Getting Warmer: Predictive Processing and the Nature of Emotion

Sam Wilkinson, George Deane, Kathryn Nave and Andy Clark

Abstract

Predictive processing accounts of neural function view the brain as a kind of prediction machine that forms models of its environment in order to anticipate the upcoming stream of sensory stimulation. These models are then continuously updated in light of incoming error signals. Predictive processing has offered a powerful new perspective on cognition, action, and perception. In this chapter we apply the insights from predictive processing to the study of emotions. The upshot is a picture of emotion as inseparable from perception and cognition, and a key feature of the embodied mind.

1. Predictive Processing and Emotion – The Story So Far

Emotion and cognition are typically thought of in contrast to one another, sitting on opposite sides of a divide between passion and reason, the hot and the cold. But what does our best theory of the brain and central nervous system (CNS) tell us about the nature of emotion?

According to an increasingly popular framework in computational neuroscience, the brain is a hierarchically arranged prediction machine (Clark (2013a)). Contrary to once-popular feedforward approaches, the brain does not simply take inputs from the outside world, process them, and pass them deeper and deeper into the processing economy. Instead, whenever information from the world impacts on your sensory surfaces, it is already, even at the earliest stages, greeted by a downward-flowing prediction on the part of your nervous system. This prediction comes from your brain's best model of what is going on in the world, and this model is constantly being updated by the mistakes it makes, by the so-called 'prediction error signal', which it constantly tries to keep to a minimum (Lee and Mumford

(2003), Rao and Ballard (1999)). In recent versions, this signal is weighted according to how reliable or salient the brain estimates the sensory information to be, relative to its best predictions. This 'precision-weighting' device operates at every level of processing. It implements attention, and allows us flexibly to balance top-down prediction and bottom-up sensory information (see Feldman and Friston (2010) and Clark (2013b))

The core business of brains like ours, if these stories are on track, is the minimization of precision-weighted errors in the prediction of sensory inputs (see Friston (2005) – and for comprehensive reviews, see Howhy (2013), Clark (2013a)). Importantly, the minimization of precision-weighted prediction error isn't always achieved by the brain updating its models of the world (which results in perception and belief). Instead it is sometimes achieved by bringing the world, usually the body, in line with the model (Feldman and Friston (2010), Clark (2016) Chapter 4). The result of this is bodily action.

According to early work in predictive processing (e.g Lee and Mumford (2003), Friston (2005)), what you *perceptually* experience is determined by the model that your brain adopts so as to best predict *exteroceptive* sensory signals such as incoming visual and auditory information. Building on this basic idea, it has recently been suggested (Seth (2013)) that what we *emotionally* experience is determined by the model that your brain adopts so as to best predict *interoceptive* signals – signals carrying information about the states of gut, viscera, hydration, vasomotor system, air-supply, muscular system, glucose and plasma levels, etc.

Here, the predictive processing (PP) account adds important dimensions to the well-known James-Lange model of emotional states as arising from the perception of our own bodily responses to external stimuli and events. The idea there, in a nutshell,

was that our emotional 'feelings' are nothing but the perceptions of our own varying physiological responses. According to James it is our interoceptive perception of the bodily changes characteristic of fear (sweating, trembling etc.) that constitutes the very feeling of fear, giving it its distinctive psychological flavor. From a subjective viewpoint, interoceptive awareness manifests as a differentiated array of feelings including those of 'pain, temperature, itch, sensual touch, muscular and visceral sensations . . . hunger, thirst, and "air hunger"' (Craig, 2003, p. 500). The feeling of fear, if James is right, is thus essentially the detection of an interoceptive physiological signature that has already been induced by exposure to the threatening situation.

A popular (and useful) way to think about James' proposal is to see it as suggesting a kind of 'subtraction test'. This is a thought experiment in which you are invited to subtract all the bodily stuff (detection of your own racing heart etc.) away from the emotional experience, and ask yourself 'what would be left?'. James' claim is that you would be left with nothing that is worth counting as an experience or emotion. What an emotion really *is*, James argument suggests, is the *self-perception of changes in our own bodily states*.

