Modeling subjective belief states in computational psychiatry: Interoceptive inference as a candidate framework

Xiaosi Gu[1,2,3*], Thomas HB FitzGerald[4,5,6], Karl J. Friston[5]

1. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY
2. Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY
3. Mental Illness Research, Education, and Clinical Center (MIRECC VISN 2) at the James J. Peter Veterans Affairs Medical Center, Bronx, NY.
4. School of Psychology, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR4 7TJ, UK
5. Wellcome Trust Centre for Neuroimaging, University College London, London, England
6. Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, Russell Square House, London, WC1B 5EH, UK

Correspondence should be addressed to

Xiaosi Gu, PhD

Icahn School of Medicine at Mount Sinai, New York, NY

1 Gustave L Levy Place, Box 1230

New York, NY 10029

xiaosi.gu@mssm.edu

Abstract

The nascent field of computational psychiatry has undergone exponential growth since its inception in the early 2010s. To date, much of the published work has focused on choice behaviors, which are primarily modeled within a reinforcement learning framework. While this initial normative effort represents a milestone in psychiatry research, the reality is that many psychiatric disorders are defined by disturbances in subjective states (e.g. depression, anxiety, etc.) and associated beliefs (e.g., dysmorphophobia, paranoid ideation, etc.), which are not considered in normative models. In this paper, we present interoceptive inference as a candidate framework for modeling subjective – and associated belief – states in computational psychiatry. We first introduce the notion and significance of modeling subjective states in computational psychiatry. Next, we present the interoceptive inference framework, and in particular focus on the relationship between interoceptive inference (i.e. belief updating) and emotions. Lastly, we will use drug craving as an example of subjective states to demonstrate the feasibility of using interoceptive inference to model the psychopathology of subjective states.

Part I. Modeling subjective belief states in computational psychiatry research: significance and challenges

Computational psychiatry is a new and interdisciplinary field that seeks to understand the mechanisms underlying mental function and dysfunction using computational approaches (Friston et al. 2014; Maia and Frank 2011; Montague et al. 2012). This nascent field has enjoyed much success since its inception, providing powerful explanations and predictions for a wide range of disorders; including schizophrenia and psychosis (Adams et al. 2013b; Braver et al. 1999; Powers et al. 2017), depression (Rutledge et al. 2017), bipolar disorder (Mason et al. 2017), anxiety (Browning et al. 2015), autism (Lawson et al. 2017; Lawson et al. 2014), ADHD (Hauser et al. 2016; Hauser et al. 2014; Hauser et al. 2017b), OCD (Gillan et al. 2016; Hauser et al. 2017a), and addiction (Fiore et al. 2018; Gu 2018; Gu and Filbey 2017; Redish and Johnson 2007). The majority of these studies and theories have focused on overt behaviors, and in particular choice behaviors, using computational models of learning and decision-making. A handful of studies constitute first efforts to model subjective feelings using a decision theoretic approach [i.e. consider subjective states as corollaries of choice behaviors and decision processes; see (Rutledge et al. 2017) for an example and (Bach and Dayan 2017) for a review]. However, there is a substantial body of computational work that models subjective data directly in other domains of cognitive neuroscience – such as metacognition (Allen et al. 2016; Fleming and Daw 2017) and pain research (Jepma et al. 2018; Wiech 2016) – all of which could be crucial for modeling subjective states in computational psychiatry.

As Helen Keller famously said, "*The best and most beautiful things in the world cannot be seen nor even touched, but just felt with the heart.*" (Keller 1881). The subjective nature of internal feelings and beliefs presents several major challenges to the scientific investigation of the mental processes that underwrite them – and the definitions of subjective states are often loose and inconsistent. For instance, the definition of emotion differs drastically between different theories, disciplines, and studies (Bach and Dayan 2017; Gu et al. 2013; James 1884; Lange and James 1922; Lazarus 1991; LeDoux 2000; Papez 1937; Pessoa and Adolphs 2010; Phelps 2006; Picard et al. 2001; Schachter and Singer 1962). Here, we define subjective states as an agent's beliefs about the world – either external or internal (see Part II for more detailed discussions). Second, subjective states are difficult to measure. Self-reports are probably the more direct and closest to one's true internal feelings; yet these reports could sometimes be partially obscured by the individual's ability to introspect and access their internal world (e.g. awareness or metacognition). On the other hand, objective measures, such as functional magnetic resonance imaging (fMRI) or electroencephalography (EEG), often provide signals that are not specific to distinct subjective states. For example, there is a considerable overlap between brain activations related to different emotions such as fear, disgust, and anger [see (Fusar-Poli et al. 2009; Murphy et al. 2003; Phan et al. 2002) for meta-analyses]. Lastly, existing analyses of subjective states are largely correlational and lack mechanistic or computational explanations. One might argue that these challenges must be overcome if affective neuroscience and computational psychiatry are to move forward.

