Neuroscience of Consciousness, 2020, 00(0): niaa026

doi: 10.1093/nc/niaa026
Review Article

Interoceptive active inference and self-representation in social anxiety disorder (SAD): exploring the neuro-cognitive traits of the SAD self

Philip Gerrans[1,*] and Ryan J. Murray[2]

[1]Department of Philosophy, University of Adelaide, Adelaide, Australia; and [2]Department of Psychiatry, Faculty of Medicine, University of Geneva, Geneva, Switzerland

*Correspondence address. Department of Philosophy, University of Adelaide, Adelaide, Australia. Tel: +61 0 409 385 662; Fax: +61 8 831 343 41; E-mail: philip.gerrans@adelaide.edu.au

Abstract

This article provides an interoceptive active inference (IAI) account of social anxiety disorder (SAD). Through a neurocognitive framework, we argue that the cognitive and behavioural profile of SAD is best conceived of as a form of maladaptive IAI produced by a negatively biased self-model that cannot reconcile inconsistent tendencies to approach and avoid social interaction. Anticipated future social interactions produce interoceptive prediction error (bodily states of arousal). These interoceptive states are transcribed and experienced as states of distress due to the influence of inconsistent and unstable self-models across a hierarchy of interrelated systems involved in emotional, interoceptive and affective processing. We highlight the role of the insula cortex, in concert with the striatum, amygdala and dorsal anterior cingulate in the generation and reduction of interoceptive prediction errors as well as the resolution of social approach-avoidance conflict. The novelty of our account is a shift in explanatory priority from the representation of the social world in SAD to the representation of the SAD self. In particular, we show how a high-level conceptual self-model of social vulnerability and inadequacy fails to minimize prediction errors produced by a basic drive for social affiliation combined with strong avoidant tendencies. The result is a cascade of interoceptive prediction errors whose attempted minimization through action (i.e. active inference) yields the symptom profile of SAD. We conclude this article by proposing testable hypotheses to further investigate the neurocognitive traits of the SAD self with respect to IAI.

Keywords: social anxiety; philosophy; psychiatry; theories and models; interoception; active inference

Introduction

In this article, we apply the concept of active inference to explain the nature of social anxiety disorder (SAD). Convergences between active inference theories of emotion, interoception, self-representation and psychiatric disorder allow us to explain the complex relationship between neural correlates of SAD and its phenomenological/clinical profile. In this respect, we provide a systematic account of the neural correlates of SAD (Chalmers 1996; Hohwy and Seth 2020). Systematic accounts explain correlations between neural anatomy and activity in terms of the representational and processing roles of neural circuitry. We argue that an *interoceptive active inference* (IAI) account of the nature of self-modelling and emotional processing provides that framework. IAI treats the brain as a computational device engaged in a constant iterative process of modelling the world and its own states in order to optimise basic physiological regulation through action (Friston *et al.* 2013; Pezzulo *et al.* 2015; Barrett *et al.* 2016; Seth and Friston 2016; Barrett 2017; Kirchhoff *et al.* 2018; Von Mohr and Fotopoulou 2018). The models relevant to SAD are (i) a hierarchal multidimensional model of the self, represented as the hidden endogenous cause of affective

Received: 18 September 2020; **Revised:** 27 October 2020. **Accepted:** 30 October 2020

experience and (ii) models of the emotionally salient properties of the social world (other people's attitudes and actions). Subjects use predictive models of affective consequences of social interaction to reduce negative affect and produce/prolong positive affect (Paulus and Stein 2010; Joffily and Coricelli 2013).

The specific thesis we defend is that neural and psychological differences between SAD and non-SAD minds reflect differences in self-representation (self-modelling) that determine the individual's capacity to navigate her social world. In effect, we provide a psychiatric case study of an idea expressed by Moutoussis and Fearon.

> ... *inferred* representations of the self have a normative function: to *predict* and *optimise* the likely outcomes of social interaction... people make inferences about themselves—and others—to minimize interpersonal surprise, enabling them to make *decisions that are most consistent with their model of the inter-personal world*. (our italics) (Moutoussis *et al.* 2014)

In this treatise, we do not provide any new empirical evidence, *per se*. Rather, we introduce the active inference theoretical framework to conceptualize SAD. We use this framework to describe neural systems and regions implicated in predictive modelling of future affective states and the signalling and resolution of discrepancies between observed and predicted states (i.e. error processing). At the neural level, we focus mainly on key structures of the cortico-striatal-limbic system that, in concert with anterior insula functioning, permit the individual to confidently interpret her bodily states, integrate this experience with her ongoing mental schema of self-representation and predict future affective states in light of associated contexts (Paulus and Stein 2010; Campbell-Sills *et al.* 2011). These structures include the salience network [anterior insula and dorsal anterior cingulate (dACC)] (Isomura *et al.* 2003; Medford and Critchley 2010; Klumpp *et al.* 2012); dopaminergic system (striatum) and the limbic system (amygdala) (Bechara *et al.* 1999; Guyer *et al.* 2008; Blackford *et al.* 2014), systems linked to interoceptive predictive signalling (Ainley *et al.* 2013; Barrett and Simmons 2015; Barrett 2017; Tsakiris and De Preester 2018; Velasco and Loev 2020) and prediction error processing (Hayden *et al.* 2011; Shenhav *et al.* 2013; Ribas-Fernandes *et al.* 2019), reward prediction (Knutson *et al.* 2001; Knutson and Cooper 2005) and relevance appraisal and approach-avoidance regulation (Sander *et al.* 2003; Guex *et al.* 2020), respectively. Together, this interdependent circuitry allows for the representation of self via internal bodily cues, generation of predictions of anticipated interoceptive and affective states, appraisal of the importance of such states to subjective goals and well-being and the management of discord between predictions and actual outcomes through active inference. We will argue that, due to maladaptive self-representation (predictive modelling of hidden endogenous causes), interoceptive prediction errors [i.e. signals of discrepancy between anticipated and actual physiological states (Paulus and Stein 2010)] are poorly interpreted in SAD. The result is negatively biased predictions of future affective states resulting from unstable models of the social self.

Importantly, extant literature attests to consistent alterations in such neural systems within SAD, particularly when anticipating social threat (Guyer *et al.* 2008; Boehme *et al.* 2013; Cremers *et al.* 2015), likely pointing to altered self-representation, interoceptive predictive signalling and error processing in anticipation of aversive (i.e. social) events. In a paper currently in preparation, we outline in greater detail a neurobiological framework to explain unstable conceptual self-processing in SAD and its relation to threat predictions. Here, however, we focus specifically on the above-mentioned neural systems likely underlying IAI that may mediate social threat predictions (STP) symptomatic of SAD.

On the view we propose, the emphasis on explanation of SAD shifts from the (mis)representation of the threat occasioned by the social world to the way the subject constructs and maintains a self-representation. Intuitively, and clinically, SAD presents as a suite of exaggerated fear responses to anticipated social interactions (STP). However, we argue that this misrepresentation of the social world as intractably hostile derives from a representation of the self as vulnerable and unable to cope with social adversity. These STP misrepresentations stem from an inability to manage anticipatory distress deriving from a weak and uncertain self-model. This exemplifies a general feature of emotional processing: self-representation conditions the way the emotional world is represented (Mellings and Alden 2000; Wong and Moulds 2011).

Here, we first explain the nature of SAD. We then introduce the concept of IAI and show how it explains the constitutive interdependence between emotional processing and self-representation. Throughout, we highlight related neural differences in SAD, relative to healthy individuals as well as other anxiety disorders. We demonstrate how these neural differences can potentially be explained largely in terms of the effects on hierarchical IAI of an inconsistent/bistable self-model that regulates basic mechanisms of social approach/avoidance. The fundamental inability to adaptively resolve conflicting motivations to approach/avoid social encounters cascades through the hierarchy of social–emotional processing, producing rigid maladaptive cognitive biases and intense and unmanageable distress.

Social anxiety disorder

SAD is defined by a persistent and intense anticipatory fear of negative social evaluation and manifests as heightened social threat sensitivity (Alfano *et al.* 2006; Kambouropoulos *et al.* 2014) and anticipatory anxiety to social situations (Mellings and Alden 2000, Wong and Moulds 2011; Boehme *et al.* 2013; Mills *et al.* 2014). Adverse outcomes include avoidance, social withdrawal, and negative rumination about self. This suite of responses ultimately reinforces maladaptive cognition and behaviour in a vicious circle.