But the standard Jamesian story remains somewhat inadequate. For it seems to require a one-to-one mapping between distinct emotional states and distinctive 'brute-physiological' signatures, and it seems to suggest that whenever the physiological state is induced and detected, the same emotional feeling should arise. Neither of these implications (see Critchley, 2005) has been borne out by observation and experiment. The basic story can, however, be refined and extended by adding a 'predictive twist'. Thus Seth (2013) suggests that a neglected core component may be the match (or mismatch) between a cascading series of top-down predictions of our

own interoceptive states, and the forward-flowing information contained in sensory prediction error. Our interoceptive predictions, this story suggests:

> arise from multiple hierarchical levels, with higher levels integrating interoceptive, proprioceptive, and exteroceptive cues in formulating descending predictions. (Seth (2013) p.567.)

A single inferential process here integrates all these sources of information, generating a context-reflecting amalgam that is experienced as emotion. Felt emotions thus integrate basic information (e.g. about bodily arousal) with higher-level predictions of probable causes and preparations for possible actions. In this way:

> The close interplay between interoceptive and exteroceptive inference implies that emotional responses are inevitably shaped by cognitive and exteroceptive context, and that perceptual scenes that evoke interoceptive predictions will always be affectively coloured. (Seth (2013) p. 563)

Physiologically, the Anterior Insular Cortex is remarkably well-positioned to play a major role in such a process by encoding what Craig (2003, p. 500) describes as ‘a meta-representation of the primary interoceptive activity’. Emotion and subjective feeling states arise, this story suggests, as the result of multilevel inferences that combine sensory (interoceptive, proprioceptive, and exteroceptive) signals with top-down predictions to generate a sense of how things are for us and of what we might be about to do. Such a sense of ‘action-ready being’ encompasses our background physiological condition, estimations of current potentials for action, and

the perceived state of the wider world. This delivers a grip upon both the nature and the significance our own embodied state.

Importantly, such a grip must integrate basic information (e.g. about bodily arousal) with higher-level predictions of probable causes. This provides a very natural way of accommodating large bodies of experimental results showing that the character of our emotional experience depends both on the interoception of brute bodily signals and higher-level 'cognitive appraisals' (see Schacter and Singer (1962), Prinz (2004)). An example of a brute bodily signal is generic arousal as induced by – to take the classic example from Schacter and Singer – an injection of adrenaline. Such brute signals combine with contextually-induced 'cognitive appraisals' leading us to interpret the very same bodily 'evidence' as either elation, anger, or lust according to our framing expectations.

2. Emotions as "constructs" (models)

The account of emotion just sketched fits perfectly with the *theory of constructed emotion* (Barrett, 2017). This mechanises Barrett's preceding *conceptual act theory* (Barrett, 2014) within a PP framework. The central claim is that in each waking moment the brain is integrating past experience to generate concepts to guide actions and give meaning to sensations. When the generated concepts involved relate to physiological imperatives, your brain constructs instances of emotion.

Following from the accounts of emotion in the PP literature, each instance of an emotion arises as a categorisation of bodily signals, according to context, in terms of past experiences:

> 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). (Barrett, 2017, p.9)

The theory of constructed emotion makes a sharp distinction between emotion *instances,* and emotion *categories.* An emotion *instance* is the in-the-moment construction of an emotion given the current context. What we usually describe as an emotion, (e.g. fear) is better described as an emotion 'category', which unifies diverse and highly variable instances under a single classificatory umbrella (Clark-Polner, Johnson & Barrett, 2016). Emotion categories, Barrett argues, do not exist in nature – they are assigned according to functional and socially constructed roles. Motivation for this view comes from what has been dubbed the "emotion paradox" (Barrett, 2006). The emotion paradox refers to the fact that while the existence of emotions such as "sadness", "anger", "happiness" is assumed by the scientific community and supported by common sense, the empirical literature calls into question this assumption due to the absence of any signature – be it a facial expression, physiological response or neural activity - that reliably indexes *any* emotion category. This leads to Barrett's claim that emotion categories are collections of diverse instances that are clumped together in terms of their functional role, lacking dedicated facial expressions, physiological responses or neural signatures, Barrett states:

> 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, 2017, p.13).