Aberrant subjective states are prevalent across almost all disorders. Disturbances in emotional feelings, for example, are hallmarks of mood disorders. While some subjective states can affect overt behaviors and related brain responses – and can thus be assessed indirectly through these choice behavior-related measures (Chiu and Deldin 2007; Huys et al. 2015) – many might not directly affect nor be inferred using choice behaviors (Chung et al. 2017; Rutledge et al. 2017). Take depression as an example. The literature suggests that depression affects many overt behaviors (and related neural circuits) such as error detection

(Chiu and Deldin 2007; Pizzagalli et al. 2006), processing stimuli related to the self (Lemogne et al. 2009), accuracy in a reward reversal task (Robinson et al. 2012), and attention (Tian et al. 2016). However, two recent studies – using computational modeling and monetary decision-making tasks – demonstrate that depression did not affect reward sensitivity (Chung et al. 2017) or reward prediction error encoding (Rutledge et al. 2017). These studies represent some of the first efforts to formally quantify depression-related behaviors with computational modeling (i.e., reinforcement learning models). Nevertheless, the null findings from both studies could indicate that a decision-theoretic neuroscience approach alone may not help us disclose the complex nature of subjective feelings, and that direct modeling of subjective data and related neural and physiological concomitants might be necessary.

Part II. Interoceptive inference as a candidate framework

Here, we introduce a framework of interoceptive inference that has recently emerged from the computational neuroscience literature to explicitly account for subjective states. As it is typically described (Seth and Friston 2016), interoceptive inference makes two related but dissociable claims. The first claim is that approximate Bayesian inference about physiological states of the body underlies feeling states [see (Barrett and Simmons 2015; Gu and FitzGerald 2014; Gu et al. 2013; Pezzulo 2014; Pezzulo et al. 2015; Seth 2013; Seth 2014; Seth and Friston 2016) for a sample; also see a special issue in Phil. Trans. R. Soc. B (Tsakiris and Critchley 2016)]. This entails an important claim that 'feelings' are a certain kind of belief states that are updated during inference to provide the best explanation for interoceptive sensations (e.g., "I am anxious" is the best explanation current bodily sensations). The second claim is that physiological states, generating interoceptive sensations, are controlled using active inference (Seth and Friston 2016). In other words, autonomic reflexes work to align internal states with descending predictions from an agent generative model of herself. This is precisely analogous to active inference as a tool for controlling the motor system (Adams et al. 2013a). Taken together, interoceptive inference, with these two major claims, reconciles the conflict between the James-Lange and the Canon-Bard theories in the sense that they are both right – that subjective feeling or belief states are both cause and consequence of autonomic states.

Understanding mental states as Bayesian inference has a number of compelling aspects. In particular, it prescribes 'first principle' strategies for integrating existing beliefs with new information and for dealing with uncertainty; problems that are of key importance for enabling complex, adaptive behavior. It also naturally leads to mechanistic hypotheses about how such processes might be implemented (Aitchison and Lengyel 2016; Friston 2005; Friston et al. 2017; Ma et al. 2006). This approach has been successful in accounting for a wide range of cognitive and neural phenomena; such as perception, learning, and memory [see (Doya 2007; Friston 2010; Knill and Pouget 2004; Moutoussis et al. 2014) for a sample of papers]. With the rise of the new field of computational psychiatry, Bayesian approaches have also proved to be successful in explaining aberrant perception and cognition in many psychiatric disorders such as schizophrenia (Moutoussis et al. 2011), autism (Pellicano and Burr 2012), and addiction (Gu and Filbey 2017; Schwartenbeck et al. 2015). In contrast to the fruitful results of Bayesian formulations in other areas of neuroscience and psychology, however, its application to subjective (feeling) states is rare.

In this article, we aim to describe how emotions – either basic (e.g. joy, fear) or social (e.g. shame, guilt) – can be formulated in the setting of interoceptive inference. Specifically, we consider emotions as Bayesian beliefs about interoceptive states. The mismatch between one’s prediction and the actual interoceptive signal contributes to an “interoceptive prediction error” signal. Agents seek to minimize this prediction error and can do so in one of two ways. First, they can alter their beliefs about internal states (perception). Second, they can alter their internal states to fit their beliefs (action, under active inference). This corresponds to the two aspects of interoceptive inference described above. If true, this model resolves a longstanding debate about the relationship between bodily signals and emotions, the center of debate between James–Lange and Cannon–Bard theories.