The clinical profile of SAD includes a range of features that point to basic cognitive-affective problems that *underlie* the STP symptoms. Any systematic account of SAD thus needs to explain these features: (i) relevance and motivational salience. SAD patients have a strong drive to engage socially and are also over-dependent on rewarding social feedback for self-respect (Gilboa-Schechtman *et al.* 2000; Aderka *et al.* 2009; Weisman *et al.* 2011; Aderka *et al.* 2012). As noted above, however, they have exaggerated aversive response to anticipated social engagement and the fear of the threat of punishment, such as rejection or ridicule. Thus, their basic motivational/reward structure for social engagement is unstable and inconsistent; (ii) SAD subjects have deficiencies in producing imagery of rewarding engagement, such as autobiographical memory or imaginative rehearsal (Stopa and Jenkins 2007; Stopa *et al.* 2010), likely reinforcing unstable self-images and models; (iii) SAD subjects have explicit symbolic/linguistic representations of themselves as vulnerable and unable to cope with adversity (Hope *et al.* 1990; Wilson and Rapee 2005, 2006), (iv) reflected in the inability to tolerate uncertain social situations (Boelen and Reijntjes

2009; Carleton *et al.* 2010; Campbell-Sills *et al.* 2011; Carleton 2012; Carleton *et al.* 2013; Campbell-Sills *et al.* 2015).

As we argue in a paper currently in preparation, these features give rise to models of the self that are conflicted, inconsistent and unable to reliably predict future social self-efficacy and coping. Ultimately, this results in a maladaptive social self-model troubled by inconsistent/bistable motivational/affective processing. Here, bistability refers to the competing motivational forces of approach (reward-seeking) and avoidant (punishment-averting) behaviours, both of which operate concurrently yet unstably. Excessive dependence on social reward, combined with lack of predicted ability to obtain those rewards, makes the SAD self, therefore, vulnerable and dependent. That vulnerability and dependency put her in a maladaptive state of hypervigilance and distress when she anticipates social encounters. This feeling of hypervigilance, produced by interoceptive signalling, is interpreted by interdependent higher-level predictive models (i) of the self as vulnerable and unable to cope with social adversity and (ii) of the social world as intractably hostile. The result is that these signals are transcribed as states of extreme anticipatory distress. These models bias emotional processing at all levels from basic approach/avoid tendencies to explicit deliberation. As we will show below, SAD derives from an inability to minimize interoceptive prediction errors due to the interaction of biased bodily state predictions and poorly defined models of self-efficacy and coping.

Interoceptive active inference

The concept of IAI unites two influential ideas. The first is that cognition is a form of action designed to minimize prediction error: the discrepancy between state predicted by a generative model and actual state. The second is that the ultimate function of cognition is basic physiological regulation (allostasis as the process is technically known). As Tsakiris and de Preester put it, 'cognition is enslaved to embodiment' (Tsakiris and De Preester 2018). Combining these ideas yields the thought that the mind constructs and uses probabilistic models of the organism's internal and external world to enable it to act adaptively (cf. Murray *et al.* 2015b for a discussion). The test of successful adaptation is allostatic optimization, signalled by production or maintenance of positive affect and reduction of negative affect. Thus, affective states unpredicted in context are signals of error and consequent action to modulate affect is allostatic active inference (Joffily and Coricelli 2013; Barrett *et al.* 2016; Corcoran and Hohwy 2017; Kleckner *et al.* 2017; Van de Cruys 2017; Velasco and Loev 2020).

Allostasis is a refinement of the concept of homeostasis that refers to context-dependent *anticipatory* regulation of states of the *internal milieu* (Corcoran and Hohwy 2017; Seth and Tsakiris 2018) and can be considered 'the activity of achieving homeostasis through physiological and behavioural change' (Velasco and Loev 2020). It thus extends the concept of homeostasis, which implies a reflexive return to optimal set points. Allostasis entails that for some variables, set points can be changed according to predicted consequences of action in context *and* that the relevant regulatory mechanisms include active cognition and overt behaviour as well as homeostatic reflex arcs. Interoceptive experience informs us at the systemic level of success or failure of allostatic regulation and entrains the next round of regulatory activity. This regulative role for interoception is reflected in the way it models hidden causes of changing endogenous states as fluctuations in the global feeling state. The feeling of fatigue for example is a way of experiencing the result of a complex suite of metabolic interactions that are opaque to introspection and cannot be directly regulated. However, we can rest, which reprograms the suite of lower-level processes. Thus, interoceptive regulation at a systemic level is a proxy for allostatic regulation at the subsystemic and molecular levels (Stephan *et al.* 2016).

The relevance to the IAI explanation of SAD is that the concept of allostasis extends the mechanisms of basic bodily regulation to include not just mechanisms of autonomic regulation but any form of cognitive or behavioural activity whose ultimate goal is optimizing physiological regulation.

> ... interoceptive processing should be extended beyond "homeostatic control of the internal milieu" to incorporate "allostatic actions on the external world" (Gu and Fitzgerald, 2014 p. 269; Corcoran and Hohwy 2017).

Thus, for example, having a cold drink or going for a swim on a hot day are forms of IAI. Similarly, affective regulation through action (social avoidance or approach) programs a suite of lower-level regulatory systems bottoming out in allostatic regulation. These affective states are produced by processing of the interoceptive signal using predictive models that generate a sense of self, modelled as the entity that feels the consequences of action (Limanowski and Blankenburg 2013; Seth 2013; Gerrans 2014; Moutoussis *et al.* 2014; Sel 2014; Hohwy and Michael 2017; Letheby and Gerrans 2017). We argue below the signature of SAD is the way this modelling process generates anticipatory distress for social situations.

As suggested by Hohwy and Michael (2017), interoception gives rise to the experience of a fundamental level of embodied selfhood via inference about the hidden internal causes of interoceptive experience. This means the self is fundamentally a predictive model of an entity generated to explain coherence between allostatic variables that track vital physiological fluctuations. Experience of internal visceral responses is critical for informing the organism of its current and anticipated cognitive-affective states, as it acts in its environment and allows for a sense of self, separate from its physical (and social) environment. In this respect, this concept put forth by Howhy and Michael extends an idea nicely put by Seth (2013):

> Mental representations of selfhood are ultimately grounded in representations of the body, with the internal physiological milieu providing a primary reference—a "material me". Seth (2013).

Hohwy and Michael, like Seth, build on previous somatic accounts of self-awareness (e.g. Damasio 2006; Blanke and Metzinger 2009). However, they situate their account of embodied selfhood in a framework that treats cognition as the minimization of prediction error through IAI. *Active inference* refers to searching in an evidential space whose dimensions are defined by predictive models or visiting unfamiliar states in an attempt to optimize a model (Friston et al. 2013; Pezzulo *et al.* 2015; Barrett *et al.* 2016; Seth and Friston 2016; Kirchhoff et al. 2018; Hohwy and Seth 2020). Thus, active inference involves *acting* to gain new evidence to support a predictive model. Acting here includes internal regulatory, overt behavioural (e.g. Friston *et al.* 2013) and cognitive (e.g. Murray *et al.* 2015b) activity. In the context of SAD, active inference includes behavioural strategies such as social withdrawal designed to reduce interoceptive prediction error. By focusing on interoceptive information about states of the visceral *milieu* (e.g. dry throat, thumping heart), the individual can infer the likely causes of her bodily responses and affective states. These physio-affective experiences then serve targets of modulation through action. Cognitive forms of

active inference include affectively scaffolded recollection and anticipation in episodes of called 'mental time travel'. In these episodes, subjects engage in 'affective sampling' of past or future states to help them act adaptively both in the present and for the future (Suddendorf and Corballis 2007; Murray *et al.* 2015a). Mental time travel is subtended by circuitry specialized for this type of self-referential gathering of evidence about expected affective consequences of action (Gusnard *et al.* 2001; Sridharan *et al.* 2008; Broyd *et al.* 2009; Spreng *et al.* 2009; Carhart-Harris and Friston 2010).

Interoceptive prediction error in SAD

Altered interoception is highly common in anxiety disorders (Paulus and Stein 2006), including SAD (cf. Domschke *et al.* 2010 for a review). In anxiety, the individual engenders a normal baseline physiological condition but an exaggerated expected body state (Paulus and Stein 2010). This large interoceptive prediction error [i.e. the discrepancy between actual and expected physiological states (Paulus and Stein 2006; Seth *et al.* 2012)] often triggers specific cognitions (e.g. worry) and behaviour (e.g. avoidance) in an effort to attenuate the error signals and improve prediction precision (Paulus and Stein 2006). In social anxiety, this worry manifests as STPs and likely stems from a belief-based assumption that the expected interoceptive afferents signal imminent social threat (Collimore and Asmundson 2014). In anxiety disorders, *a priori* beliefs influence strongly anticipatory interoceptive signalling *and* the attempts to attenuate the ensuing prediction error (Paulus and Stein 2006). What is unique in the SAD case, however, is that these beliefs centre on the view that the social self is inadequate, i.e., socially ineffective (Koban and Pourtois 2014) and unable to cope with unexpected social challenges (Iancu *et al.* 2015a,b). In addition, bistable self-representations and conflicting social motivations increase stress and cognitive-affective conflict, contributing to feelings of worry and doubt. The SAD individual thus responds to errors in predicted physiological states in future social interactions (e.g. racing heart, dry throat, flushed cheeks) with increased (biased) belief-based cognitions that would signal imminent social threat (i.e. STPs).