This many-to-one mapping of physical states to emotion categories - called 'degeneracy' - is the primary argument behind the lack of any kind of emotional "essence". Degeneracy is borne out by the empirical literature. A meta-analysis of facial expressions indicates that many different facial expressions can be observed for the same category, and many different emotional categories can be understood by the same facial expression (Duran et al, 2017) – the meaning of a facial expression largely depends on context. Physiological signatures for any emotion category have proved to be similarly elusive, with a recent meta-analysis (Siegel et al, 2018) showing that there are no physiological signatures that reliably correspond to any one emotion category – for instance, when you're angry, your blood pressure can go up, down, or remain the same. On Barrett's view the determining factor is what kind of *action* the brain is preparing the body for – getting ready to fight requires recruitment of different resources than some other anger-related course of action, despite the emotion categorisation ('anger') being the same (Barrett, 2017). Similarly, a meta-analysis on the neurophysiological basis of emotion categories are not contained within any one brain region or system, but are represented as configurations across multiple brain networks (Wager et al, 2015).

From the perspective of evolution, degeneracy in the brain makes sense as an adaptive engineering principle. A key result of degeneracy is that a single brain can create a vast number of spatiotemporal patterns. These high complexity systems are preferred by natural selection as they can as they can reconfigure themselves into a multitude of different states (Whitacre, 2010; Whitacre and Bender, 2010). This

reconfiguration ability is what makes our brains, on this account, radically flexible according to culture and environment.

Emotions, then, are not reactions to the world, not even *interoceptively informed* reactions to the world. Rather, they are out-and-out constructions of the world. Emotions are constructed in just the same way that percepts are constructed; that is, they are predictive models of the likely causes of the sensory input, made by re-stitching together past experiences and then classifying the current experience as an amalgam of past experiences of a similar nature. These emotional predictions are made always in the service of regulating the body's internal milieu, that is, in the service of *allostasis (*Barrett and Simmons, 2015; Barrett, 2017) Predictive processing, Barrett suggests, provides the mechanism underlying these categorisations.

On this more 'action-oriented' predictive processing account, the top-down flow of predictions anticipate 1) upcoming interoceptive and exteroceptive signals and 2) the best action or bodily response to deal with the upcoming sensory flow. In order to create these 'concepts' (embodied, whole-brain representations), the brain creates predictions by using past experience to answer *"What is this new sensory input most similar to?"* (Barrett, 2017). The incoming sensory evidence, in the form of prediction error, helps to select and shape the distributions of predictions that are activated that best fit the sensory array, thereby minimising prediction error – resulting in a *categorisation* of the incoming sensory information in terms of past experiences (Barrett, 2006). That means that the predictions activated in the present are an instance of what Barsalou refers to as 'ad hoc' concepts (Barsalou, 1983). In the brain, a concept looks like a distributed pattern of activity across populations. These ad hoc concepts or predictions, that categorize present sensory flux in terms of

past experience, are the mechanism of construction of any given instance of emotion. This predictive cascade – the interpretation of the sensory flux in terms of its expected utility to allostasis - is the process of meaning-making in the brain.

Notice also that emotion and cognition are here performed in exactly the same way, that is, in reference to allostasis, and sensory inputs (prediction error) are used as information to guide the sculpting of concepts that engender adaptive action. This process is an approximation of Bayesian inference (Deneve, 2008) to decide amongst which simulation (interlocked web of predictions) should be implemented in order to maximise allostatic efficiency across multiple body systems (e.g. need for glucose, oxygen, salt etc.), and activate appropriate metabolic expenditure in the service of action (tiger, run!).