Specifically, William James (1842-1910) and Carl Lange (1834-1900) proposed that bodily responses, including physiological responses of the muscles, skin, and viscera, precede conscious emotional feelings (e.g. “my heart races therefore I perceive fear”) (James 1884; Lange and James 1922). In other words, emotions are the result of changing bodily states. On the contrary, Walter Cannon (1871-1977) and his student Philip Bard (1898-1977) proposed that physiological responses of the body and subjective experiences of emotion can occur independently from each other (Bard 1928; Cannon 1927). This view was developed based on experimental findings contradicting the James-Lange theory. For example, physiological responses of the body could manifest more slowly than human subjects’ reports of changed feelings (Bard 1928; Cannon 1927). While these early theories tried to determine a unidirectional relationship between bodily responses and subjective experience, under interoceptive inference, both processes are in play concurrently, suggesting that the brain-body relationship is a two-way street (Barrett and Simmons 2015; Gu and FitzGerald 2014; Ondobaka et al. 2017; Seth and Friston 2016). This explains the *prima facie* plausibility, but also incompleteness, of these two classic approaches to understanding emotion.

Cognitive labeling theories of emotion, first developed in the 1960s and 1970s, offered a more nuanced view of emotions and bodily states (Dutton and Aron 1974; Schachter 1964;

Schachter and Singer 1962; Slochower 1976; Valins 1966). Despite some minor differences, these theories, in essence, all suggest that emotions do not solely depend on changes in bodily states; instead, cognitive processes are involved such that the same bodily responses can be interpreted in different ways depending on context. For example, increased heart rate could be felt as excitement in the context of a positive event (e.g., job promotion) but fear in a threatening situation (e.g., robbery). In the Capilano Suspension Bridge experiment – a notably study that examined sexual arousal and romantic relationships (Dutton and Aron 1974) – heterosexual male participants were approached by an attractive female on either a fear-arousing suspension bridge or a non-fear-arousing bridge and were asked to complete questionnaires related to the Thematic Apperception Test pictures (that lacked obvious sexual content). The sexual content of reports and attempts to contact the female experimenter post-experiment with greater in the suspension bridge condition. These effects were not observed if the (male) participants were approached by a male. Findings from this study were interpreted as evidence for cognitive labeling theory – that bodily arousal was labeled by people in the context sensitive fashion. Here, we propose that contextual cues and autonomic signals are assimilated as sensory evidence during interoceptive inference – to produce a posterior belief that best explains both; namely, a cognitive label or explanation. In this sense, interoceptive inference can be viewed as a formal (Bayesian) description of cognitive labeling. On this reading, interoceptive inference may provide an account of the mental and neuronal belief updating that underwrites cognitive labeling.

From the perspective of computational psychiatry, interoceptive inference is especially prescient as it provides a formal, quantitative approach (at least in principle) to understanding subjective feeling states. Existing computational work looking at affective disorders typically appeals to the framework of reinforcement learning and accompanying claims about dopaminergic function (Huys et al. 2015; Nestler and Carlezon 2006). However, while there is likely to be a close relationship between reward and certain aspects of affective states (Huys et al. 2015; Rutledge et al. 2017; Whitton et al. 2015), affective and motivational states (or 'liking' and 'wanting') are known to be dissociated at both behavioral and neurobiological levels (Berridge 2012; Robinson and Berridge 1993). Crucially, subjective (feeling) states have attributes that cannot be captured by a simple scalar reward. For these reasons, there is a natural need to develop and test formal accounts of interoceptive inference – and how it may relate to emotion in both normal and pathological cognition.

Interoceptive inference is a young field, and has undergone substantial theoretical developments over the past few years (Barrett and Simmons 2015; Gu and FitzGerald 2014; Ondobaka et al. 2017; Owens et al. 2018; Petzschner et al. 2017; Pezzulo et al. 2015; Seth 2013). These advances include the delineation of computational blocks involved in interoception (Petzschner et al. 2017), its relationship to self-perception (Seth 2013), emotion (Barrett and Simmons 2015; Gu et al. 2013), and embodied cognition and decision-making (Allen and Friston 2018; Gu and FitzGerald 2014), and importantly, its relevance to psychosomatic and psychiatric disorders (Owens et al. 2018; Petzschner et al. 2017). As such, empirical evidence that directly supports the predictions of this account is still limited (see (Allen et al. 2016; Gentsch et al. 2019; Jepma et al. 2018) for a few examples), but one might anticipate empirical results that endorse or refute its utility, over the next few years. One clear prediction is that emotional impairments should exist in people with deficits in the processing of bodily signals (i.e., due to imprecise sensory data) and emotional function could improve in individuals with better interoceptive accuracy; i.e., due to more precise sensory data; as reported in (Barrett et al. 2004). Relatedly, this model also predicts that strategies – that can help individuals recognize or infer explicitly (i.e., mentalize) the causes of bodily signals correctly – should also improve affective processing and responses. For example, psychotherapy that can re-interpret the emotional content and meaning of increased heart rate may help individuals with panic attacks.