Model precision in SAD

In order to better explain SAD, we need to describe another aspect of the IAI framework: the estimation of precision. Precision refers to the estimation of signal-to-noise ratio in a signal of prediction error, given a model, in order to reduce uncertainty. In general, high precision stabilizes a predictive model whereas low precision initiates or encourages active inference. Furthermore, in the sense of allocation of cognitive and behavioural resources, the mind prefers models that maximize precision *across a processing hierarchy*. Which is just to say that the mind prefers the model estimated to be the 'best fit' for the signal estimated to be prediction error, given a model (Friston and Kiebel 2009; Hohwy 2013; Limanowski and Blankenburg 2013). In the case of SAD, for example, an imprecise interoceptive prediction error generated in anticipation of a social encounter is estimated to be a signal of a potential social threat given the prior model of the self as unable to cope with social adversity. Unfortunately for the SAD subject, there is no precise model available that *stably* minimizes error *across the hierarchy*. The rigid conceptual level self-model that predicts rejection and humiliation is inconsistent with her fundamental drive for social engagement. Thus in SAD, the inability to minimize prediction error across the hierarchy results from the interaction between (i) high-level predictive models that assign high precision to exaggerated interoceptive prediction errors and interpret them as intractable STPs and (ii) the low-level systems that generate those errors as part of allostatic regulation. The contribution of each of these mechanisms can be analysed independently (indeed, we do so in the section on neural correlates), but predictive coding models suggest that the phenomenology of SAD is an emergent result of reciprocal interaction and mutual reinforcement between high-level self-models and signals of interoceptive prediction error.

In order to initiate active inference, the precision of a model must be relaxed, which increases uncertainty. Intuitively, to initiate a search for new evidence, the subject needs to reduce the strength of belief in an existing hypothesis, thus leaving more room for ambiguity and doubt. For the SAD subject, relaxing precision on her self-model would thus (temporarily) reinstate a state of imprecise but intense distress. An adaptive strategy would be to inhibit processes generating that distress in order to enable active inference (e.g. cognitive and behavioural exploration of potentially socially rewarding actions). However, because of rigidities in processes that generate these distressing bodily signals, coupled with an intolerance to uncertainty (Carleton 2012), the SAD subject cannot escape the anticipatory distress, does not explore cognitive alternatives to challenging existing self-models (Carleton 2012; Tanovic *et al.* 2018) and hence reverts to the maladaptive, comparatively precise, conceptual model that fits the distressing experience. This results not only in avoidant cognitions but also avoidant social behaviour, whereby the SAD subject denies herself rich meaningful interactions, which would provide positive experiences that could dispel negative biases and/or reinforce positive aspects of the self-model. This model, however, is actually unstable because it does not account for the underlying strong motivation to engage socially, despite the overt social avoidance, which is another basic feature of SAD psychology (Aderka *et al.* 2009; Weisman *et al.* 2011). The SAD subject therefore, remains simultaneously attracted to, and repelled by, social interaction. In SAD, consequently, aversive self-models remain rigid and resistant to change.

Unfortunately, however, that aversive model does not cancel the error signals emanating from below. Although she has a strong aversive reaction to anticipated social engagement, she retains a fundamental basic drive for social engagement and affiliation. At lower levels of self-model processing, the SAD individual possesses high motivation for bonding and affiliation to build and strengthen attachments and a fundamental sense of belonging (Aderka *et al.* 2009; Weisman *et al.* 2011). At higher levels of processing, however, she demonstrates an excessive dependence on social feedback to construct a conceptual model of the self as adequate and competent. Thus, the SAD subject is trapped by inconsistent motivational and cognitive biases deriving from her inability to construct a coherent multilevel self-model. She is condemned to interpret exaggerated interoceptive prediction errors as evidence of STPs. The rigidity of her higher-level self-model both reinforces these interpretations and prevents the SAD individual from reducing the ensuing prediction error by employing IAI.

Active inference and self-modelling in SAD

In SAD, individuals manifest maladaptive allostatic regulation in the anticipation of social situations perceived as unpredictable and/or uncontrollable. This reflects a poor predictive model

of future cognitive, behavioural and affective states (e.g. anxiety, stuttering, blushing, panic) in light of impending and uncertain social demands (e.g. date with a potential romantic partner, job interview). These predictions thus reinforce an already negatively biased self-model, resulting in STP cognitions. These STPs produce an imprecise state of anticipatory hyperarousal, which is symptomatic of SAD. The product of an inability to resolve social uncertainty and a negatively biased self-model is, therefore, a representation of the self as (i) ineffective in social situations, (ii) incapable of coping with social uncertainty and (iii) highly vulnerable to costly social loss in an unpredictable social event (Mellings and Alden 2000; Stopa and Jenkins 2007; Stopa *et al.* 2010; Wong and Moulds 2011; Boehme *et al.* 2014; Wong and Rapee 2016).

States of bodily arousal are interoceptive states. The experience of distress is the result of emotional processes that redescribe/reinterpret and contextualize interoceptive signals producing the experience of *affect*. This level of active inference models the self as a hidden cause of affective experience. Such self-modelling is crucial since it provides the interface between physiological/allostatic regulation and personal experience of the value and significance of life events. For example, it allows physiological arousal (e.g. blushing) to be experienced as part of an episode of humiliation which programs a suite of aversive behaviours designed to restore allostatic equilibrium.

At still higher levels the narrative self, the protagonist of a recountable autobiography, is an explicit conceptual representation of the integrated physical and cognitive life of the organism over time. Each level of self-modelling is a form of active inference that minimizes predictive error referred from lower levels. The relationship between interoception and affect thus reflects the general principle of hierarchical organization, with each level of the hierarchy modelling the self as the hidden cause of endogenously caused experience. As with interoceptive experience, affective experience serves as a regulative proxy for overall organismic well-being. Within the IAI framework, negative affect signals a failure to reduce prediction error across the system and instigates regulatory action via a cascade of downward commands that ultimately bottom out in allostatic regulation (Joffily and Coricelli 2013; Van de Cruys 2017; Velasco and Loev 2020). The result of the failure to reduce prediction error in SAD is that the mind is unable to incorporate stable, reliable and epistemically accurate models of uncertain future social events. Given the elevated degree of social motivation toward social acceptance and affiliation (Aderka *et al.* 2009, Cremers *et al.* 2015) coupled with a baseline fear of social loss and rejection, the SAD mind interprets poorly modelled bodily perturbations as intimations of impending and unpredictable threat and felt as anxiety, hypervigilance or panic; i.e., as affective states (Boehme *et al.* 2013; Garfinkel and Critchley 2013; Terasawa *et al.* 2013).

At the highest narrative or conceptual levels of hierarchical processing, affective states are interpreted using explicit models of the self and social world. The role of higher-level processing is not just to determine the relevance of interoceptive predictive error *post hoc* but to set the parameters *a priori* that determine which allostatic variations eventually become prediction errors. Higher-level models of self and world determine whether and how physiological changes are experienced. For example, aspects of the lethargy of fatigue and depression have the same bodily and molecular signature but the latter is experienced differently as part of the avolition and apathy of a depressive episode (Paulus and Stein 2010; Barrett *et al.* 2016; Stephan *et al.* 2016). In the latter case, the experience reflects the downward influence of models of self (hopeless and inadequate) and the social world (a site of irretrievable disconsolation). Furthermore, given biased and unstable models of self, integration of new empirical priors based on previous episodes (e.g. positive social interactions, praise, etc.) is altered. The prior self-model means that positive social information does not become salient due to poor attentional orienting and, as a result, social threat cues monopolize cognitive resources (Heinrichs and Hofmann 2001). The result is consolidation of the maladaptive SAD self-model. In other words, rigid higher-level models disable active inference across the hierarchy while at the same time failing to resolve inconsistencies generated by conflicting approach/avoid motivations.