Barrett's theory is supplemented with a compelling neurobiological implementation story, where the default mode network represents efficient, multimodal summaries, which, when activated, cascade through the entire cortical sheet, terminating in primary sensory and motor regions. The cascade as a whole is an instance of a concept, or an emotion (Barrett, 2017). That said, the link between the neurobiology and the conceptual argument is not altogether clear: the empirical evidence is open to interpretation and amenable to other conceptual theories of emotion (including other conceptual theories with PP as the underlying mechanism).

The theory of constructed emotion offers a plausible account of how diverse instances of emotion come to be placed together under unifying conceptual umbrellas. It also fleshes out how emotion categories are cleaved apart according to context, and how the categories are more socially determined conceptual categories than categories existing in nature. Furthermore, the theory partially fleshes out the conception of emotion as interoceptive inference, both with a more specific mechanism of diverse

instances of emotion, and in setting out how different emotion categories come to be formed.

3. From Embodied Emotion to Embodied Valence

So how do we make sense of affective value or valence? What determines the evaluative dimension of an emotion instance? Here is an initial approach we might take to accounting for valence in terms of the properties of an action-oriented predictive processing system.

The core imperative of a predictive processor is the successful prediction of incoming sensory evidence. Thus it may initially seem that the successful minimisation of prediction error should be what determines an overall state of positive valence. Though this may seem promising at first, such a proposal quickly falls apart. Any account of valence that is state-based, that equates positive valence to a state of minimized prediction error, fails to do justice to the fact that prediction error minimisation is necessarily a dynamic and continuous process, constantly engaging action, and designed to account for the on-going maintenance of an organism in an ever changing world. Only from this perspective can we avoid the 'dark room' objection to predictive processing (Friston, Thornton, and Clark, 2012). This states that if my goal is solely the minimisation of prediction error, then surely I should just seek out a dark, empty room and stay there. Perfect prediction, it seems, is attainable by avoiding action and practicing sensory (and nutritional) deprivation until death. Such a policy is, of course, wholly inconsistent with the actual behaviour of living things.

An initial response to this might be that the various demands of survival (as ultimately signalled in the form of prediction error) would move you onwards. But note that even were your dark room to come equipped with a life support machine (consider an unending night in an abandoned hospital ward) it is unlikely that you would find this to be an endlessly pleasurable experience. Humans not only find a lack of novel stimulation boring, they actively seek out and take delight in a rich repertoire of aesthetic, humorous, or thrilling situations, from skydiving to stand-up, that are specifically engineered to generate a rush of prediction error through the violation of prior expectations.

A more promising strategy is as follows. Instead of tying valence to the achievement of some particular error-minimized state, Joffily and Coricelli (2013) propose a dynamic alternative in which valence is taken to be the *rate* at which this error is being reduced. In mathematical terms valence is recast as the first time-derivative of error: a matter of *velocity*, rather than position. We seek out surprising states, then, in as much as they offer us the opportunity to engage in a faster (rather than slower) rate of reduction in prediction error. Drawing on Carver and Scheier's (1990) control theoretic account of emotion, Van De Cruys (2017) improves and extends this story by suggesting that, rather than being straightforwardly a matter of a positive rate of error reduction, pleasure (positive valence) occurs when our actual rate of error reduction is higher than we had predicted it would be. If it is lower, we experience negative valence.

An upshot of explaining valence in terms of these processing characteristics, rather than specific content, is that it is no longer tied to any particular set of causes, error modality, or level of inference. We can thus describe a relationship between allostasis and valence that is not constrained (as it is in Seth (2013)) to inference over

the causes of interoceptive signals alone. This seems like the correct route to take. Homeostatic maintenance is served not only by the direct monitoring and regulation of physiological variables, but also indirectly, by the anticipatory regulation of our external environment. Whether intero or extero-ceptive, persistent unreduced prediction error is a sign that we do not have a grip on our self or surroundings, and adjustments need to be made.

