

How are emotions organized in the brain?

Tor D. Wager¹
Anjali Krishnan^{1,2}
Emma Hitchcock¹

¹ University of Colorado Boulder

² Brooklyn College of the City University of New York

Running Head: THE EMOTIONAL BRAIN

Please address correspondence to:

Tor D. Wager
Department of Psychology and Neuroscience
University of Colorado, Boulder
345 UCB

Boulder, CO 80309

Email: tor.wager@colorado.edu

Telephone: (303) 895-8739

SUMMARY FOR THIS DRAFT (OPTIONAL)

5 figures; 6 supp. Figures (online); 4 supp. Tables (online)

Character count:

38,830 characters with spaces

Figures: est. 10000 characters.

Note: for guidelines, see:

<http://flash1r.apa.org/apastyle/basics/index.htm>

Acknowledgements

Thanks to Choong-Wan Woo for help with Figure 2C.

Are there emotion systems in the brain? We have to admit we are very tempted to just say yes. Of course there are emotion systems, because...we feel emotions. But that does that mean that there is a system or systems dedicated only to emotion? We also shop, and canoe, and stargaze, but we did not evolve a 'purpose-built' canoeing system. Instead canoeing requires a range of related capacities. Likewise, emotion systems did not evolve specifically to create emotional experiences, but to allow us to respond to the most important threats and opportunities in our ancestral environment.

Emotions are emergent properties that reflect activity in layers of systems evolved for action. For instance, threats related to body temperature are solved by thermoregulation, a set of coordinated, multi-system actions taken to mitigate cold or heat, including complex 'voluntary' behavior. Imagine falling into a frozen lake: Your body hair might stand on end, your blood vessels constrict, and your metabolism may slow down. You will shiver, light a fire, ask for help, and form long-term memories that will prevent a reoccurrence. These are affective behaviors, and they include both skeletomotor actions and physiological responses in the autonomic and neuroendocrine systems. They constitute some of the building blocks of emotion. Other systems that contribute to emotions evolved to respond to insufficiencies in oxygen, salt, glucose, and water; to maintain bodily integrity; to pursue mating opportunities; and to avoid microbial infection, predators, loss of property or damage to social relationships.

Many of the mechanisms that support these affective processes have their oldest roots in the brainstem and even the spinal cord, because they are so fundamental that they pre-date the development of the brain itself. But our emotional life is also built upon the brain's newer innovations, particularly our ability to represent abstract concepts, recall specific episodes from our past, imagine future possibilities in vivid detail, and simulate alternative present or

future realities. Thus, the space of mechanisms that contribute to emotion is vast, from basic affective processes related to tissue damage or infection (e.g., pro-inflammatory cytokines appear to be powerful drivers of emotion) to complex capacities such as gratitude, which allow us to contribute to causes larger than ourselves. Some of these processes are affective (related to survival and well being) while others are cognitive. In sum, emotions are combinations of affective and cognitive processes deployed in particular ways in particular situations.

Here, we explore the brain basis for emotional experiences. Some aspects we consider are their multiple points of origin in the brain, the diversity of brain regions that are likely involved in a single emotional episode, and the role that conceptualization plays in shaping emotion. Although there is not one simple system for emotion, we describe some ways of moving towards identifying brain representations related to emotional experience and other affective outcomes.

An emotional brain without borders, but not formless

In 1949, Paul MacLean coined the term 'limbic system' (MacLean, 1949). The term refers to a border, or 'limbus', between the cortex and the brainstem regions governing our physiology and basic, 'reflexive' behaviors. But the limbic system is not a boundary; it is a connector, a set of bridges between abstract thought and the oldest environment-action contingencies stored in our brains. The 'limbic system' came to be synonymous with 'emotional brain'. However, its very position as a connector of systems implies that it is not the sole seat of emotion.

Paul Maclean also developed the concept of the 'triune brain,' which divides the brain into three broad zones. We also find it useful to divide the brain into zones associated with different classes of functions, though they differ from MacLean's zones in their specifics (Figure 1).

First, the brainstem and spinal cord govern physiological homeostasis and behavioral responses to environmental challenges that are autonomous and reflexive. The isolated spinal cord is capable of ‘emotional behaviors’ such as shock avoidance and affective learning (Grau, 2014), and the brainstem is capable of coordinated responses to the environment that involve multiple skeletomotor and visceromotor systems (Berntson & Micco, 1976; Saper, 2002).