Within this framework, we can see higher levels of emotional processing and self-modelling as forms of active inference designed to optimize affective regulation on behalf of a self that feels the consequences of action. At this level, the subject can explicitly interpret the emotional world, reflect and plan responses over long time scales using circuitry specialized for so-called mental time travel. At this level, the self is explicitly represented as the subject of an autobiographical history. This type of explicit self-modelling provides a relatively precise conceptual or linguistic interpretation of an imprecise affective experience.

Because the SAD subject models herself as invariably lacking self-efficacy and coping potential in social interactions, she cannot perform exploratory active inference (regulatory, cognitive or behavioural) to regulate distressing bodily experiences occasioned by anticipating social encounters. She cannot imaginatively inhabit a future in which she competently navigates the social world (Spurr and Stopa 2002; Stopa and Jenkins 2007). The SAD individual is likely to inhibit any exploratory cognition that would aim at reducing the ambiguity from signalling errors and thus improve prediction precision (Carleton 2012; Tanovic *et al.* 2018). This inability to minimize interoceptive prediction error cascades through a processing hierarchy of social cognitive-affective processing. We argue, therefore, that unlike other anxiety disorders, SAD individuals generate imprecise predictive social self-models, resulting from inaccurate predictive modelling of interoceptive signals.

SAD provides a compelling example of how an inability to reduce interoceptive prediction error results in increased negative affect (hyperarousal and hypervigilance in the case of SAD) and unstable models of self and world. Unlike other anxiety disorders, SAD is unique in that negative affect, worry and feelings of imminent threat stem not from external causes, *per se*, but from biased self-conceptual schema representing the SAD subject as an incapable social agent and vulnerable to intolerable social loss. Coupled with bistable self-motivations unique to SAD, poor mental imagery of self and one's coping potential, the social milieu is rendered as hostile, threatening and potentially destructive. Thus, in anticipation of ambiguous or uncertain social scenarios, SAD individuals manifest poor regulation of anticipated affective states. This is an instance of dysfunctional allostatic regulation.

We now turn to a more detailed description of the neurobiological mechanisms involved, showing how neural correlates of SAD function as components of an IAI hierarchy biased by inconsistencies in self-modelling.

Insular cortex and self-representation

The insular cortex, located in the lateral hemispheres of the brain, spanning both poster and anterior regions, is an established interoceptive hub (Sridharan *et al.* 2008; Bennett and

Baird, 2009; Craig 2009; Singer *et al.* 2009; Medford and Critchley 2010; Gasquoine 2014). As we noted earlier, regulating affective experience allows the organism to optimize allostasis. A primary hub of processing at this level of self-modelling is the anterior portion of the insula cortex, or anterior insula cortex (AIC). The AIC is specialized to re-represent and integrate information about body state afferents from the posterior portion of the insula to allow us to feel the significance of interoceptive states as affects (Craig 2009). The AIC thus plays an important role in transcribing interoceptive visceral states into felt affective states. It is consistently active in processing threat uncertainty (Morriss *et al.* 2019) and is considered to be a substrate of conceptual self-processing (Murray *et al* 2012; Murray *et al.* 2015a). Finally, the insular cortex exhibits resting functional connectivity with the striatum, the amygdala, dACC (Cauda *et al.* 2011), making the insula a key substrate not only for subserving conceptual self-representation but for facilitating reward predictions, affective relevance processing and error monitoring and resolution, respectively (Sridharan *et al.* 2008; Singer *et al.* 2009; Barrett and Simmons 2015). Thus, it is not surprising to see that contemporary affective neuroscience treats activation of the insula, particularly the AIC, as a representation of the integrated functioning of the organism, evaluated against emotionally salient goals *and creating a sense of self in the process*. As Craig (2009) put it:

> The integration successively includes homeostatic, environmental, hedonic, motivational, social and cognitive activity to produce a "global emotional moment", *which represents the sentient self at one moment of time.* (Craig 2009)

In other words, the affective self at any moment is the body under an emotional mode of presentation (to import some philosophical jargon).

The AIC's role here is as an integrative hub. It collates interoceptive and emotional information from across the hierarchy, allowing the organism to experience the 'feeling of what matters' to adapt a nice expression coined by another theorist of somatic self-representation (Damasio and Dolan 1999). This 'feeling of what matters' is consistent with the role of the AIC at the top level of a salience processing hierarchy that enables acquisition of adaptive patterns of response. The AIC likely allows for anticipatory processing, providing information about future aversive bodily states associated with specific contextual events (Paulus and Stein 2010), and it allows the organism to feel the subjective relevance of the prediction error between current and predicted states (Seth 2013).

SAD individuals show consistent alterations in insular cortex functioning and structure (Shah *et al.* 2009; Klumpp *et al.* 2012; Tang *et al.* 2012; Garfinkel and Critchley 2013; Terasawa *et al.* 2013; Klumpp *et al.* 2014; Kawaguchi *et al.* 2016; Duval *et al.* 2018; Wang *et al.* 2019). For instance, hyperactive insula activity is linked to increased SAD symptom severity (Schmidt *et al.* 2010) and is present more often in SAD than other anxiety disorders like post-traumatic stress disorder (PTSD) (Etkin and Wager 2007), making insula hyperactivity a potential biomarker for SAD. Furthermore, social anxiety is associated with altered AIC functional connectivity with the striatum (Blackford *et al.* 2014; Clauss *et al.* 2014), amygdala (Liao *et al.* 2010) and dACC (Tang *et al.* 2012), regions implicated in reward prediction (Knutson and Cooper 2005; Knutson *et al.* 2001) affective relevance appraisals (Sander *et al.* 2003), and prediction error monitoring and resolution (Ribas-Fernandes *et al.* 2019) respectively. It is, thus, not surprising that altered insula activity in SAD, particularly within the anterior portion, is consistently implicated in the emotional experience of anticipated social interactions (Gilboa-Schechtman *et al.* 2000, Singer *et al.* 2009, Medford and Critchley 2010, Weisman *et al.* 2011, Terasawa *et al.* 2013, Moayedi 2014). During anticipation, altered insula activity often presents parallel alterations in the striatum (Boehme *et al.* 2014), the amygdala (Boehme *et al.* 2013; Davies *et al.* 2017) and dACC (e.g. Davies *et al.* 2017). As we describe below, these alterations may likely give way not only to inaccurate interoceptive predictions but also to poor reward prediction, motivational regulation and error prediction that would be required to address imprecise interoceptive prediction error.

Cortico-striatal-limbic system and predictive processing

The generation of an affective interoceptive state likely occurs thanks to the insula's functional connectivity with the cortico-striatal-limbic system, which allows for reward prediction, relevance and goal-oriented appraisals and prediction error monitoring and resolution. The cortico-striatal-limbic system is consistently implicated in prediction error processing and minimization as well as affect regulation (Murray *et al.* 2020), and alterations in this system appear consistent in SAD (Furmark 2009; Davies *et al.* 2017). Key substrates include the striatum (Becker *et al.* 2017; Richey *et al.* 2017), amygdala (e.g. Meffert *et al.* 2015) and dACC (Davies *et al.* 2017).

Striatum

The striatum is a key subcortical substrate subserving prediction- and goal-related functions, such as anticipating reward (Knutson *et al.* 2001; Knutson and Cooper 2005) and processing unexpected reward (Stalnaker *et al.* 2012). Relative to healthy controls, SAD individuals exhibit inhibited ventral striatum activation when anticipating social situations, while also exhibiting increased insula activity (Boehme *et al.* 2014), suggesting elevated self-referential processing with diminished reward prediction. This blunted reward prediction of the ventral striatum was later demonstrated by Richey *et al.* (2017), who showed that relative to healthy controls, SAD individuals showed inhibited ventral striatal response in anticipation of positive social information, whereas they did not exhibit such inhibition when anticipating negative social information (Richey *et al.* 2017). This would suggest that when anticipating social situations, SAD individuals likely minimize the possibility of socially rewarding outcomes. This blunted striatal activation to future rewarding social information also persists during the reception of positive performance feedback, whereby SAD subjects show no increased striatal response to positive social reward, relative to healthy controls (Becker *et al.* 2017). This may contribute to a critical SAD feature of impaired social information processing and encoding (Heinrichs and Hofmann 2001). Altered processing of self-affirming social information may thus contribute to persistently and negatively biased self-models of inefficacy and vulnerability to social loss. Importantly, this striatal alteration in SAD is likely domain-specific, i.e., to social situations, and does not generalize to other non-social rewards, like monetary gain (Richey *et al.* 2017).

