning can change the conditions and intensity with which basic emotions are activated and experienced, they cannot create, do novo, a basic emotion that is not already possible via genetically encoded neural structures," where each emotion has its own "distinct neurology" (p. 398). Yet, the voxels constituting a pattern were not unique to and dedicated to any single emotion category, as we have already demonstrated. Also emotion essences are presumed to be homologous across species (i.e., evolutionarily old and heritable). For example, from Tracy and Randles (2011, p. 398) "All agree that cross-species generalization. . .. is a clear indicator".

The results reported by Saarimaki et al. do not meet either of these criteria. The voxels constituting a pattern were not unique to and dedicated to any single emotion category, as we have already demonstrated. Furthermore, it is unclear how sparsely distributed patterns can be homologous across different species (and therefore easily heritable) [An additional flaw in the logic of basic emotion theories is that even if different species share the same brain structures at a macro level, structural homology does not guarantee functional homology. For example, macaque monkeys have a definable default mode network (Mantini et al. 2013), although there is no evidence that they use this network to mentalize or engage in mental time travel as humans do. Chimpanzees have the equivalent of a human language network, although they do not possess the capacity for human language (Sherwood et al. 2012). Human, chimp, and macaque brains differ in their microwiring (Finlay and Uchiyama 2015) because as brains grow larger, they re-organize (like companies) rather than just lay down new tissue on old (like sedimentary rock; Striedter 2005)]. Moreover basic emotion theories are typically based on Maclean's outdated (but still frequently cited) "triune brain" (Maclean 1990) which localizes emotion essences to limbic circuitry deep within subcortical regions (Panksepp 1998), sandwiched between the "reptilian brain" in the brainstem (for hunger, thirst, etc.) and the isocortex (sometimes called the neo-cortex) where cognitive functions are typically localized. Again from Tracy and Randles (2011, p. 398), "basic emotions are primitive in that they must originate in subcortical brain structures." The patterns described by Saarimaki et al. clearly violate this requirement because they are distributed throughout the brain and are concentrated relatively more heavily within the isocortex. Thus, the results reported in Saarimaki et al. do not meet the most elementary criteria for basic emotions laid out by recent discussions of basic emotion theories. More generally, the ability to categorize a set of instances in no way implies that the underlying processes are themselves categorical (which is also assumed by basic emotion theory). This is a common misperception that interferes with scientific thinking and can hinder efforts to better understand the nature of emotion and its neural bases. For example, we perceive human speech as a string of discrete sounds (phonemes), but the underlying acoustical signals are continuous.

The Saarimaki et al. paper is elegant in method, and the results are impressive in that they do provide empirical support for a theory of emotion that relies on population thinking. This is the conceptual act theory of emotion (Barrett 2012, 2013; Lindquist and Barrett 2012; Barrett et al. 2015), which belongs to a family of constructionist emotion theories (for a review, see Barrett and Russell 2015; Cunningham 2013; Oosterwijk et al. 2015). The conceptual act theory explicitly hypothesizes that an emotion word like "anger" refers to a population of variable, situation-specific instances, and that emotion categories do not have neural essences, although, as categories, they certainly could be successfully diagnosed by pattern classification. According to the conceptual act theory each instance of emotion is hypothesized to arise from an interaction of core networks within the intrinsic architecture of the brain (for the latest version, see Barrett and Satpute 2013; Barrett and Simmons 2015; Barrett et al. 2015). The conceptual act theory specifically hypothesizes that regions of the default mode network, the salience network, and other intrinsic networks are important for constructing instances of emotion. For example, one specific hypothesis is that medial prefrontal and posterior cingulate cortices/precuneus are important for predicting and conceptualizing sensory inputs, including interoceptive inputs from the body's internal milieu; Barrett 2012; Barrett and Simmons 2015); this is exactly what Saarimaki and colleagues report. In fact, their findings are consistent with a growing body of literature that has failed to find evidence that emotion categories are biological basic and instead provide direct evidence for the conceptual act theory (e.g., see our pattern classification analysis of an emotion meta-analytic database, Wager et al. 2015; see recent meta-analyses of traditionally analyzed neuroimaging studies; Kober et al. 2008; Lindquist et al. 2012; see also Hamann (2012)'s discussion of Vytal and Hamann 2010; intrinsic connectivity evidence from resting-state fMRI analyses, Touroutoglou et al. 2015; a review of intracranial stimulation in humans, Guillory and Bujarski 2014; a review of lesion studies in humans, Lindquist et al. 2012). In fact, there are no neurons (configured as regions, circuits, or networks) that are specific to the category of emotion more broadly (Barrett and Satpute 2013).