Furthermore, tying valence to the regulation of exteroceptive error reduction rate allows us to characterise more 'cognitive' experiences of positive valence – those that are not easily described in terms of basic physiological reactions - such as responses to art, narrative or humour. These can now be understood as achieving their emotional effects by engineering pleasurable trajectories in the creation and violation of expectations, followed by the subsequent pleasurable release in the eventual reduction of resulting prediction error. This fits nicely with descriptions of humour, as resulting from the creation and resolution of tension (Sroufe and Waters, 1976) and, as Van De Cruys and Wagemans (2011) suggest, provides a potential explanation of the failure of aesthetic principles (such as harmony, fluency, or balance) to account for the success of celebrated works of art which regularly display the deliberate violation of such rules.

4. Emotion and Cognition

Summing up the previous sections, what predictive processing reveals is a world permeated by affect – a world of opportunities for action, geared to current tasks, modulated by information about our own bodily states. But to see just how radical the PP picture turns out to be, we still need to add one final ingredient. It's that

PP rejects the picture of emotion and cognition as fundamentally different kinds - at least insofar as they are causally active parts of the cognitive machinery.

According to a popular view, often associated with Hume, a fundamental divide among all things mental is one that divides the informational and the motivational. The former is about the organism ("coldly") coming to a view about what's going on in the world, whereas the latter is about ("hotly") driving the organism to bring about change in the world. Hume's central point was that without the latter, without passions, no action would ever take place. A hypothetical creature only capable of having informational states would stay still, inert and unmoved to do anything, regardless of what it learnt about the world. In this sense, according to Hume, emotion (affect, passion) broadly construed, is the driving force behind all action, but completely distinct from belief (and insulated from "reason").

The idea that informational states on the one hand, and motivational states on the other, are fundamentally different kinds of state whose interaction is required to bring about action, is widely embraced in daily life. It forms not only a core part of common-sense (or 'folk') psychology, but is deeply embedded in some more scientific frameworks too. Statistical decision-theory (including neuroeconomics and work on reinforcement learning) inherits this Humean picture, since in standard realizations it works with a firm separation between encodings of value or 'utility' and encodings of probability. In these frameworks, decisions are made and actions selected only when utility and probability align, revealing viable opportunities for worldly interventions that deliver weighted rewards at calculated costs (for a useful review, see Sanfey et al (2006)).

By contrast, PP posits only predictions, informed by multiple inner and outer sources of information. In PP motivational states are realized as predictions about our

own future actions and states. To see how, let's return to the PP treatment of action. Action is making the world conform to some of your predictions, and is just another way of reducing long-term prediction error. At the bottom level, PP makes sensori-muscular (proprioceptive) prediction into a proxy for motor commands (Shipp, Adams and Friston (2013)). Predicting the flow of sensori-muscular effects that would occur if you hit the tennis ball just right actually brings the 'good hitting' about. In a little more detail, the brain predicts the flow of states of muscle spindles, tendons, and joints that the action demands, and the resulting errors (since those states are not yet actual) are systematically quashed by moving the body so as to make that flow of prediction come true. This is an elegant and economical means of delivering basic motor control (see e.g. Shipp et al. (2013)).

PP deploys the same kind of story 'all the way up'. Our action-guiding proprioceptive predictions are themselves caused by even higher-level and longer time-scale predictions – predictions about our own future behaviors and future states. These entrain actions when good opportunities arise (see Pezzulo, Rigoli and Friston (2015)). The picture is of nested beliefs that entrain actions by bringing about predicted sensory flows. For example, suppose I believe/predict that I will meet you at 7:00 at the movie-theatre. This (combined with prior knowledge and any newly gleaned information) leads me to believe/predict that I will get the 6:30 bus. That last prediction then acts as a kind of mini-policy that enslaves motor action (by means of proprioceptive predictions) when it is time to leave the house.

Simple action-entraining motor intentions here cash out as precise proprioceptive predictions, while higher-level intentions, including standing goals, are realized by higher-level predictions of whole swathes of sensory information, which likewise entrain actions (by yielding precise proprioceptive predictions) when they

themselves are assigned high enough precision. These nested, interacting predictions arise and dissolve – in ways that realize the phenomenological flux of shifting drives and desires - as we move around the world, acting and harvesting new sensory information. If PP is on track, the causally potent play of human motivation is not an illusion - but it is realized using only the common currency of multi-level, multi-area prediction. In this picture, prior beliefs (resulting in predictions) combine with sensory evidence to bring about action. This is just the bedrock (Bayesian) move – one that turns everything into a form of prediction-based inference.

