PhD Thesis

Interoception in Social Cognition under Uncertainty: A Computational and Psychophysiological Approach

Vassilis Kotsaris

February 2025

This thesis is submitted in fulfilment of the requirements for a degree of Doctor of Philosophy in Psychology at University of Kent.

University of Kent

Declaration of Authorship

I hereby declare that the work presented in this thesis is my own. All work and materials that are drawn from others are always clearly attributed.

Vassilis Kotsaris

Acknowledgments

Firstly, I would like to express my deepest gratitude to my supervisor, Dr. Ruben Azevedo, for granting me the opportunity to undertake this thesis and for his continuous, invaluable support to complete this endeavour.

I should also like to thank the School of Psychology at the University of Kent for the financial support and the collaborative working environment provided as well as the technical support team whose assistance allowed us to conduct our experiments with fewer concerns.

A sincere thank you to Dr. Dimitris Bolis and Dr. Alex Sel for their help in data analysis for the studies in Chapter 3 and 4.

My PhD work at the University of Kent, however, was not only an intellectual journey but also a place where I met wonderful people, particularly Laura Hunt, Kimberly Sequeira, Chiara Zazzarino, Aarohi Khare, who were a vital source of support and unpretentious fun.

I would also like to thank all my friends for being there during both the difficult and good times throughout this process. Lastly, I want to express my heartfelt gratitude to Myrto Karavakou for her priceless presence in this endeavour.

Abstract

Interoception - the perception and monitoring of physiological states - plays a critical role in affective processes, intuitive decision-making, and social interactions. However, how interoceptive processing influences emotional and social processes under uncertainty remains empirically underexplored. Within a predictive processing framework, this dissertation combined behavioural, computational, and electrophysiological methods to investigate how individuals integrated interoceptive and external cues to infer others' emotions and adapt empathic responses in dynamic and uncertain environments.

In the first study (Chapter 2), questionnaires and behavioural experiments were employed, along with Structural Equation Modelling and Diffusion Drift Modelling, to explore if intolerance of uncertainty (IU) and alexithymia mediated the relationship between interoceptive sensibility and anxiety, as well as empathic distress and emotion perception. The second study (Chapter 3) further investigated how emotion perception is influenced by interpersonal contextual uncertainty. Specifically, it explored how emotion egocentricity biases in emotion recognition are modulated by learned interpersonal emotion contingencies and perceptual ambiguity. Computational modelling using the Hierarchical Gaussian Filter (HGF) illustrated how individuals adapted their beliefs about interpersonal emotional congruency, with learning trajectories modulated by individual differences in interoceptive sensibility and physiological reactions.

In the third study (Chapter 4), we measured heart-evoked potentials (HEPs), a cortical index of cardiac interoceptive processing, to probe interoceptive predictive processing during an adaptive empathy task. Results indicated that HEP amplitudes tracked precision-weighted prediction errors during social feedback, while showing that interoceptive modulations were associated with autistic traits. HEPs also predicted empathic accuracy during decision-making in this task. Lastly, in chapter 5, transcutaneous vagus nerve stimulation was shown to modulate adaptive empathic learning, providing causal evidence for the role of afferent vagal input in social interactions.

This research advances the understanding of how interoception shapes social and emotional functioning. It uncovers neural, cognitive, and physiological mechanisms underlying empathy and interoceptive processing using computational approaches. These findings have significant implications for our understanding of the role of interoceptive processing on social perception under uncertainty and, potentially, for the development of targeted interventions to improve emotional resilience and social functioning in clinical and non-clinical populations. This thesis enriches emerging perspectives of social cognition as an embodied anticipatory process, where interoception and uncertainty dynamically shape how we perceive, learn, and adapt to the social world.