Amygdala

Within the limbic system, the amygdala is the primary hub of processing responsible for the coordination of perception and

interoception in response to social/emotional information. It drives low-level semiautomatic categorization, cognition and active inference (Adolphs *et al.* 2002; Sander *et al.* 2003; Yaniv *et al.* 2004). To do so, it integrates sensorimotor processing with activity in systems that reinforce adaptive exploratory or approach behaviour, the striatum and ventral tegmental area. These circuits operate as a dopaminergically driven reward prediction system (Bayer and Glimcher 2005; Pessiglione *et al.* 2006; Ruff and Fehr 2014; Cremers *et al.* 2015). At these levels of processing, the self is modelled as the target of adaptive physiological regulation and action tendencies consequent on low-level evaluation of properties of the social–emotional world. How the self is predicted to respond (e.g. to emotional expression) determines the balance of approach/avoidance behaviour and the consolidation of patterns thereof.

The amygdala is shown not only to respond to negative social prediction error, i.e., viewing unexpected negative reactions from others, but to orchestrate avoidant-approach behaviours following such prediction errors (Meffert *et al.* 2015). This suggests that the amygdala may mediate emotional responding to negative interoceptive prediction errors and may play a critical role in organizing approach-avoidant behaviours. In SAD, we see pervasive amygdala dysfunction, at rest (Pannekoek *et al.* 2013; Dodhia *et al.* 2014) and when anticipating the social interactions (Boehme *et al.* 2014; Davies *et al.* 2017). Furthermore, Blackford *et al.* (2014) showed increased positive resting functional connectivity between the insula and bilateral amygdala related to increased social inhibition in individuals with SAD traits. Although conducted in clinical healthy individuals, this suggests that interoceptive afferents from the insula, coupled with motivational processing of the amygdala, may mediate approach-avoidance behaviour in SAD. This thus indicates emotional regulation of responses to cues relevant for both approach and avoidance tendencies. These responses each elicit (are coupled with) increased body state predictions, which are then misinterpreted as threatening in SAD. Manipulating avoidance-approach motivations in social and non-social anticipatory processing in future research would contribute greatly to our understanding of the specificity by which the amygdala responds to opposing outcome possibilities (reward vs. punishment), both of which would implicate the SAD self.

This approach-avoidance bistability that is likely primed in the SAD subject may elicit a degree of ambiguity in assessing outcomes as well as behavioural planning. The amygdala exhibits elevated sensitivity to uncertainty (Morriss *et al.* 2019). Additionally, both the insula and amygdala appear reactive when anticipating information that is uncertain (Morriss *et al.* 2019), suggesting a likely coupling of interoceptive signalling and approach-avoidance regulation under uncertainty. Note that uncertainty has at least two independent but related dimensions: nature of the potential threat and capacity to cope. It is currently not easy to parse the relative contribution of these dimensions to anticipatory anxiety in SAD, but literature does attest to altered amygdala reactivity to negative and uncertain, or ambiguous, the information in SAD (Brühl *et al.* 2011). Nonetheless, future fMRI research would benefit from investigations manipulating reward and punishment anticipation certainty in SAD as well as measuring interoceptive sensitivity within the SAD individual in order to gauge to what extent anticipatory amygdala reactivity relates to altered goal-oriented behavioural planning stemming from bistable motivations or biased/exaggerated interoceptive signalling.

Dorsal anterior cingulate

The dACC is a frontal cortical midline structure located superior to the corpus callosum in the anterior cingulate, between the pregenual anterior and midcingulate cortices. It is consistently implicated in anticipatory pain processing (Kalisch *et al.* 2006; Mee *et al.* 2006; Drabant *et al.* 2011; Kalisch 2009) and is thus established as a critical hub for conscious threat predictions (Kalisch *et al.* 2006; Maier *et al.* 2015). It is, therefore, not surprising that anticipatory STPs relate to increased dACC activations in SAD (Clauss *et al.* 2014; Davies *et al.* 2017). Importantly, the dACC illustrates altered functioning in SAD during anticipatory social threat processing (Davies, *et al.* 2017; Lorberbaum *et al.* 2004).

The dACC is also implicated in goal-related motivational behavioural planning (Heilbronner and Hayden 2016) and value outcome estimations (Shenhav *et al.* 2013). In SAD, evidence of bistable motivational processing is suggested in resting-state dACC functional connectivity. That is, expectations of both social reward and punishment elicit increased negative striatum-dACC connectivity at rest, with respect to neutral outcome expectation. This would suggest elevated self-regulatory motivational processing when predicting both rewarding and punishing social outcomes (Cremers *et al.* 2015). This may thus speak to the bistable social motivations likely inherent in SAD, whereby the dACC mediates the regulatory response from deficient afferent reward-punishment predictions projecting from the striatum. This dual social motivational processing may ascribe elevated self-relevance to the potential for both bonding and rejection, requiring future state predictions and expected control estimations (cf. Shenhav *et al.* 2013) via striatal-dACC pathways.

The dACC is also highly sensitive to information relative to one's current state, likely thanks to inputs from the insula and the prefrontal cortex that it uses to estimate outcome values, in terms of their likelihood and costs, ultimately adjusting parameters to improve prediction precision (Shenhav *et al.* 2013). Indeed, the dACC may play an important role in monitoring and adjusting for prediction error (Hayden *et al.* 2011; Ribas-Fernandes *et al.* 2019; Shenhav *et al.* 2013), particularly when anticipating future states and situational outcomes (Shenhav *et al.* 2013). Importantly, the SAD subject exhibits a tendency to exaggerate the likelihood and costs of a negative social experience (Foa *et al.* 1996; Lucock and Salkovskis 1988). Unfortunately, to our knowledge, the SAD literature has yet to conduct seed-based resting-state functional connectivity within the dACC when anticipating both positive and negative social scenarios. Nonetheless, we can infer that estimates of one's own self-regulatory capacity to achieve social reward and avoid punishment (cf. Shenhav *et al.* 2013) are impaired in SAD, likely due to either deficient dACC function or altered dACC connectivity with substrates subserving self-state representations, e.g., the insula. The SAD literature would benefit from future studies conducting seed-based resting-state functional connectivity in the insula, amygdala, striatum and dACC when in anticipation of various social scenarios (e.g. public speaking) that assure only rewards (e.g. positive feedback) or punishments (e.g. negative feedback), and those that assure both rewards and punishments in order to discern motivationally driven responding in the insula and its connected corticolimbic structures.

The integrated model: from neuroanatomy to the SAD mind

Against this background, we summarize this theoretical analysis in an effort to provide greater clarity on the differences

between SAD and non-SAD minds when anticipating social encounters. Relative to non-SAD individuals, SAD subjects present with bistable social motivational tendencies and negatively biased social self-representations. When imagining or planning for future unstructured social events, the SAD mind receives interoceptive afferents signalling increased physiological response to a likely imminent threat. This reflects an interoceptive prediction error, i.e. an unexpectedly excited allostatic state relative to baseline. This prediction error between actual and anticipated bodily states creates feelings of uncertainty, fear and threat. In order to increase precision and reduce such errors, and coupled with underlying social approach-avoidance conflict and negatively biased self-models, the SAD mind regulates allostatic imbalance by generating STPs.

This maladaptive allostatic regulation of anticipatory bodily states and interoceptive prediction error plays out in the SAD brain via an interdependent network of insular and cortico-striatal-limbic functioning. In anticipation of imminent social events, the insula allows for online current and future bodily state representations, via interoceptive afferent signalling and semantic integration, allowing for a coherent predictive model of the self's ability to effect change and to cope in its social environment. Elevated anticipatory insula functioning in SAD may suggest exaggerated interoceptive signalling of future bodily states that are specific to SAD, relative to other anxiety disorders.

When anticipating social events, altered striatal activity may occur in tandem with increased interoceptive afferents from the insula to dampen expectations of experiencing or receiving positive social information. Blunted anticipatory striatal activity to impending positive external stimuli appears specific to social situations and may result from domain-specific models of biased social self-representations, likely constructed in part by impaired insula dynamics.

The amygdala, a critical limbic system hub, moderates approach-avoidance behaviour and likely processes and signals self-relevant threat in face of imminent aversive uncertain outcomes. Although sufficient data is presently lacking to draw conclusions in SAD, we can nonetheless hypothesize that SAD-related approach-avoidance motivational conflict and elevated interoceptive prediction error are likely imbued with increased affective and threat-relevance for the SAD subject via altered insula and amygdala dysfunction.

Finally, the dACC likely plays a self-regulatory function, assembling current state information with outcome value estimations, to orient the individual toward goal-relevant behaviours whilst mediating approach-avoidant tendencies arising from bottom-up neural substrates. In SAD, intrinsic bistable motivations imbue both rewarding and punishing social outcomes as self-relevant, creating increasing demands on approach-avoidance mediation and value outcome estimations, likely subserved in the dACC.