In our view, Saarimaki et al. is an example of how essentialism—the belief that all instances of a category possesses fundamental, necessary features (i.e., an essence) to define what is, and what is not, a category member—interferes with the interpretation of pattern classification, leading to a basic misunderstanding of the neural basis of emotions. They are not alone. Other neuroimaging studies using classifiers to empirically distinguish one emotion category from another make similar errors (Kassam et al. 2013; Kragel and LaBar 2015), as do studies of facial movements (i.e., "expressions") or autonomic nervous system responses (e.g., Yuen et al. 2012; Park et al. 2013; Kragel and LaBar 2014)]. Essentialism's tendency to interfere with scientific progress is not unique to the science of emotion. It interferes with scientific understanding more generally, particularly in relation to natural selection and evolution (Gelman and Rhodes 2012; also, see a number of essays in the recent volume, "This Idea Must Die," Brockman 2015).

Some critics might suggest that the criteria we have laid out here set an unrealistically high bar for evidence supporting the existence of basic emotions. We would not disagree, but would only add that the criteria we listed do not originate with us; they reside in the claims of basic emotions theories themselves. If the criteria seem unrealistic, it is because the theories themselves are so.

Funding

This work was supported by a US National Institute on Aging grant (R01AG030311) and contracts from the US Army Research Institute for the Behavioral and Social Sciences (Contract W5J9CQ-12-C-0049, W5J9CQ-11-C-0046) to L.F.B. The views, opinions, and findings contained in this article are those of the authors and should not be construed as an official position, policy, or decision of the US National Institutes of Health or Department of the Army unless so designated by other documents.

Notes

Conflict of Interest: None declared.

References

Barrett LF. 2006. Are emotions natural kinds? Perspect Psychol Sci. 1:28–58.

Barrett LF. 2012. Emotions are real. Emotion. 12:413–429.

Barrett LF. 2013. Psychological construction: a Darwinian approach to the science of emotion. Emot Rev. 5:379–389.

Barrett LF, Lindquist KA, Bliss-Moreau E, Duncan S, Gendron M, Mize J, Brennan L. 2007. Of mice and men: natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. Perspect Psychol Sci. 2:297–312.

Barrett LF, Russell JA. 2015. The psychological construction of emotion. New York: Guilford.

Barrett LF, Satpute A. 2013. Large-scale brain networks in affective and social neuroscience: Towards an integrative architecture of the human brain. Curr Opin Neurobio. 23:361–372.

Barrett LF, Simmons WK. 2015. Interoceptive predictions in the brain. Nat Rev Neuro. 16:419–429.

Barrett LF, Wilson-Mendenhall CD, Barsalou LW. 2015. The conceptual act theory: a road map. In: Feldman Barrett E, Russell J, editors. The psychological construction of emotion. New York: Guilford Press. p. 83–110.

Brockman J ,editor. 2015. This idea must die. New York: Harper.

Cunningham WA. 2013. Introduction to special section: psychological constructivism. Emot Rev. 5:333–334.

Ekman P, Cordaro D. 2011. What is meant by calling emotions basic. Emot Rev. 3:364–370.

Finlay BL, Uchiyama R. 2015. Developmental mechanisms channeling cortical evolution. Trends Neurosci. 38:69–76.

Gelman SA, Rhodes M. 2012. "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding. In: Rosengren KS, Brem S, Evans EM, Sinatra G, editors. Evolution challenges: integrating research and practice in teaching and learning about evolution. New York: Oxford University Press.