It has been suggested (Holton (2016), Klein (2016)) that this picture is too impoverished to be a satisfying story about human minds. Part of their reasoning is roughly Humean. The Humean worry is that beliefs (or predictions) without motivations are inert, unable to mandate actions. That's already taken care of by the PP story though, since high-precision predictions that have proprioceptive (hence motoric) consequences are immediately poised to entrain actions to make themselves come true. Holton also worries that assimilating desires to predictions "doesn't do justice to the multiplicity and malleability of human desire" noting that we need to accommodate cases where we desire X and may even do X while believing that X won't bring us happiness or pleasure. However, PP accommodates this very simply, by separating predictions about the hedonic consequences of actions from the full set of predictions that interactively entrain actions. Specifically, the predictive processing story firmly distinguishes (Friston, Shiner et al. (2012)) between sub-personal action-entraining high-precision predictions concerning what I will do and predictions of the hedonic (interoceptive) outcomes of those very actions. PP thus accommodates the fact, highlighted by Holton, that drug users often do not believe/predict that taking the drugs will actually lead to happiness. But what they do predict is seeking and

ingesting the drug. PP thus easily reconstructs the useful distinction between 'wanting' and 'liking' (Berridge (2007)). The PP picture thus turns out to be a neat fit with important work on the nature and mechanisms of addiction (Berridge (2007), Friston et al (2012)). More generally, even given that the addict need not predict that the drugs will bring pleasure, PP remains poised to explore a wide variety of promising accounts in which the addict's experiences and actions are the results of interacting sub-personal (non-conscious) predictions.

This replaces Hume's two interacting kinds (reason and passion) with a picture of large numbers of subtly different and modifiably interacting elements. All of those elements are somewhat belief-like (consisting in predictions) but somewhat desire-like too (as they help select and entrain actions at multiple time-scales). So, while it may look like a simplifying move, what PP finally delivers will in fact be a far richer palette for explaining human behaviour. That palette allows a full spectrum of possibilities that reach far beyond the simple, constrained interactions suggested by crude folk psychological distinctions between 'cognition' and 'conation'.

We have seen how this collapses belief and desire, and desire is often construed as a "hot" or "impassioned" state, but it is clearly a mistake to equate emotion with desire. As several theorists have noted (e.g. Marks 1982, Oakley 1992), emotion has both belief-like and desire-like elements. Experiencing fear of the spider simultaneously tells you about the world (e.g. that there is a spider there), while also motivating you to act in a certain way (run away from said spider). But whereas the standard way of thinking of emotions is as *composed of* these belief and desire-like elements, PP construed things very differently. Just because the belief and desire-like elements can be "read off" the emotional state, it is not to say that psychologically (or indeed ontologically) they are somehow the primitive building

blocks of a hybrid and less primitive state called emotion. On the contrary, according to PP, it is the emotional state, which simultaneously informs and moves, that is primitive, and, in predictive processing terms, this is all fleshed out in the common currency of predictions and predictive models: predictions generated by complex hierarchical models concerning, in an interconnected manner, the organism, the world, and the organism's place in that world.

The same point can be made in terms of direction of fit. Whereas it has been common to think of beliefs, with their mind-to-world (or descriptive) direction of fit, and desires, with their world-to-mind (or directive) direction of fit, as being the fundamental building blocks of the mind, what is actually fundamental in the PP architecture is prediction, which can vary across a spectrum as to the extent to which it should be fulfilled by the world (perception/belief) or the self (action/desire). This means that pure belief (or cold perception), or pure desire (or blind action), is actually idealization, a mere theoretical construct. What we are actually left with is a wide variety of what Millikan (1995) calls "pushmi-pullyu representations", states that simultaneously describe and direct.