Second, the upper brainstem, including the periaqueductal gray, and a set of regions including the hypothalamus, ventral striatum and, amygdala, and others, regulate affective responses based on a more complex set of organism-environment interactions. These regions govern behaviors such as fighting, nurturing, and sex. Their processes depend on appraisal of the environment: whether another organism is displaying aggression, whether it is weak (fight) or strong (flee), and whether you are injured (if losing blood, vasoconstrict, immobilize, and ‘pass out’). Additionally, information from episodic memories, encoded initially in hippocampal systems, begins to enter the equation here, as do processes related to other evolutionarily ancient portions of the cortex like the agranular (ventral) insula.

Third, the cortex governs these organism-environment interactions and shapes the brain’s responses based on increasingly complex contingencies, including analysis of memories and background knowledge (medial temporal systems), inferences about the intentions of others (dorsomedial prefrontal cortex, dmPFC), situational meaning and value based on one’s goals (frontal poles), and conceptual value and thoughts about the future (ventromedial prefrontal cortex, vmPFC; (Schacter, Addis, & Buckner, 2007). In naming these regions, we do not mean to imply that any region acts alone in implementing one function—such as emotion— because these high-level functions require complex networks.

Though every brain region is an integrator of circuits, we believe the vmPFC plays a special role in constructing *affective meaning* and using it to generate

strong emotional states that include physiological changes. It is perhaps unique in integrating (a) multisensory affective information from the orbitofrontal cortices, (b) episodic and semantic memory from the medial temporal lobes, (c) imagination and prospection, and (d) viscerosensory and interoceptive information from the insula.

There cannot be ‘an emotional brain’ that is less than the brain in its entirety, because so many processes contribute to the generation of emotional experiences. But are there certain processes that we should consider *necessary* for emotion? Affective responses can be altered by processing in many areas of the cortex, but must be connected with systems that drive motivation and physiology in order to be truly ‘affective.’ Affect requires the participation of a specialized subset of classical ‘emotional brain’ regions: vmPFC, dorsal anterior cingulate and insula, periaqueductal gray, amygdala, ventral striatum and thalamus, parabrachial complex and solitary nucleus, and other vagal sensory and motor nuclei. Thus, the ‘emotional brain’ is not circumscribed to these regions, but they give affect the vigor and motivational qualities that differentiate it from other forms of cognition.

When we think about the ‘emotional brain’, we should keep in mind the varieties of prototypical emotions that we experience at three levels—brainstem, basal forebrain, and cortical. Many emotions might depend on interactions across all levels. Others may exist primarily at one level, with minimal involvement from others.

Different shades, different systems

It is very tempting to think of discrete emotion categories, such as ‘anger’ and ‘fear,’ as being instantiated in circuits dedicated to those categories, but such conclusions might be misleading. Some of our recent work (Wager et al., 2015)

and others (Kassam, Markey, Cherkassky, Loewenstein, & Just, 2013; Kragel & LaBar, 2015; Saarimäki et al., 2015; Vytal & Hamann, 2010) suggests that it is possible to differentiate patterns of brain activity for different emotion categories in a way that allows us to predict which ‘type’ of emotion a person is feeling based on the brain pattern in a way that generalizes across the types of materials used to evoke emotion and across individuals and studies. At first blush, if ‘anger’ produces a distinct neural ‘signature’, one might infer that this is the ‘anger system.’ However, as we explain below, this may not be the right conclusion.

First, the systems involved in these emotion ‘patterns’ are distributed across the brain, including contributions from cortical areas that are unlikely to contain emotion category-specific neurons, such as the lateral prefrontal and motor cortices. Second, though different categories of affective experience are encoded in different neurons in some brain regions, in others there is substantial convergence, with individual neurons that respond to multiple distinct affective challenges (Morrison & Salzman, 2009; Xiu et al., 2014). Third, it is unclear whether the emotion-predictive patterns are *consistent* across these studies. If three studies identify three different neural patterns related to ‘anger’, they may not be capturing the essence of ‘anger’ at all, but ideomotor and conceptual processes that go along with ‘anger’ only in a limited experimental context. Fourth, there is an issue of *generalization* of these emotional ‘signatures’ across the range of experiences of a given category. This issue is just beginning to be explored, but there are some important theoretical limitations.