There has been one fMRI study that directly applied Bayesian modeling to quantify subjective feelings (Xiang et al. 2013). Participants were asked to rate how happy they felt about a monetary offer they received while being scanned. Using an ideal Bayesian observer model, the authors found that feeling prediction errors were encoded in the nucleus accumbens, vmPFC, and posterior cingulate, whereas feeling variance prediction errors were computed in bilateral anterior insula. To date, this remains still the first and only fMRI study to use the Bayesian framework to model subjective feelings directly.

A recent study using EEG also provided direct evidence supporting the interoceptive inference hypothesis (Gentsch et al. 2019). In this study, participants viewed either neutral or angry faces followed by either the same (repetition; e.g. neutral-neutral, angry-angry) or different (alternation; e.g. neutral-angry, angry-neutral) emotional faces. Interoceptive responses were measured with heartbeat-evoked potential (HEP), which is the EEG response time-locked to the electrocardiogram (ECG) R-peak of the heartbeat. The authors found that angry faces, preceded by angry faces, elicited attenuated HEPs compared to those proceeded by neutral faces. In other words, the affective prediction of upcoming stimuli modulated interoceptive responses to those stimuli. Such differences between predicted and unpredicted emotional stimuli were not found for neutral stimuli; suggesting that the top-down predictions of interoception are based on bottom-up interoceptive responses. Taken together, these results shed light onto the intricate interactions between interoceptive predictions and inputs and provide initial support for the interoceptive inference framework. More empirical studies are needed to echo such efforts.

Part III. Drug craving as an example

In this last section, we consider a recent computational theory of drug craving, as an example of how interoceptive inference can be used to account for subjective states in particular disorders. Craving refers to an intense, urgent, or abnormal desire or longing, according to the Merriam-Webster dictionary. Naturally, this definition suggests that craving is a subjective state and is much more difficult to measure objectively and quantitatively. Clinically, craving directly relates to the subjective well-being and life quality of the individual, and can be dissociated from drug-related choice behaviors (Tiffany et al. 2012; Tiffany and Wray 2012). In humans laboratory studies, craving has been extensively studied using cue-exposure paradigms [see (Chase et al. 2011; Engelmann et al. 2012; Jasinska et al. 2014; Tang et al. 2012; Yalachkov et al. 2012) for reviews and meta-analyses]. However, it remains controversial as to what psychological processes are actually elicited by these paradigms and how they relate to real-life craving (Shiffman et al. 2015). For instance, drug cues are inherently valuable to addicted individuals and could thus induce reward processing, along with craving. Cue-elicited response studies typically contrast brain activities elicited by drug cues with those induced by non-drug cues (e.g. cigarette vs. pencil) and have reported widespread activations in dopaminergic and limbic regions including the midbrain (ventral tegmental area, VTA), ventral striatum, insula, anterior cingulate cortex (ACC), ventromedial prefrontal cortex (vmPFC), amygdala etc. (Chase et al. 2011; Engelmann et al. 2012; Jasinska et al. 2014; Tang et al. 2012; Yalachkov et al. 2012). Many of these regions are involved in value encoding (Rangel et al. 2008; Rushworth and Behrens 2008). It is thus difficult to tease apart the neural mechanisms underlying subjective craving vs. those supporting value encoding in cue-exposure paradigms.

We recently proposed the first Bayesian model of craving (Gu and Filbey 2017), building upon the interoceptive framework (Barrett and Simmons 2015; Gu and FitzGerald 2014; Gu et al. 2013; Seth and Friston 2016). Specifically, we consider craving as a posterior belief about bodily states associated with the availability of addictive substances (Gu and Filbey 2017). This inference model of drug craving has proven effective in accounting for several important experimental findings that are not explained by purely pharmacological models of drug addiction. In the human addiction literature, for example, it has been shown that craving depends on not only the availability of the addictive substance in the body, but also people's beliefs about the presence of that substance (Gu et al. 2015; Juliano et al. 2011; Kelemen and Kaighobadi 2007; McBride et al. 2006). In addition to the well-known placebo effects where people respond to 'sugar pills' without active substances (McCusker and Brown 1990; Robinson and Berridge 1993), recent studies have demonstrated that – in nicotine addiction – craving was reduced (Gu et al. 2016) and ventral striatal responded to value (Gu et al. 2015) only when smokers had a nicotine cigarette and also believed there was nicotine in the cigarette, but not when they thought otherwise. These findings are complementary to previous studies reporting placebo effects; in the sense that both sets of results point to the importance of considering both prior beliefs (i.e., expect drug or no drug) and sensory evidence (drug delivered or not). These results are also interesting, in the sense that placebo effects primarily describe how beliefs can 'up' responses in the absence of active substances, while the Gu et al studies demonstrate how prior expectations can also 'down-regulate' behavioral and neuronal responses in the presence of active substances, such as nicotine. Using an interoceptive inference framework, we were able to simulate these findings (Gu et al. 2016) by systematically manipulating prior beliefs (e.g., whether the smoker expected to receive a cigarette with nicotine or placebo cigarette) and the likelihood of drug administration (e.g. whether the cigarette has nicotine or not). This work (Gu and Filbey 2017) suggests that the posterior beliefs – that undergird craving – are ameliorated when a drug is administered and when the individual expects the drug, compared to the other three alternatives (i.e. expect drug & drug absent, expect no drug & drug present, expect no drug & drug absent), corroborating empirical findings (Gu et al. 2015; Juliano et al. 2011; Kelemen and Kaighobadi 2007; McBride et al. 2006).