Established cognitive models of social anxiety

Our predictive processing account of social anxiety, we believe, juxtaposes nicely with prominent cognitive models of social anxiety proposed by Clark and Wells (1995) and Hofmann (2007).

The influential cognitive model of Clark and Wells emphasizes a shift to 'self-focused attention' and the reliance on internally generated information to regulate social behaviour in intrinsically unpredictable and uncertain social encounters. They further emphasize the role of 'safety behaviours' in maintaining the causes of anxiety by installing avoidant tendencies, ranging from avoiding eye contact to social withdrawal. Such behaviours make the subject appear unfriendly and disengaged at the same time as her internal focus makes both distressing feelings and maladaptive self-models highly salient. In the active inference framework, this translates as the role played by a rigid self-model in inhibiting the possibility of exploratory behaviour and cognition.

The cognitive model of social anxiety proposed by Hofmann (2007) states that social anxiety emphasizes the role of negatively biased self-models that represent the social self as ineffective and unable to cope with social threats as well as creates unrealistic standards in social situations and poorly defined social goals, among others (Hofmann 2007). This account gels with our predictive processing model, arguing for poor self-models that leave the SAD individual fearful of her ability to emotionally survive a future social event, but also poorly articulated social goals, which leave the SAD subject in a state of bistability and, ultimately, distress when attempting to regulate approach-avoidance motivations.

Testing predictions

Here, we provide a few avenues that may allow for the empirical testing of the presence of maladaptive IAI in SAD, via behavioural, cognitive and neurobiological measures.

To test the bistability of SAD goal-oriented motivations, it would behove researchers to consider investigating bistable motivational processing in SAD, in terms of behavioural approach-avoidance dynamics (e.g. reaction time, attention), neural responses to anticipating social events with an equal chance of reward and punishment. Furthermore, it would be important to investigate insula activation more closely during anticipatory processing, for both certain and uncertain future social events, measure physiological activity, such as heart rate and skin conductance and measure the degree to which the SAD individual recognizes changes in her bodily state and to what degree she deems them threatening.

Another area of potential interest would be to examine the effect of treatment on the interpretation of neutral or imprecise interoceptive and social stimuli. We might predict that SAD patients would experience and report interoceptive states as evidence of STPs. The social analogue would be the interpretation of a neutral or imprecise (e.g. blurred or ambiguous) facial expression. As in the interoceptive case, SAD individuals are predicted to interpret such stimuli as social threat and to avoid active inference such as exploratory saccades. In each case, it would be interesting to compare patterns of activation in areas of interest, particularly anterior insula, a hub of self-modelling, before and after treatment. We would predict reduced activity in AIC correlates with reduced STP and more accurate interpretation of interoceptive perturbation and social stimuli.

Conclusion

The last three decades have seen a theoretical convergence across disciplines such as psychiatry, neuroscience, psychology and philosophy arguing that basic bodily regulation is the fundamental cognitive imperative. Consequently, there is renewed emphasis on interoception as a process that underlies and unifies self-representation and emotional processing. However, disentangling the relationships between emotion, affect, body representation, self-awareness and cognition is not straightforward.

In this article, we pursued a suggestion that active inference theories of interoception can provide a unifying and clarifying perspective. In particular, interoception can be thought of as allostatic active inference and emotional processing as IAI. This conceptualizes a hierarchical processing structure ultimately grounded in basic bodily regulation.

In this framework, the affective experience is a signal of success or failure at realizing organismic goals, represented at different levels ranging from basic bodily maintenance (allostasis) to explicitly represented personal and social goals. Predictive processing theory tells us that prediction error is best minimized over the long term by a model that makes endogenous experience coherent by attributing it to a stable unified entity. A self: this self-model has different levels and dimensions of which perhaps the most important is affective self-modelling. This level of modelling attributes fluctuations in affective experience to a persisting self: the person who feels better or worse as her goals are realized or frustrated in action.

We applied this idea to the explanation of SAD, a psychiatric disorder characterized by extreme aversive response and distress at the prospect of social interaction. Superficially, SAD presents as excessive fear of social interaction. However, we argued that, while this is true, that fear is generated by a self-model that predicts uncontrollable distress when anticipating social encounters. This distress arises from imprecise interoceptive prediction error. That experience is then interpreted using self-models that represent the subject as vulnerable and unable to cope. These self-models interact with and reinforce models of the social world as intractably hostile. Finally, we showed how this idea unifies the range of clinical and neural evidence about SAD symptoms and their neural correlates and provides possible avenues for future empirical testing of our theory of IAI and self-representation in SAD.

Funding

Funding for this article was provided by Australian Research Council Discovery Grant DP190101451.

Conflict of interest statement. None declared.

References

Aderka IM, Hofmann SG, Nickerson A, *et al*. Functional impairment in social anxiety disorder. *J Anxiety Disord* 2012;**26**:393–400.

Aderka IM, Weisman O, Shahar G, *et al*. The roles of the social rank and attachment systems in social anxiety. *Pers Individ Differ* 2009;**47**:284–8.

Adolphs R, Baron-Cohen S, Tranel D. Impaired recognition of social emotions following amygdala damage. *J Cogn Neurosci* 2002;**14**:1264–74.

Ainley V, Maister L, Brokfeld J, *et al*. More of myself: manipulating interoceptive awareness by heightened attention to bodily and narrative aspects of the self. *Conscious Cogn* 2013;**22**:1231–8.

Alfano CA, Beidel DC, Turner SM. Cognitive correlates of social phobia among children and adolescents. *J Abnormal Child Psychol* 2006;**34**:182–94.

Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc Cogn Affect Neurosci* 2017;**12**:1–23.

Barrett LF, Quigley KS, Hamilton P. An active inference theory of allostasis and interoception in depression. *Philos Trans R Soc B* 2016;**371**:20160011.

Barrett LF, Simmons WK. Interoceptive predictions in the brain. *Nat Rev Neurosci* 2015;**16**:419–29.

Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 2005;**47**:129–41.

Bechara A, Damasio H, Damasio AR, *et al*. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *J Neurosci* 1999;**19**:5473–81.

Becker M, Simon D, Miltner W, *et al*. Altered activation of the ventral striatum under performance-related observation in social anxiety disorder. *Psychol Med* 2017;**47**:2502.

Bennett CM, Baird AA. The processing of internally-generated interoceptive sensation. *Neuroimage* 2009;**47**:S84.

Blackford JU, Clauss JA, Avery SN, *et al*. Amygdala–cingulate intrinsic connectivity is associated with degree of social inhibition. *Biol Psychol* 2014;**99**:15–25.

Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn Sci* 2009;**13**:7–13.

Boehme S, Miltner WH, Straube T. Neural correlates of self-focused attention in social anxiety. *Soc Cogn Affect Neurosci* 2014;**10**:856–62.

Boehme S, Ritter V, Tefikow S, *et al*. Brain activation during anticipatory anxiety in social anxiety disorder. *Soc Cogn Affect Neurosci* 2013;**9**:1413–8.

Boelen PA, Reijntjes A. Intolerance of uncertainty and social anxiety. *J Anxiety Disord* 2009;**23**:130–5.

Broyd SJ, Demanuele C, Debener S, *et al*. Default-mode brain dysfunction in mental disorders: a systematic review. *Neurosci Biobehav Rev* 2009;**33**:279–96.

Brühl AB, Rufer M, Delsignore A, *et al*. Neural correlates of altered general emotion processing in social anxiety disorder. *Brain Res* 2011;**1378**:72–83.

Campbell-Sills L, Espejo E, Ayers CR, *et al*. Latent dimensions of social anxiety disorder: a re-evaluation of the Social Phobia Inventory (SPIN). *J Anxiety Disord* 2015;**36**:84–91.

Campbell-Sills L, Simmons AN, Lovero KL, *et al*. Functioning of neural systems supporting emotion regulation in anxiety-prone individuals. *Neuroimage* 2011;**54**:689–96.

Carhart-Harris RL, Friston KJ. The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. *Brain* 2010;**133**:1265–83.

Carleton RN. The intolerance of uncertainty construct in the context of anxiety disorders: Theoretical and practical perspectives. *Expert Rev Neurother* 2012;**12**:937–47.

Carleton RN, Collimore KC, Asmundson GJ. "It's not just the judgements—It's that I don't know": intolerance of uncertainty as a predictor of social anxiety. *J Anxiety Disord* 2010;**24**:189–95.

Carleton RN, Fetzner MG, Hackl JL, *et al*. Intolerance of uncertainty as a contributor to fear and avoidance symptoms of panic attacks. *Cogn Behav Ther* 2013;**42**:328–41.