We believe that the patterns that distinguish ‘anger,’ ‘fear,’ and others in these studies do not likely apply to all *instances* of those emotions (Barrett & Satpute, 2013), if for no other reason than our language is too vague to capture the meaningful distinctions between emotions. An emotion label such as ‘anger’ can be used to describe many heterogeneous experiences that are likely to be supported by different brain systems. For example, we might use the word

'anger' to describe both 'hot' shades of aggression that can be triggered by remembering a past injustice and 'cold' shades that do not involve immediate physiological arousal.

To make this concrete, imagine a dog that, backed into a corner, postures and snarls in defensive rage. This is a form of 'anger' that can be produced by brainstem systems alone, without any input from the cortex or forebrain. (Berntson & Micco, 1976). In dogs, it usually lasts for moments, often without any repercussions for long-term aggression. It occurs in humans, too: Imagine your young child crying over a broken cookie (to five-year-olds, cookies must be whole to be edible). You experience a spike of annoyance, which quickly disappears. Is this the same emotion as the dog's defensive aggression? Now imagine a colleague that has stolen a research idea from another and published it in a prestigious journal. Or imagine that you are consistently excluded by your compatriots at work or school. You might report feeling 'angry' at any moment during the weeks or months of these ordeals, yet have no immediate impulse to aggression and none of the physiological arousal that marks the defensive posturing of the cornered dog. It is an open question whether all these types of 'anger' are more similar to each other than they are to any other instances of any other emotion categories. For example, perhaps the anger of moral injustice is more closely related to disgust than annoyance, and the anger of rejection more akin to sadness than rage.

Other emotion categories, such as 'fear' and 'sadness,' are similarly diverse. Fear can be a "that snake startled me" fear; an "I'm afraid that people won't like me" fear; or an "I'm afraid of not having enough money when I retire" fear, among others. Neuroscience literature has focused overwhelmingly on the first kind. It is sometimes assumed that the neural circuitry involved in the startle response generalizes to other types of fear and anxiety, though different circuits are involved depending on the context, level of ambiguity, and behavioral and

autonomic responses afforded. Similarly, sadness can range from being extremely *visceral* (e.g., feeling like one has been “punched in the stomach” after a bad break-up) to being *existential* (e.g., feeling as though the “color has been drained from the world”). The sadness of melancholy may have little in common with the deep, desperate grief that accompanies great loss. It is unlikely that these instances rely on activation of the same neural circuits. It is even possible that some instances of ‘anger,’ ‘sadness,’ and other emotions are *entirely* subcortical, whereas others are *entirely* cortical, with no immediate physiological consequences.

The joints at which we carve emotion into categories are so compelling in part because they are embedded deeply in our language. Some of the emotional experiences described above do not fit our previous description of affective responding (i.e., characterized by motivational vigor), and perhaps they would be better classified as persistent cognitive styles. However, they are often grouped under the categories of mood and emotion because people describe them using emotion words. In fact, it is possible that the category-specific brain responses described above are patterns related to learned emotion *concepts* rather than core experiences. We have concepts of ‘anger’ that transfer across stimulus modalities—an angry shape is spiky, angry music is loud and abrupt, and angry movement is jerky (Sievers, Polansky, Casey, & Wheatley, 2013). Concepts of ‘anger,’ ‘fear,’ and other emotion categories are reliably distinct from one another across cultures, though seeing an ‘angry shape’ or an angry face may have little in common with the experience of rage. This is why there might be no coherent neuroscience of emotions as defined purely by self-report. For all the variety and precision our language affords in some ways, it is unlikely that it captures many of the fundamental distinctions among the varieties of emotional processes likely to map onto different brain systems.

This does not mean that affect and emotion cannot be studied scientifically. It

only means that the prototypical emotion categories are not the only way (or, perhaps, the best way) to organize how we study the emotional brain. Rather than lumping emotions into categories—whether the categories are ‘anger’ and ‘fear’ or ‘positive’ and ‘negative’—we may need to group emotional responses based on their antecedents, behavioral and physiological response tendencies, time course, and other variables.

Towards understanding emotional brain representations

It is not possible to identify a brain representation for an *emotion* per se, because an emotion is not a single process, but rather a collection of processes. We typically associate emotions such as ‘sadness’ with subjective experience, but changes in physiology, action tendencies, motivated behavior, memory, and perception are also often included in the definition. Thus, ‘sadness’ is a word that we use to describe the sum of many representations in many systems.