5. Concluding remarks

Emotions, we have argued, are built from predictions. They reflect inner and outer sources of information, combined in flexible ways, and are answerable to the full world knowledge (generative model) of an agent. But they are not a special cognitive kind. Instead, they are part and parcel of an integrated processing system whose core functionality is to reduce precision-weighted prediction error by

maintaining dynamic engagements with the world. These engagements display trajectories both marked and determined by valence, where positive valance reflects better-than-predicted slopes of error minimization. What emerges is a picture of mind as an action-oriented system all of whose states are somewhat belief-like, and somewhat desire-like too.

Another way of looking at this is as follows. In so far as full-blown emotions as we typically understand them are the most prominent and consciously detectable (and hence categorized) of these action-oriented states, one could say that PP renders emotion, construed more broadly to include even the very subtlest of these, ever-present in cognition. In other words, the embodied predictive mind is, by necessity, an emotional mind.

Acknowledgements

All authors were supported by the European Research Council (XSPECT – DLV-692739).

Reference List

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social cognitive and affective neuroscience*.

Barrett, L. F. (2014). The Conceptual Act Theory: A Precis. *Emotion Review* 6 (4):292-297.

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

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition* 11:211-277.

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* (Berl.) 191: 391–431.

Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. Psychological review, 97(1), 19.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press USA.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (3):181-204.

Clark, A. (2013b). Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences* 36 (3):233-253

Clark-Polner, E., Johnson, T. D., and Barrett, L. F. (2016). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cereb. Cortex* 27, 1944–1948. doi: 10.1093/cercor/bhw028

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Curr. Opin. Neurobiol.* 13, 500–505. doi: 10.1016/s0959-4388(03)00090-4

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective and cognitive integration. *J. Comp. Neurol.* 493, 154–166. doi: 10.1002/cne.20749

Denève, S. (2008). Bayesian spiking neurons I: inference. *Neural Comput.* 20, 91–117. doi: 10.1162/neco.2008.20.1.91

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci*. 4:215. doi: 10.3389/fnhum.2010.00215

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston K. (2009). The free-energy principle: a rough guide to the brain? Trends Cogn. Sci. 13, 293–301

Friston K, Shiner T, FitzGerald T, Galea JM, Adams R, et al. (2012) Dopamine, Affordance and Active Inference. *PLoS Comput Biol* 8(1): e1002327

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol.* 3:130. doi: 10.3389/fpsyg.2012.00130

Gerken, L., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid 'labouring in vain'by attending more to learnable than unlearnable linguistic patterns. Developmental science, 14(5), 972-979.

Holton, R. 2016. Review of Surfing Uncertainty: Prediction, Action, and the Embodied Mind, *Times Literary Supplement* October 7, 10-11

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. PLoS computational biology, 9(6), e1003094.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PloS one, 7(5), e36399.

Klein, C. (2016). What do predictive coders want? *Synthese.*

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A.* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Marks, J. (1982). A theory of emotion. *Philosophical Studies* 42 (1):227-242.

Millikan, R. (1995). Pushmi-pullyu representations. *Philosophical Perspectives* 9:185-200.

Oakley, Justin, 1992. *Morality and the Emotions*, London: Routledge and Kegan Paul.

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(3), 902-911.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford University Press.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Sanfey, A. G., Loewenstein, G., McClure, S. M., and Cohen, J. D. (2006). Neuroeconomics: cross-currents in research on decision-making. *Trends Cogn. Sci.* 10, 108–116. doi: 10.1016/j.tics.2006.01.009

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379-399.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565-573.

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Phil. Trans. R. Soc. B*, *371*(1708), 20160007.

Shipp, S., Adams, R., & Friston, K. J. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences*, *36*(12), 706–16

Srofe, L. A., & Waters, E. (1976). The ontogenesis of smiling and laughter: a perspective on the organization of development in infancy. Psychological review, 83(3), 173.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: a tentative prediction error account of visual art. i-Perception, 2(9), 1035-1062.

Van de Cruys, S. (2017). Affective value in the predictive mind. MIND Group.