Word count: 64431

Number of pages: 241

Table of Contents

List of Tables

List of figures

Chapter 1. General Introduction

1.1. Social cognition under uncertainty

1.1.1. Empathy and emotion perception

Interpreting and predicting the emotions and behaviour of others lies at the heart of human social interaction, forming the foundation for empathy, cooperation, and effective communication. Despite its significance, the affective, cognitive and neural mechanisms underlying the ability to empathetically understand others remain a central and contentious topic in cognitive neuroscience. Most scholars define empathy as the capacity to perceive and share another's emotional state while maintaining awareness of the self-other distinction (Singer & Klimecki, 2014; Cuff et al., 2016). Another approach is to view empathy as an umbrella term encompassing distinct yet interrelated social cognition functions, each involving varying degrees of affect sharing and overlapping self-other boundaries. From this perspective, Zaki and Ochsner (2012) propose the following empathic processes: experience sharing, prosocial concern and mentalizing. Experience sharing, also referred to as emotion contagion, describes the 'vicarious sharing of the target's internal states', entailing extended overlap between self and other's affective experiences. Prosocial concern, also known as empathic concern or compassion, also pertains to affective empathy, but without necessarily involving sharing other's emotions. Instead, it describes 'a motivation to improve the target's experience (e.g. alleviate their distress or suffering). Lastly, mentalizing, refers to cognitive empathy functions, involving explicit inferential reasoning, often described in terms of perspective-taking and theory of mind, to explain the target's mental states and behaviour.

Importantly, prosocial behaviour, and particularly goal directed like compassion, does not only involve regulation of the target's emotional state but also requires self-regulation to remain effective and sustainable rather than psychologically and physically depleting (Hunt et al., 2017; Weilenmann et al., 2018). Therefore, interpersonal emotion regulation has been increasingly recognised as a critical aspect of empathy (Zaki, 2020; Thompson et al., 2019). While research into these prosocial functions has surged in recent years, the underlying mechanisms remain not clearly understood, particularly in the context of real-world, dynamic, interactive contexts.

Most models and experimental paradigms conceive and examine emotion perception and empathy as largely static phenomena, but effective and flexible prosocial behaviour often demands dynamic adaptation to others' shifting emotional needs. To address this, Shamay-Tsoory and Hertz (2022) developed a reinforcement learning (RL) model of empathy and to test it they introduced an adaptive empathy task (Arbel et al. (2021), in which individuals learn and refine their empathic responses based on feedback from the target. Specifically, participants should select between two emotion regulation strategies (reappraisal or distraction) to alleviate another's distress, receiving feedback about the strategy's effectiveness. Their results showed that participants optimised their empathic responses over time, with learning performance correlating with cognitive but not affective empathy, suggesting that mentalizing plays an important role in adaptive compassionate behaviour. These findings indicate an interactive, learning-based nature of empathic concern as individuals update their social predictions to provide effective support.

Emotion recognition, a critical precursor to empathy, is a similarly complex, context-sensitive process requiring the decoding of dynamic and ambiguous social cues, such as facial expressions, vocal tone or body language. Emotions are not universally conveyed through fixed cues, such as facial configurations; rather, contextual factors such as situational dynamics, cultural norms, personality, and past experiences further shape how emotions are expressed and interpreted (Barrett et al., 2011; Gross et al., 2000; Jack et al., 2013; 2014). Thus, to assess someone's affective state, the observer needs to integrate information from various sources, each associated with some degree of uncertainty. For example, a frown may signal sadness, concentration, or even intimidation and subtle facial muscle contractions can reflect fleeting emotional shifts. Using contextual factors to resolve perceptual uncertainty can be helpful in some occasions but it may also lead to the frame problem as there is an infinite number of contextual factors that could influence an emotional interpretation (Wilkerson, 2021), making it challenging to determine which cues are most relevant in any given social interaction. Therefore, emotional expressions, and particularly facial ones, are characterized with high variability resulting in high contextual and perceptual uncertainty (Barrett et al., 2013; Hassin et al., 2013).

The challenges of emotion perception can also be casted in terms of reverse inference uncertainty, where an observed cue, like a facial expression, is used to infer an underlying emotional state (Saxe & Houlihan, 2017). However, this process is inherently uncertain, as multiple causes can produce the same expression. Social interactions introduce further layers of uncertainty (FeldmanHall & Shenhav, 2019), including intention uncertainty (perceived intentions of others), action uncertainty (selecting the best response), reward uncertainty (predicting valence of outcomes) and volatility (how environmental contingencies change over time). The implications of these uncertainties are two-fold: while the context-dependent and temporally extended nature of socio-affective expressions imbues them with social meaning, their interconnections pose inferential and behavioural challenges, as perceptual uncertainty about a stimulus compounds uncertainty about its outcomes and, ultimately, the optimal course of action.