Guillory SA, Bujarski KA. 2014. Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. Soc Cogn Affect Neurosci. nsu002.

Hamann S. 2012. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn Sci. 16:458–466.

Haynes JD, Rees G. 2006. Decoding mental states from brain activity in humans. Nat Rev Neuro. 7:523–534.

Izard CE. 2011. Forms and functions of emotions: matters of emotion–cognition interactions. Emot Rev. 3:371–378.

Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. Nat Neuro. 8:679–685.

Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA. 2013. Identifying emotions on the basis of neural activation. PLoS ONE. 8:e66032.

Kober H, Barrett LF, Joseph J, Bliss-Moreau E, Lindquist K, Wager TD. 2008. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage. 42:998–1031.

Kragel PA, LaBar KS. 2015. Multivariate neural biomarkers of emotional states are categorically distinct. Soc Cogn Affect Neurosci. 9:1880–1889.

Kragel PA, LaBar KS. 2014. Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. Emotion. 13:681–690.

Levenson RW. 2011. Basic emotion questions. Emot Rev. 3:379–386.

Lindquist KA, Barrett LF. 2012. A functional architecture of the human brain: insights from Emotion. Trends Cogn Sci. 16:533–540.

Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF. 2012. The brain basis of emotion: a meta-analytic review. Behav Brain Sci. 35:121–143.

MacLean PD. 1990. The triune brain in evolution. New York: Plenum Press.

Mantini D, Corbetta M, Romani GL, Orban GA, Vanduffel W. 2013. Evolutionarily novel functional networks in the human brain? J Neurosci. 33:3259–3275.

Mayr E. 2004. 80 years of watching the evolutionary scenery. Science. 305:46–47.

Norman KA, Polyn SN, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 10:424–430.

Oosterwijk S, Touroutoglou A, Lindquist KA. 2015. The neuroscience of construction: What neuroimaging can tell us about how the brain creates the mind. In: Barrett LF, Russell JA, editors. The psychological construction of emotion. New York: Guilford.

Panksepp J. 1998. Affective neuroscience: the foundations of human and animal emotions. Oxford, UK: Oxford University Press.

Panksepp J, Watt D. 2011. What is basic about basic emotions? Lasting lessons from affective neuroscience. Emot Rev. 3:387–396.

Park BJ, Jang EH, Chung MA, Kim SH. 2013. Design of prototype-based emotion recognizer using physiological signals. ETRI J. 35:869–879.

Posner MI, Keele SW. 1968. On the genesis of abstract ideas. J Exp Psychol. 77:353.

Saarimäki H, Gotsopoulos A, Jääskeläinen IP, Lampinen J, Vuilleumier P, Hari R, Nummenmaa L. 2015. Discrete neural signatures of basic emotions. Cereb Cortex. 26:2563–2573.

Sherwood CC, Bauernfeind AL, Bianchi S, Raghanti MA, Hof PR. 2012. Human brain evolution writ large and small. In: Hofman MA, Falk D, editors. Progress in brain research. Amsterdam: Elsevier.

Striedter GF. 2005. Principles of brain evolution. Sunderland, MA: Sinauer Associates, Inc.

Touroutoglou A, Lindquist KA, Dickerson BC, Barrett LF. 2015. Intrinsic connectivity in the human brain does not reveal networks for "basic" emotions. Soc Cogn Affect Neuro. doi:10.1093/scan/nsv013/.

Tracy JL, Randles D. 2011. Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. Emot Rev. 3:397–405.

Vytal K, Hamann S. 2010. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J Cogn Neurosci. 22:2864–2885.

Wager TD, Kang J, Johnson TD, Nichols TE, Satpute AB, Barrett LF. 2015. A Bayesian model of category-specific emotional brain responses. PLoS Comput Bio. 11:e1004066.

Yuen KSL, Johnston SJ, De Martino F, Sorger B, Formisano E, Linden DEJ, Goebel R. 2012. Pattern classification predicts individuals responses to affective stimuli. Transl Neurosci. 3: 278–287.