Cauda F, D'Agata F, Sacco K, *et al*. Functional connectivity of the insula in the resting brain. *Neuroimage* 2011;**55**:8–23.

Chalmers D. On the search for the neural correlate of consciousness. 1996.

Clark DM, Wells A. A cognitive model of social phobia. *Soc Phobia* 1995;**41**:00022–3.

Clauss JA, Avery SN, VanDerKlok RM, *et al*. Neurocircuitry underlying risk and resilience to social anxiety disorder. *Depress Anxiety* 2014;**31**:822–33.

Collimore KC, Asmundson GJ. Fearful responding to interoceptive exposure in social anxiety disorder. *J Anxiety Disord* 2014; **28**:195–202.

Corcoran AW, Hohwy J. Allostasis, interoception, and the free energy principle: feeling our way forward. PsyArXiv 2017 doi 31234/osf.io/zbqnx.

Craig AD. How do you feel—now? The anterior insula and human awareness. *Nat Rev Neurosci* 2009;**10**:59–70.

Cremers HR, Veer IM, Spinhoven P, *et al*. Neural sensitivity to social reward and punishment anticipation in social anxiety disorder. *Front Behav Neurosci* 2015;**8**:439.

Damasio AR. Descartes' error. Emotion reason and the human brain. Random House New York. 2006.

Damasio A, Dolan RJ. The feeling of what happens. *Nature* 1999; **401**:847.

Davies CD, Young K, Torre JB, *et al*. Altered time course of amygdala activation during speech anticipation in social anxiety disorder. *J Affect Disord* 2017;**209**:23–29.

Dodhia S, Hosanagar A, Fitzgerald DA, *et al*. Modulation of resting-state amygdala-frontal functional connectivity by oxytocin in generalized social anxiety disorder. *Neuropsychopharmacology* 2014;**39**:2061–2069.

Domschke K, Stevens S, Pfleiderer B, *et al*. Interoceptive sensitivity in anxiety and anxiety disorders: an overview and integration of neurobiological findings. *Clin Psychol Rev* 2010;**30**:1–11.

Drabant EM, Kuo JR, Ramel W, *et al*. Experiential, autonomic, and neural responses during threat anticipation vary as a function of threat intensity and neuroticism. *Neuroimage* 2011;**55**: 401–410.

Duval ER, Joshi SA, Block SR, *et al*. Insula activation is modulated by attention shifting in social anxiety disorder. *J Anxiety Disord* 2018;**56**:56–62.

Etkin A, Wager TD. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *Am J Psychiatry* 2007;**164**:1476–1488.

Foa EB, Franklin ME, Perry KJ, *et al*. Cognitive biases in generalized social phobia. *J Abnormal Psychol* 1996;**105**:433–9.

Friston K, Kiebel S. Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B: Biol Sci* 2009;**364**:1211–1221.

Friston K, Schwartenbeck P, FitzGerald T, *et al*. The anatomy of choice: active inference and agency.*Philosophical Transactions of the Royal Society B: Biological Sciences* 369, no. 1655 (2014): 20130481

Furmark T. Neurobiological aspects of social anxiety disorder. *Israel J Psychiatry Relat Sci* 2009;**46**:5.

Garfinkel SN, Critchley HD. Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on: "Anterior insular cortex mediates bodily sensibility and social anxiety" by Terasawa et al. (2012). *Soc Cogn Affect Neurosci* 2013;**8**:231–234.

Gasquoine PG. Contributions of the insula to cognition and emotion. *Neuropsychol Rev* 2014;**24**:77–87.

Gerrans P. *All the Self We Need: Open MIND*. Frankfurt am Main: MIND Group, 2014.

Gilboa-Schechtman E, Franklin ME, Foa EB. Anticipated reactions to social events: differences among individuals with generalized social phobia, obsessive compulsive disorder, and non-anxious controls. *Cogn Ther Res* 2000;**24**:731–746.

Gu X, Fitzgerald TH. Interoceptive inference: homeostasis and decision-making. Trends in Cognitive Sciences 2014;**18**:269–70. 10.1016/j.tics.2014.02.001

Gusnard DA, Akbudak E, Shulman GL, *et al*. Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proc Natl Acad Sci USA* 2001;**98**: 4259–4264.

Guex R, Méndez-Bértolo C, Moratti S, *et al*. Temporal dynamics of amygdala response to emotion-and action-relevance. *Sci Rep* 2020;**10**:1–16.

Guyer AE, Lau JY, McClure-Tone EB, *et al*. Amygdala and ventrolateral prefrontal cortex function during anticipated peer evaluation in pediatric social anxiety. *Arch Gen Psychiatry* 2008;**65**: 1303–1312.

Hayden BY, Heilbronner SR, Pearson JM, *et al*. Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J Neurosci* 2011;**31**:4178–4187.

Heilbronner SR, Hayden BY. Dorsal anterior cingulate cortex: a bottom-up view. *Annu Rev Neuroscience* 2016;**39**:149–170.

Heinrichs N, Hofmann SG. Information processing in social phobia: a critical review. *Clin Psychol Rev* 2001;**21**:751–770.

Hofmann SG. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. *Cogn Behav Ther* 2007;**36**:193–209.

Hohwy J. *The Predictive Mind*. Oxford: Oxford University Press, 2013.

Hohwy J, Michael J. Why should any body have a self? In: *The Subject's Matter: Self-Consciousness and the Body*. 2017, 363.

Hohwy J, Seth A. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. PsyArXiv doi 10.31234/osf.io/nd82g.2020.

Hope DA, Rapee RM, Heimberg RG, *et al*. Representations of the self in social phobia: vulnerability to social threat. *Cogn Ther Res* 1990;**14**:177–189.

Iancu I, Bodner E, Ben-Zion IZ. Self esteem, dependency, self-efficacy and self-criticism in social anxiety disorder. *Compr Psychiatry* 2015a;**58**:165–171.

Iancu I, Lupinsky Y, Barenboim D. Negative and positive automatic thoughts in social anxiety disorder. *Israel J Psychiatry Relat Sci* 2015b;**52**:129.

Isomura Y, Ito Y, Akazawa T, *et al*. Neural coding of "attention for action" and "response selection" in primate anterior cingulate cortex. *J Neurosci* 2003;**23**:8002–8012.

Joffily M, Coricelli G. Emotional valence and the free-energy principle. *PLoS Comput Biol* 2013;**9**:e1003094.

Kalisch R. The functional neuroanatomy of reappraisal: time matters. *Neurosci Biobehav Rev* 2009;**33**:1215–1226.

Kalisch R, Wiech K, Critchley HD, *et al*. Levels of appraisal: a medial prefrontal role in high-level appraisal of emotional material. *Neuroimage* 2006;**30**:1458–1466.

Kambouropoulos N, Egan S, O'Connor EJ, *et al*. Escaping threat. *J Individ Differ*35(1), 47-53 2014.

Kawaguchi A, Nemoto K, Nakaaki S, *et al*. Insular volume reduction in patients with social anxiety disorder. *Front Psychiatry* 2016;**7**:3.

Kirchhoff M, Parr T, Palacios E, *et al*. The Markov blankets of life: autonomy, active inference and the free energy principle. *J R Soc Interface* 2018;**15**:20170792.

Kleckner IR, Zhang J, Touroutoglou A, *et al*. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nat Hum Behav* 2017;**1**:0069.

Klumpp H, Angstadt M, Phan KL. Insula reactivity and connectivity to anterior cingulate cortex when processing threat in generalized social anxiety disorder. *Biol Psychol* 2012;**89**:273–276.

Klumpp H, Fitzgerald DA, Cook E, *et al*. Serotonin transporter gene alters insula activity to threat in social anxiety disorder. *Neuroreport* 2014;**25**:926.

Knutson B, Cooper JC. Functional magnetic resonance imaging of reward prediction. *Curr Opin Neurol* 2005;**18**:411–417.

Knutson B, Fong GW, Adams CM, *et al*. Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport* 2001;**12**:3683–3687.

Koban L, Pourtois G. Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neurosci Biobehav Rev* 2014;**46**:71–84.

Letheby C, Gerrans P. Self unbound: ego dissolution in psychedelic experience. *Neurosci Conscious* 2017;**2017**: 1–11.

Liao W, Qiu C, Gentili C, *et al*. Altered effective connectivity network of the amygdala in social anxiety disorder: a resting-state FMRI study. *PLoS One* 2010;**5**:e15238.

Limanowski J, Blankenburg F. Minimal self-models and the free energy principle. *Front Hum Neurosci* 2013;**7**:547.

Lorberbaum JP, Kose S, Johnson MR, *et al*. Neural correlates of speech anticipatory anxiety in generalized social phobia. *Neuroreport* 2004;**15**:2701–2705.