Incubation of craving, which refers to the effect that craving increases rather than decreases during early abstinence, is another important finding that remained unexplained by any previous computational framework. In one recent paper, we used the same model to further simulate previous experimental findings (Bedi et al. 2011; Conrad et al. 2008; Grimm et al. 2001; Lu et al. 2005; Parvaz et al. 2016) that craving could increase over time during early abstinence (Gu 2018) . Taken together, this Bayesian framework has been proven powerful in accounting for craving in addiction, and provides a more general illustration of the potential for interoceptive inference models in building novel and formal hypotheses about the processes underlying psychiatric disorders.

Conclusion

In this short paper, we offer a viewpoint that interoceptive inference can serve as a candidate framework for modeling subjective states in computational psychiatry research. While we emphasize the importance of modeling subjective data directly [e.g. following similar practice in other areas of cognitive neuroscience (Allen et al. 2016; Fleming and Daw 2017;

Jepma et al. 2018; Wiech 2016)], this view does not diminish the significance of modelling overt behaviors in terms of their subjective causes. Instead, we suggest that subjective experience can – and should be – modelled directly, and the interoceptive inference framework is a natural fit for this purpose. For example, in depression research, models of sad feelings and models of aberrant learning and decision-making are both necessary in order to achieve a comprehensive view of depression. More empirical studies will be needed to directly examine the validity of the interoceptive inference framework in both health and disease.

Acknowledgments: XG is supported by the National Institute on Drug Abuse (grant 1R01DA043695) and the Mental Illness Research, Education, and Clinical Center (MIRECC VISN 2) at the James J. Peter Veterans Affairs Medical Center, Bronx, NY. THBF is supported by a European Research Council (ERC) Starting Grant under the Horizon 2020 program (Grant Agreement 804701). KF is a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z).

Declaration of interest: The authors declare there are no conflicts of interest.

Reference

Adams RA, Shipp S, Friston KJ (2013a) Predictions not commands: active inference in the motor system. Brain Struct Funct 218: 611-43.

Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ (2013b) The computational anatomy of psychosis. Front Psychiatry 4: 47.

Aitchison L, Lengyel M (2016) The Hamiltonian Brain: Efficient Probabilistic Inference with Excitatory-Inhibitory Neural Circuit Dynamics. PLoS Comput Biol 12: e1005186.

Allen M, Frank D, Schwarzkopf DS, Fardo F, Winston JS, Hauser TU, Rees G (2016) Unexpected arousal modulates the influence of sensory noise on confidence. Elife 5.

Allen M, Friston KJ (2018) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese 195: 2459-2482.

Bach DR, Dayan P (2017) Algorithms for survival: a comparative perspective on emotions. Nat Rev Neurosci 18: 311-319.

Bard P (1928) A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. American Journal of Physiology--Legacy Content 84: 490-515.

Barrett LF, Quigley KS, Bliss-Moreau E, Aronson KR (2004) Interoceptive sensitivity and self-reports of emotional experience. J Pers Soc Psychol 87: 684-97.

Barrett LF, Simmons WK (2015) Interoceptive predictions in the brain. Nat Rev Neurosci 16: 419-29.

Bedi G, Preston KL, Epstein DH, Heishman SJ, Marrone GF, Shaham Y, de Wit H (2011) Incubation of cue-induced cigarette craving during abstinence in human smokers. Biol Psychiatry 69: 708-11.

Berridge KC (2012) From prediction error to incentive salience: mesolimbic computation of reward motivation. Eur J Neurosci 35: 1124-43.

Braver TS, Barch DM, Cohen JD (1999) Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. Biol Psychiatry 46: 312-28.

Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. Nat Neurosci 18: 590-6.