Therefore, various kinds of uncertainties, or else unknowns, are present in any social interaction. These social unknowns are critical for successfully and safely navigating the world (FeldmanHall & Shenhav, 2019), while inability to reduce social uncertainty can contribute to social and mental health difficulties (Godinić & Obrenovic, 2020; Smith et al., 2020). Next, I will briefly discuss the various affective reactions to uncertainty and theoretical frameworks that could address how we resolve social ambiguity to understand others' emotions, intentions and desires.

1.1.2. Affective reactions to uncertainty

High uncertainty usually triggers negative affect, manifesting as worry, anxiety and fear (Anderson et al., 2019). These responses stem from the brain's sensitivity to unpredictability, as uncertain outcomes might hide potential threats (Radoman et al., 2019). Two theoretical models related to psychological uncertainty, namely The Entropy Model of Uncertainty (Hirsh et al., 2012) and the Behavioural Inhibition System Theory (Gray & McNaughton, 2000), suggest that uncertainty induces perceptual and behavioural conflicts, triggering heightened arousal and anxiety to facilitate adaptive responses such as vigilance or avoidance. Furthermore, Anderson & colleagues (2019) argue that there is a cognitive bias toward simulating negative outcomes under uncertainty, as these are evolutionarily prioritised to avoid danger. This negativity bias amplifies the perception of threat and fuels negative affect (Baumeister et al., 2001).

Yet, uncertainty is not always aversive. In controlled or low-stakes situations, uncertainty can evoke positive affect, manifesting as curiosity or excitement (Anderson et al., 2019; Vazard, 2024). Uncertain outcomes in games, sports, or storytelling enhance engagement and enjoyment. This occurs when the brain perceives uncertainty as an opportunity for exploration rather than a threat (Anderson et al., 2019; Kurtz et al., 2007). These positive emotions help people overcome boredom and acquire more knowledge leading to better chances to reduce uncertainty in the long run. Uncertainty might also be favourable in anticipation of a negative outcome. Interestingly, though, evidence suggests that most people would rather prefer receiving a certain painful stimulation than enduring the anticipation and unpredictability of whether they will receive it or not (Pavy et al., 2024).

While uncertainty is typically perceived negatively, individuals differ in the degree of their aversive responses to it. This disposition is called intolerance of uncertainty (IU) and describes a cognitive and emotional trait reflecting heightened sensitivity to ambiguity and an exaggerated perception of uncertainty as threatening, regardless of the actual likelihood or severity of potential outcomes (Anderson et al., 2019; Carleton; 2016). Associated with increased worry towards uncertainty, IU is considered a transdiagnostic factor linked to several anxiety disorders (Carleton, 2012). For individuals with generalized anxiety, this intolerance drives excessive worry and rumination, as they struggle to tolerate the unpredictable aspects of daily life, such as unforeseen changes or ambiguous decisions. In social anxiety, IU is particularly challenging due to the inherently ambiguous nature of social interactions (Carleton et al., 2010). Socially anxious individuals often face uncertainty about how they are perceived by others, whether their behaviour will be judged, or how others will respond to their actions. This ambiguity amplifies fear of negative evaluation and rejection, reinforcing hypervigilance to social cues, like facial expressions or tone of voice. These affective reactions hinder emotion regulation leading to avoidance behaviours, such as declining social invitations or minimising interactions, further preventing corrective learning, leaving IU and anxiety intact (Jazaieri et al., 2015). Thus, improving tolerance for ambiguity may be critical for reducing anxiety and enhancing emotional resilience in uncertain social environments.