Lucock MP, Salkovskis PM. Cognitive factors in social anxiety and its treatment. *Behav Res Ther* 1988;**26**:297–302.

Maier SU, Makwana AB, Hare TA. Acute stress impairs self-control in goal-directed choice by altering multiple functional connections within the brain's decision circuits. *Neuron* 2015; **87**:621–631.

Medford N, Critchley HD. Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response. *Brain Struct Funct* 2010;**214**:535–549.

Mee S, Bunney BG, Reist C, *et al*. Psychological pain: a review of evidence. *J Psychiatric Res* 2006;**40**:680–690.

Meffert H, Brislin SJ, White SF, *et al*. Prediction errors to emotional expressions: the roles of the amygdala in social referencing. *Soc Cogn Affect Neurosci* 2015;**10**:537–544.

Mellings TM, Alden LE. Cognitive processes in social anxiety: the effects of self-focus, rumination and anticipatory processing. *Behav Res Ther* 2000;**38**:243–257.

Mills AC, Grant DM, Judah MR, *et al*. The influence of anticipatory processing on attentional biases in social anxiety. *Behav Ther* 2014;**45**:720–729.

Morriss J, Gell M, van Reekum CM. The uncertain brain: a coordinate based meta-analysis of the neural signatures supporting uncertainty during different contexts. *Neurosci Biobehav Rev* 2019;**96**:241–249.

Moutoussis M, Fearon P, El-Deredy W, *et al*. Bayesian inferences about the self (and others): a review. *Conscious Cogn* 2014;**25**: 67–76.

Moayedi M. All roads lead to the insula. *Pain* 2014;**155**:1920–1921.

Murray RJ, Apazoglou K, Celen Z, *et al*. Maladaptive emotion regulation traits predict altered corticolimbic recovery from psychosocial stress. *J Affective Disord* 2021;**280**:54–63.

Murray RJ, Debbané M, Fox PT, *et al*. Functional connectivity mapping of regions associated with self-and other-processing. *Hum Brain Mapp* 2015a;**36**:1304–1324.

Murray RJ, Gerrans P, Brosch T, *et al*. When at rest: "event-free" active inference may give rise to implicit self-models of coping potential. *Behav Brain Sci* 2015b;**38**:41–42.

Murray RJ, Schaer M, Debbané M. Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neurosci Biobehav Rev* 2012;**36**:1043–1059.

Pannekoek JN, Veer IM, van Tol M-J, *et al*. Resting-state functional connectivity abnormalities in limbic and salience networks in social anxiety disorder without comorbidity. *Eur Neuropsychopharmacol* 2013;**23**:186–195.

Paulus MP, Stein MB. An insular view of anxiety. *Biol Psychiatry* 2006;**60**:383–387.

Paulus MP, Stein MB. Interoception in anxiety and depression. *Brain Struct Funct* 2010;**214**:451–463.

Pessiglione M, Seymour B, Flandin G, *et al*. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 2006;**442**:1042.

Pezzulo G, Rigoli F, Friston K. Active inference, homeostatic regulation and adaptive behavioural control. *Progr Neurobiol* 2015; **134**:17–35.

Ribas-Fernandes JJ, Shahnazian D, Holroyd CB, *et al*. Subgoal-and goal-related reward prediction errors in medial prefrontal cortex. *J Cogn Neurosci* 2019;**31**:8–23.

Richey JA, Ghane M, Valdespino A, *et al*. Spatiotemporal dissociation of brain activity underlying threat and reward in social anxiety disorder. *Soc Cogn Affect Neurosci* 2017;**12**: 81–94.

Ruff CC, Fehr E. The neurobiology of rewards and values in social decision making. *Nat Rev Neuroscience* 2014;**15**:549.

Sander D, Grafman J, Zalla T. The human amygdala: an evolved system for relevance detection. *Rev Neurosci* 2003;**14**: 303–316.

Schmidt S, Mohr A, Miltner WH, *et al*. Task-dependent neural correlates of the processing of verbal threat-related stimuli in social phobia. *Biol Psychol* 2010;**84**:304–312.

Sel A. Predictive codes of interoception, emotion, and the self. *Front Psychol* 2014;**5**:189.

Seth AK. Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci* 2013;**17**:565–573.

Seth AK, Tsakiris M. Being a beast machine: the somatic basis of selfhood. *Trends Cogn Sci* 2018;**22**:969–981.

Seth AK, Suzuki K, Critchley HD. An interoceptive predictive coding model of conscious presence. *Front Psychol* 2012;**2**:395.

Seth AK, Friston KJ. Active interoceptive inference and the emotional brain. *Philos Trans R Soc B* 2016;**371**:20160007.

Shah SG, Klumpp H, Angstadt M, *et al*. Amygdala and insula response to emotional images in patients with generalized social anxiety disorder. *J Psychiatry Neurosci* 2009;**34**:296.

Shenhav A, Botvinick MM, Cohen JD. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* 2013;**79**:217–240.

Singer T, Critchley HD, Preuschoff K. A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci* 2009;**13**: 334–340.

Spreng RN, Mar RA, Kim ASN. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *J Cogn Neurosci* 2009;**21**:489–510.

Spurr JM, Stopa L. Self-focused attention in social phobia and social anxiety. *Clin Psychol Rev* 2002;**22**:947–975.

Sridharan D, Levitin DJ, Menon V. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc Natl Acad Sci USA* 2008;**105**: 12569–12574.

Stalnaker TA, Calhoon GG, Ogawa M, *et al*. Reward prediction error signaling in posterior dorsomedial striatum is action specific. *J Neurosci* 2012;**32**:10296–10305.

Stephan KE, Manjaly ZM, Mathys CD, *et al*. Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front Hum Neurosci* 2016;**10**:550.

Stopa L, Brown MA, Luke MA, *et al*. Constructing a self: the role of self-structure and self-certainty in social anxiety. *Behav Res Therapy* 2010;**48**:955–965.

Stopa L, Jenkins A. Images of the self in social anxiety: effects on the retrieval of autobiographical memories. *J Behav Therapy Exp Psychiatry* 2007;**38**:459–473.

Suddendorf T, Corballis MC. The evolution of foresight: what is mental time travel, and is it unique to humans? *Behav Brain Sci* 2007;**30**:299–313.

Tang GSM, van den Bos W, Andrade E, *et al*. Social anxiety modulates risk sensitivity through activity in the anterior insula. *Front Neurosci* 2012;**5**:142.

Tanovic E, Gee DG, Joormann J. Intolerance of uncertainty: neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clin Psychol Rev* 2018;**60**:87–99.

Terasawa Y, Shibata M, Moriguchi Y, *et al*. Anterior insular cortex mediates bodily sensibility and social anxiety. *Soc Cogn Affect Neurosci* 2013;**8**:259–266.

Tsakiris M, De Preester H. *The Interoceptive Mind: From Homeostasis to Awareness*. Oxford University Press: Oxford, 2018.

Van de Cruys S. Affective value in the predictive mind. Frankfurt am Main: MIND Group, 2017.

Velasco PF, Loev S. Affective experience in the predictive mind: a review and new integrative account. *Synthese* 2020;1–36.

Von Mohr M, Fotopoulou A. The cutaneous borders of interoception: active and social inference of pain and pleasure on the skin. In: Tsakiris, M and De Preester H.*The Interoceptive Mind: From Homeostasis to Awareness*. Oxford: OUP 2018, 102.

Wang W, Zhornitsky S, Li CS-P, *et al*. Social anxiety, posterior insula activation, and autonomic response during self-initiated action in a Cyberball game. *J Affect Disord* 2019;**255**:158–167.

Weisman O, Aderka IM, Marom S, *et al*. Social rank and affiliation in social anxiety disorder. *Behav Res Ther* 2011;**49**:399–405.

Wilson JK, Rapee RM. Interpretative biases in social phobia: content specificity and the effects of depression. *Cogn Ther Res* 2005;**29**:315–331.

Wilson JK, Rapee RM. Self-concept certainty in social phobia. *Behav Res Ther* 2006;**44**:113–136.

Wong QJ, Moulds ML. The relationship between the maladaptive self-beliefs characteristic of social anxiety and avoidance. *J Behav Ther Exp Psychiatry* 2011;**42**:171–178.

Wong QJ, Rapee RM. The aetiology and maintenance of social anxiety disorder: a synthesis of complementary theoretical models and formulation of a new integrated model. *J Affect Disord* 2016;**203**:84–100.

Yaniv D, Desmedt A, Jaffard R, *et al*. The amygdala and appraisal processes: stimulus and response complexity as an organizing factor. *Brain Res Rev* 2004;**44**:179–186.