Cannon WB (1927) The James-Lange theory of emotions: A critical examination and an alternative theory. The American journal of psychology 39: 106-124.

Chase HW, Eickhoff SB, Laird AR, Hogarth L (2011) The neural basis of drug stimulus processing and craving: an activation likelihood estimation meta-analysis. Biol Psychiatry 70: 785-93.

Chiu PH, Deldin PJ (2007) Neural evidence for enhanced error detection in major depressive disorder. Am J Psychiatry 164: 608-16.

Chung D, Kadlec K, Aimone JA, McCurry K, King-Casas B, Chiu PH (2017) Valuation in major depression is intact and stable in a non-learning environment. Sci Rep 7: 44374.

Conrad KL, Tseng KY, Uejima JL, Reimers JM, Heng LJ, Shaham Y, Marinelli M, Wolf ME (2008) Formation of accumbens GluR2-lacking AMPA receptors mediates incubation of cocaine craving. Nature 454: 118-21.

Doya K (2007) Bayesian brain: Probabilistic approaches to neural coding. MIT press

Dutton DG, Aron AP (1974) Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety. Journal of Personality and Social Psychology 30: 510-517.

Engelmann JM, Versace F, Robinson JD, Minnix JA, Lam CY, Cui Y, Brown VL, Cinciripini PM (2012) Neural substrates of smoking cue reactivity: a meta-analysis of fMRI studies. Neuroimage 60: 252-62.

Fiore VG, Ognibene D, Adinoff B, Gu X (2018) A Multilevel Computational Characterization of Endophenotypes in Addiction. eNeuro 5.

Fleming SM, Daw ND (2017) Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychol Rev 124: 91-114.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360: 815-36.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11: 127-38.

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G (2017) Active Inference: A Process Theory. Neural Comput 29: 1-49.

Friston KJ, Stephan KE, Montague R, Dolan RJ (2014) Computational psychiatry: the brain as a phantastic organ. Lancet Psychiatry 1: 148-58.

Fusar-Poli P, Placentino A, Carletti F, Landi P, Allen P, Surguladze S, Benedetti F, Abbamonte M, Gasparotti R, Barale F, Perez J, McGuire P, Politi P (2009) Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J Psychiatry Neurosci 34: 418-32.

Gentsch A, Sel A, Marshall AC, Schutz-Bosbach S (2019) Affective interoceptive inference: Evidence from heart-beat evoked brain potentials. Hum Brain Mapp 40: 20-33.

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND (2016) Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife 5.

Grimm JW, Hope BT, Wise RA, Shaham Y (2001) Neuroadaptation. Incubation of cocaine craving after withdrawal. Nature 412: 141-2.

Gu X (2018) Incubation of craving: a Bayesian account. Neuropsychopharmacology 43: 2337-2339.

Gu X, Filbey F (2017) A bayesian observer model of drug craving. JAMA Psychiatry 74: 419-420.

Gu X, FitzGerald TH (2014) Interoceptive inference: homeostasis and decision-making. Trends Cogn Sci 18: 269-70.

Gu X, Hof PR, Friston KJ, Fan J (2013) Anterior insular cortex and emotional awareness. J Comp Neurol 521: 3371-88.

Gu X, Lohrenz T, Salas R, Baldwin PR, Soltani A, Kirk U, Cinciripini PM, Montague PR (2015) Belief about nicotine selectively modulates value and reward prediction error signals in smokers. Proc Natl Acad Sci U S A 112: 2539-44.

Gu X, Lohrenz T, Salas R, Baldwin PR, Soltani A, Kirk U, Cinciripini PM, Montague PR (2016) Belief about Nicotine Modulates Subjective Craving and Insula Activity in Deprived Smokers. Front Psychiatry 7: 126.

Hauser TU, Fiore VG, Moutoussis M, Dolan RJ (2016) Computational Psychiatry of ADHD: Neural Gain Impairments across Marrian Levels of Analysis. Trends Neurosci 39: 63-73.

Hauser TU, Iannaccone R, Ball J, Mathys C, Brandeis D, Walitza S, Brem S (2014) Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. JAMA Psychiatry 71: 1165-73.

Hauser TU, Iannaccone R, Dolan RJ, Ball J, Hattenschwiler J, Drechsler R, Rufer M, Brandeis D, Walitza S, Brem S (2017a) Increased fronto-striatal reward prediction errors moderate decision making in obsessive-compulsive disorder. Psychol Med 47: 1246-1258.

Hauser TU, Moutoussis M, Iannaccone R, Brem S, Walitza S, Drechsler R, Dayan P, Dolan RJ (2017b) Increased decision thresholds enhance information gathering performance in juvenile Obsessive-Compulsive Disorder (OCD). PLoS Comput Biol 13: e1005440.