We can, however, identify brain ‘signatures’ related to specific aspects of emotion, including the intensity of self-reported emotional experience, which is a good place to start. This is what recent studies using multi-voxel pattern analysis and related multivariate statistical models have begun to do. For instance, we recently identified a pattern of fMRI activity that strongly predicts the intensity of negative emotional experience (Figure 2; (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015). The pattern is a weighted average of activity in multiple brain systems, from the brainstem to the prefrontal cortex, and the pattern as a whole responds more strongly to intensely negative than mildly negative images in 180 out of 180 participants in our sample. This brain pattern likely tracks the average of a set of processes that are related to emotion reports. Identifying this pattern does not mean we identified *the* pattern for negative emotion, or even that “negative emotion” is tracked by any single system across contexts. However, it provides a starting point for investigating the effects of interventions and

contextual variables on objective brain measures that are *strongly related to* emotional experience, and thus likely to be relevant for emotion in ways that general brain activity is not.

Another benefit of identifying brain patterns that target specific emotion-related outcomes is that it provides a basis for investigating the brain regions and systems that are *necessary* and *sufficient* for predicting the outcome. We cannot assess whether they are causal, because we cannot manipulate those precise circuits in their entirety and without off-target effects. However, we can do what might be called a ‘virtual lesion’ analysis, which looks at the effects of including or excluding different parts of the overall predictive pattern. In the case of negative emotional responses to aversive pictures, the predictive pattern involves many systems of the brain, spanning specific patterns within the prefrontal cortex, medial temporal lobes, and subcortical regions such as the amygdala, hypothalamus, and periaqueductal gray. This pattern correlates with emotion ratings above $r = 0.80$. The best correlation in a single, local region was $r = 0.44$, and this estimate is over-optimistic due to post hoc selection (unlike the whole-brain pattern). Regions typically associated strongly with emotion, like the amygdala and anterior insula, were individually very weakly correlated with experience. Thus, no single region of the brain is nearly as predictive of negative emotional experience as the entire pattern (Figure 2).

Our approach to thinking about emotion stands in contrast to approaches that place heavy emphasis on single brain regions, such as the amygdala, in understanding emotion and related forms of psychopathology (e.g., Swartz, Knodt, Radtke, & Hariri, 2015). Other analyses that have identified brain patterns that distinguish instances of one emotion category from another (Kassam et al., 2013; Kragel & LaBar, 2015; Saarimäki et al., 2015; Wager et al., 2015) so far converge similarly on highly distributed predictive patterns that span systems of the brain traditionally associated with ‘cognitive,’ ‘emotional,’ and ‘motor’

functions.

A final way that we can move forward in understanding emotional brain representations with brain imaging is to compare predictive patterns with one another. By identifying patterns that are predictive of two or more types of affective experience, we can obtain clues about the similarity of the underlying representations (Chikazoe, Lee, Kriegeskorte, & Anderson, 2014). There are many pitfalls to this approach, which can result in conflicting results. For example, many patterns will look similar when intense emotional experiences are compared with control conditions such as rest. But if we can identify patterns that are highly predictive of emotional experiences, there is a good chance we can say something meaningful about their similarity in the brain. Using this approach, we have found that brain patterns that track five types of negative emotional experience—physical pain, romantic rejection, vicarious pain (or pain empathy), negative emotion, and aversive taste—are all highly dissimilar from one another (Chang et al., 2015; Krishnan et al., submitted; Wager et al., 2013; Woo et al., 2014; Woo, Roy, Buhle, & Wager, 2015). Thus, there may be very little commonality among these types of negative emotional experience, though they are all salient, arousing, and aversive. Our findings suggest that no single brain system encodes negative affect, but rather that a set of interlocking (and superficially similar) circuits encode distinct types of negative experience in ways that are measurable and largely conserved across individuals. Though this work is in its early stages, it promises to help us disentangle the brain representations that underlie distinct types of affective experience.

Is there an ‘emotional brain’?

Recent empirical findings provide evidence that brain measures sensitively and specifically relate to particular categories of emotion, but they involve systems distributed across the brain, which likely serve basic information processing

functions that are not dedicated or evolved for emotion per se. Thus, there are many routes to (and many brain systems involved in) emotion, and which shades of experience and sub-processes are rooted in which brain systems remains largely unknown. There may be many neurophysiological varieties of different 'shades' of common emotions like anger, sadness, and joy, with different brain representations; and there is not one type of negative affect, but many. This 'view from the brain' challenges us to think about whether traditional, psychological models of affect grounded in phenomenology (including both basic emotion and dimensional models of affect) really carve the emotional brain at its joints.

References

- Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3), 361-372.
- Berntson, G. G., & Micco, D. J. (1976). Organization of brainstem behavioral systems. *Brain research bulletin*, 1(5), 471-483.
- Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLoS Biol*, 13(6), e1002180.
- Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature neuroscience*.
- Grau, J. W. (2014). Learning from the spinal cord: how the study of spinal cord plasticity informs our view of learning. *Neurobiol Learn Mem*, 108, 155-171.
- Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PLoS One*.
- Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social cognitive and affective neuroscience*, nsv032.
- Krishnan, A., Woo, C.-W., Chang, L. J., Ruzic, L., Gu, X., M., L.-S., . . . Wager, T. D. (submitted). Somatic and vicarious pain are represented by dissociable multivariate brain patterns.
- MacLean, P. (1949). Psychosomatic disease and the visceral brain; recent developments bearing on the Papez theory of emotion. *Psychosom Med*, 11(6), 338-353.
- Morrison, S. E., & Salzman, C. D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *The Journal of Neuroscience*, 29(37), 11471-11483.
- Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I., Lampinen, J., Vuilleumier, P., Hari, R., . . . Nummenmaa, L. (2015). Discrete Neural Signatures of Basic Emotions. *Cerebral cortex (New York, NY: 1991)*.
- Saper, C. B. (2002). The central autonomic nervous system: conscious visceral perception and autonomic pattern generation. *Annu Rev Neurosci*, 25(1), 433-469.
- Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9), 657-661.
- Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, 110(1), 70-75.
- Swartz, J. R., Knodt, A. R., Radtke, S. R., & Hariri, A. R. (2015). A Neural Biomarker of Psychological Vulnerability to Future Life Stress. *Neuron*, 85(3), 505-511.
- Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, 22(12), 2864-2885.

- Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388-1397.
- Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., & Barrett, L. F. (2015). A bayesian model of category-specific emotional brain responses. *PLoS Comput Biol*, *11*(4), e1004066. doi: 10.1371/journal.pcbi.1004066
- Woo, C.-W., Koban, L., Kross, E., Lindquist, M. A., Banich, M. T., Ruzic, L., . . . Wager, T. D. (2014). Separate neural representations for physical pain and social rejection. *Nature communications*, *5*.
- Woo, C.-W., Roy, M., Buhle, J. T., & Wager, T. D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS biology*, *13*(1), e1002036.
- Xiu, J., Zhang, Q., Zhou, T., Zhou, T.-t., Chen, Y., & Hu, H. (2014). Visualizing an emotional valence map in the limbic forebrain by TAI-FISH. *Nature neuroscience*, *17*(11), 1552-1559.

Figure Captions

Figure 1. Vertically integrated systems for emotion. One way to think of the interactions across brain systems is in terms of patterns that represent different types of information, which mutually constrain one another in the control of behavior. We find it useful to divide the brain into zones associated with different classes of functions. First, brainstem centers generate patterned, ‘homeostatic’ regulatory responses to environmental challenges (e.g., thermoregulatory responses and modulation of nociception). Second, affective pattern generators in the upper brainstem, including the periaqueductal gray, and a set of regions including the hypothalamus, ventral striatum and amygdala, are organized around canonical organism-environment interactions (e.g., fighting an aggressor), which influences multiple lower-level responses in service of the behavioral state. Third, the cortex contains ‘*conceptual*’ pattern generators that governs these organism-environment interactions and shapes the brain’s responses based on increasingly complex contingencies. Conceptual pattern generators represent ‘situations’—patterns of cues, goals, and perceived internal resources—which can activate behavioral states based on appraisals of the situation and available resources. Situation concepts or ‘schemas’ can involve information associated with multiple systems, including inferences about social information (dorsomedial prefrontal cortex, dmPFC), interoceptive assessments of one’s body state (insula), expectancies (orbitofrontal cortex, OFC), and autobiographical memories and places (hippocampus, Hipp). The vmPFC is positioned to integrate these elements into a coherent schema that informs (and is informed by) patterns at other processing levels.