Huys QJ, Daw ND, Dayan P (2015) Depression: a decision-theoretic analysis. Annu Rev Neurosci 38: 1-23.

James W (1884) What is an emotion? Mind os-IX: 188-205.

Jasinska AJ, Stein EA, Kaiser J, Naumer MJ, Yalachkov Y (2014) Factors modulating neural reactivity to drug cues in addiction: a survey of human neuroimaging studies. Neurosci Biobehav Rev 38: 1-16.

Jepma M, Koban L, van Doorn J, Jones M, Wager TD (2018) Behavioural and neural evidence for self-reinforcing expectancy effects on pain. Nat Hum Behav 2: 838-855.

Juliano LM, Fucito LM, Harrell PT (2011) The influence of nicotine dose and nicotine dose expectancy on the cognitive and subjective effects of cigarette smoking. Exp Clin Psychopharmacol 19: 105-15.

Kelemen WL, Kaighobadi F (2007) Expectancy and pharmacology influence the subjective effects of nicotine in a balanced-placebo design. Exp Clin Psychopharmacol 15: 93-101.

Keller H (1881) Letter from Heller Keller to Rev. Phillips Brooks.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. TRENDS in Neurosciences 27: 712-719.

Lange CG, James W (1922) The emotions. Williams & Wilkins

Lawson RP, Mathys C, Rees G (2017) Adults with autism overestimate the volatility of the sensory environment. Nat Neurosci 20: 1293-1299.

Lawson RP, Rees G, Friston KJ (2014) An aberrant precision account of autism. Front Hum Neurosci 8: 302.

Lazarus RS (1991) Progress on a cognitive-motivational-relational theory of emotion. Am Psychol 46: 819-34.

LeDoux JE (2000) Emotion circuits in the brain. Annu Rev Neurosci 23: 155-84.

Lemogne C, le Bastard G, Mayberg H, Volle E, Bergouignan L, Lehericy S, Allilaire JF, Fossati P (2009) In search of the depressive self: extended medial prefrontal network during self-referential processing in major depression. Soc Cogn Affect Neurosci 4: 305-12.

Lu L, Hope BT, Dempsey J, Liu SY, Bossert JM, Shaham Y (2005) Central amygdala ERK signaling pathway is critical to incubation of cocaine craving. Nat Neurosci 8: 212-9.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9: 1432-8.

Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14: 154-62.

Mason L, Eldar E, Rutledge RB (2017) Mood Instability and Reward Dysregulation-A Neurocomputational Model of Bipolar Disorder. JAMA Psychiatry 74: 1275-1276.

McBride D, Barrett SP, Kelly JT, Aw A, Dagher A (2006) Effects of expectancy and abstinence on the neural response to smoking cues in cigarette smokers: an fMRI study. Neuropsychopharmacology 31: 2728-38.

McCusker CG, Brown K (1990) Alcohol-predictive cues enhance tolerance to and precipitate "craving" for alcohol in social drinkers. J Stud Alcohol 51: 494-9.

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16: 72-80.

Moutoussis M, Bentall RP, El-Deredy W, Dayan P (2011) Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. Cogn Neuropsychiatry 16: 422-47.

Moutoussis M, Fearon P, El-Deredy W, Dolan R, Friston K (2014) Bayesian inferences about the self (and others): A review. Consciousness and Cognition 25: 67-76.

Murphy FC, Nimmo-Smith I, Lawrence AD (2003) Functional neuroanatomy of emotions: a meta-analysis. Cogn Affect Behav Neurosci 3: 207-33.

Nestler EJ, Carlezon WA, Jr. (2006) The mesolimbic dopamine reward circuit in depression. Biol Psychiatry 59: 1151-9.

Ondobaka S, Kilner J, Friston K (2017) The role of interoceptive inference in theory of mind. Brain Cogn 112: 64-68.

Owens AP, Allen M, Ondobaka S, Friston KJ (2018) Interoceptive inference: From computational neuroscience to clinic. Neurosci Biobehav Rev 90: 174-183.

Papez JW (1937) A proposed mechanism of emotion. Archives of Neurology & Psychiatry 38: 725-743.

Parvaz MA, Moeller SJ, Goldstein RZ (2016) Incubation of Cue-Induced Craving in Adults Addicted to Cocaine Measured by Electroencephalography. JAMA Psychiatry 73: 1127-1134.

Pellicano E, Burr D (2012) When the world becomes 'too real': a Bayesian explanation of autistic perception. Trends Cogn Sci 16: 504-10.

Pessoa L, Adolphs R (2010) Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. Nat Rev Neurosci 11: 773-83.

Petzschner FH, Weber LAE, Gard T, Stephan KE (2017) Computational Psychosomatics and Computational Psychiatry: Toward a Joint Framework for Differential Diagnosis. Biol Psychiatry 82: 421-430.