Figure 2. Case study in emotion representation: A negative emotion-predictive brain signature. A. The Picture-Induced Negative Affect signature (PINES), a brain ‘signature’ that predicts the intensity of self-reported negative

emotion in response to viewing aversive images. **B.** When applied to data from a new individual, the PINES response tracked the intensity of negative emotion across the range from neutral to highly aversive, and showed an increase with increasing negative emotion in 100% of the participants in our sample (N = 180). Percentages on the figure reflect the classification accuracy for telling which of two sets of images sampled from the same person was more aversive, as a function of the rated negative emotion (1-5) for each set. **C.** The percentage of variance explained by predictive models based on local regions vs. the whole brain when predicting negative emotion ratings induced by aversive pictures. These percentages are from an 'item analysis' examining the correlation between the group-average PINES response and group-average brain response across the set of images. Local regions include amygdala, anterior cingulate cortex (ACC), insula, and searchlights (sphere with 5 voxel radius). **D.** Distribution of prediction-outcome correlations from searchlights. The dashed line represents the maximum correlation when searchlights used, and the red line shows the prediction-outcome correlation with the whole brain. The maximum across the whole brain is upwardly biased due to voxel selection biases, but is still much lower ($r = 0.44$, 19% of variance explained) than the PINES response ($r = 0.85$, 72% of variance explained), indicating that no one region of the brain is sufficient to accurately predict negative emotional experience.

Figure 1

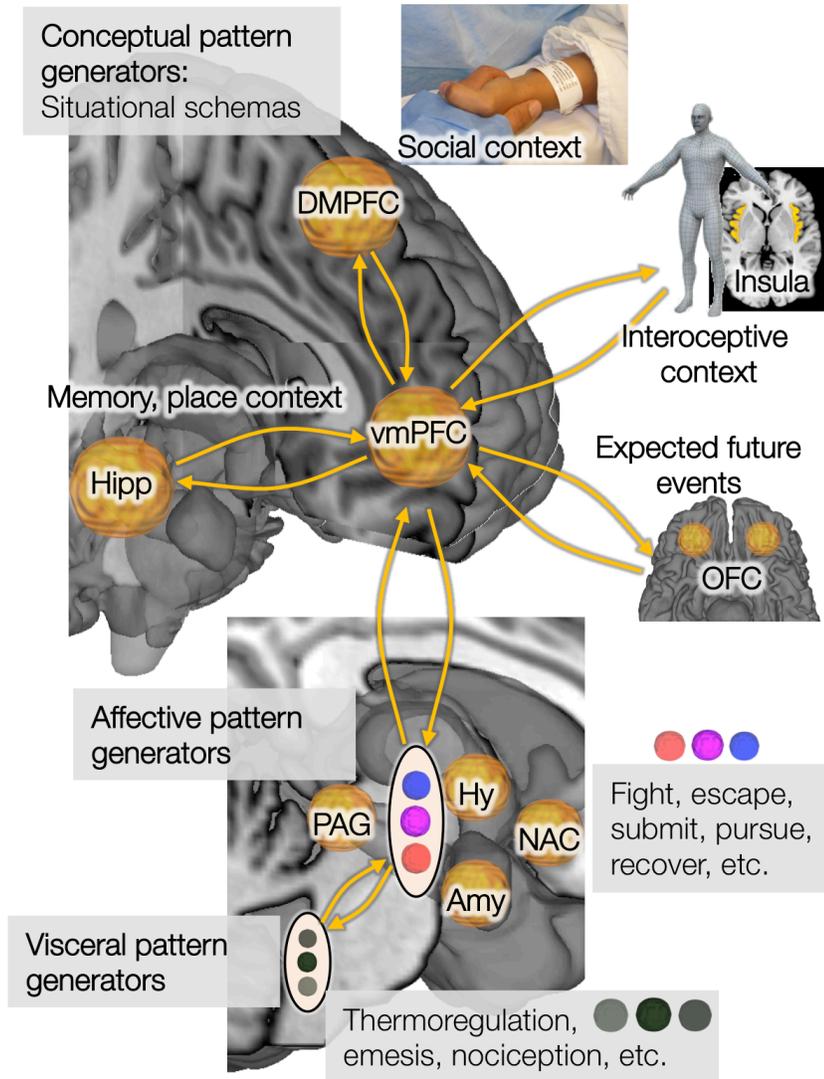


Figure 2