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

Pezzulo G, Rigoli F, Friston K (2015) Active Inference, homeostatic regulation and adaptive behavioural control. Prog Neurobiol 134: 17-35.

Phan KL, Wager T, Taylor SF, Liberzon I (2002) Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage 16: 331-48.

Phelps EA (2006) Emotion and cognition: insights from studies of the human amygdala. Annu Rev Psychol 57: 27-53.

Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: Analysis of affective physiological state. IEEE transactions on pattern analysis and machine intelligence 23: 1175-1191.

Pizzagalli DA, Peccoralo LA, Davidson RJ, Cohen JD (2006) Resting anterior cingulate activity and abnormal responses to errors in subjects with elevated depressive symptoms: a 128-channel EEG study. Hum Brain Mapp 27: 185-201.

Powers AR, Mathys C, Corlett PR (2017) Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science 357: 596-600.

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9: 545-56.

Redish AD, Johnson A (2007) A computational model of craving and obsession. Ann N Y Acad Sci 1104: 324-39.

Robinson OJ, Cools R, Carlisi CO, Sahakian BJ, Drevets WC (2012) Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. Am J Psychiatry 169: 152-9.

Robinson TE, Berridge KC (1993) The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Res Brain Res Rev 18: 247-91.

Rushworth MF, Behrens TE (2008) Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11: 389-97.

Rutledge RB, Moutoussis M, Smittenaar P, Zeidman P, Taylor T, Hrynkiewicz L, Lam J, Skandali N, Siegel JZ, Ousdal OT, Prabhu G, Dayan P, Fonagy P, Dolan RJ (2017) Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. JAMA Psychiatry 74: 790-797.

Schachter S (1964) The Interaction of Cognitive and Physiological Determinants of Emotional State. Adv Exp Soc Psychol 1: 49-80.

Schachter S, Singer JE (1962) Cognitive, Social, and Physiological Determinants of Emotional State. Psychological Review 69: 379-399.

Schwartenbeck P, FitzGerald TH, Mathys C, Dolan R, Wurst F, Kronbichler M, Friston K (2015) Optimal inference with suboptimal models: addiction and active Bayesian inference. Med Hypotheses 84: 109-17.

Seth AK (2013) Interoceptive inference, emotion, and the embodied self. Trends Cogn Sci 17: 565-73.

Seth AK (2014) The cybernetic Bayesian brain: from interoceptive inference to sensorimotor contingencies. In: Metzinger TKaW, Jennifer M (ed) Open MIND. MIND Group, Frankfurt am Main,, pp 1-24

Seth AK, Friston KJ (2016) Active interoceptive inference and the emotional brain. Phil Trans R Soc B 371: 20160007.

Shiffman S, Li X, Dunbar MS, Tindle HA, Scholl SM, Ferguson SG (2015) Does laboratory cue reactivity correlate with real-world craving and smoking responses to cues? Drug and Alcohol Dependence 155: 163-169.

Slochower J (1976) Emotional Labeling and Overeating in Obese and Normal Weight Individuals. Psychosom Med 38: 131-139.

Tang DW, Fellows LK, Small DM, Dagher A (2012) Food and drug cues activate similar brain regions: a meta-analysis of functional MRI studies. Physiol Behav 106: 317-24.

Tian Y, Du J, Spagna A, Mackie MA, Gu X, Dong Y, Fan J, Wang K (2016) Venlafaxine treatment reduces the deficit of executive control of attention in patients with major depressive disorder. Sci Rep 6: 28028.

Tiffany ST, Friedman L, Greenfield SF, Hasin DS, Jackson R (2012) Beyond drug use: a systematic consideration of other outcomes in evaluations of treatments for substance use disorders. Addiction 107: 709-718.

Tiffany ST, Wray JM (2012) The clinical significance of drug craving. Ann N Y Acad Sci 1248: 1-17.

Tsakiris M, Critchley H (2016) Interoception beyond homeostasis: affect, cognition and mental health. Philosophical Transactions of the Royal Society B: Biological Sciences 371.

Valins S (1966) Cognitive Effects of False Heart-Rate Feedback. Journal of Personality and Social Psychology 4: 400-&.

Whitton AE, Treadway MT, Pizzagalli DA (2015) Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. Curr Opin Psychiatry 28: 7-12.

Wiech K (2016) Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. Science 354: 584-587.

Xiang T, Lohrenz T, Montague PR (2013) Computational substrates of norms and their violations during social exchange. J Neurosci 33: 1099-108a.

Yalachkov Y, Kaiser J, Naumer MJ (2012) Functional neuroimaging studies in addiction: multisensory drug stimuli and neural cue reactivity. Neurosci Biobehav Rev 36: 825-35.