1.1.3. Theoretical Perspectives on Social Cognition

Simulation Theory

Simulation Theory (ST) posits that we can use our own mind as a model to understand the minds of others. Specifically, we simulate others' mental states as if we were in their situation or as if we were them and then attribute these pretend states to them (Gordon, 1995; Shanton & Goldman, 2010). Importantly, this also applies to the automatic and unconscious neural and bodily simulations of others' actions and affective states (Gallese, 2014). Support for ST initially came from the discovery of the ''mirror neuron system'' which refers to matching neural activations during action observation and execution (Gallese, & Goldman, 1998). However, this view was soon extended beyond action perception to suggest that (all) social understanding is grounded in embodied simulation (Gallese et al., 2004; 2009). Indeed, there is now extensive evidence for overlapping neural activations when individuals experience an emotion and observe the same emotion in others (Lamm et al., 2008; Singer & Lamm, 2009). For instance, the anterior insula, anterior cingulate cortex (ACC), and cerebellum are activated when both experiencing and witnessing pain in another individual (Singer et al., 2004). Similar patterns of co-activation are observed in behavioural studies reporting various phenomena of bodily mimicry and imitation (Heyes, 2011; Prochazkova & Kret, 2017). For instance, facial mimicry occurs automatically when perceiving an emotional facial expression (Sato, & Yoshikawa, 2007; Wood et al., 2016), while disruption of relevant muscle activity hinders emotion recognition (Oberman et al., 2007). This process is involuntary, partial, and occurs at a subthreshold neural level, generating sensorimotor feedback that activates associated emotional components (Wood et al., 2016). The final step involves attributing the simulated experience to the observed target. Importantly, it is argued that embodied simulation can extend beyond sensorimotor re-activation, encompassing the entrainment of related internal bodily states (Gallese, 2014). This mirroring phenomena have been considered key to social cognition, supporting a wide range of prosocial behaviours (Niedenthal et al., 2005; Van Baaren et al., 2004). However, this perspective has faced substantial criticism, since, among others, raises the question of how we avoid egocentric biases - if understanding others primarily relies on projecting one's own thoughts and emotions (Zahavi, 2008), it remains unclear how we accurately perceive others and effectively respond to their needs when there is interpersonal incongruency.

Associative Learning Theory

There are, however, alternative perspectives in social cognition research (Gallagher, 2001; Gangopadhyay, 2014; Heyes, 2010; Heyes & Catmur, 2022; Zahavi, 2011a), such as the Associative Learning Theory or Learn Matching Hypothesis (LMH), according to which, mirroring does not originate from innate modules of social brain, but is rather shaped by domain-general mechanisms of associative learning, with evolution playing a background role (Cook et al, 2014; Heyes, 2018). In the affective domain, these matching behavioural and neural activations are forged by correlating observed and experienced emotional states. That is, people match observed exteroceptive states (emotional expressions) with felt bodily states (physiological responses). These associations are created during early developmental ages and remain are malleable through the whole life, forming bidirectional excitatory social links, based on both contiguity (spatiotemporal closeness) and contingency (predictability).

At the neural level, these associations are postulated to be implemented by Hebbian learning (Cook et al., 2014; Keysers & Gazzola, 2014). Heyes (2018) argues that these learning mechanisms suffice to explain the emergence of automatic interpersonal processes of affect mirroring and emotional contagion that underpin primary empathic responses, while can also account for the flexibility and context-dependency as they are forged by socio-cultural regularities. The plasticity of these unintentional empathic reactions, like in emotional contagion, is reported in an early study demonstrating increased psychophysiological responses after congruent training and decreased after incongruent training (Englis et al., 1982). Similar training effects have been also reported for the sensorimotor mirror-like neurons (Calvo-Merino et al., 2006) which after training could also acquire counter-mirror properties (Catmur et al., 2011). However, similar studies in the domain of affect sharing are scarce.

Interaction Theory

The interaction theory (IT) expands embodied accounts arguing that we can 'directly' perceive others' emotions through bodily expressions unfolding in embodied social interactions, framing empathy as an active, participatory process of sense-making (De Jaegher, 2009; Gallagher, 2001; 2007). In this view, social meaning is co-constructed, framed and embedded within shared social contexts, thereby addressing both the frame problem and the opacity of other minds. They argue that since emotions are integrative reverberations of the whole lived body in relation to environmental conditions (Fuchs, 2017; Fuchs & Koch, 2014), we can have partial perceptual access to others' emotions much like (though not identically to) the perception of physical objects (Zahavi, 2011b). Importantly perception is conceived as a 'rich' process, intertwined with interpersonal affective engagement, social context and past experience (De Bruin, & Strijbos, 2015; Zahavi, 2011b). While the perception of inanimate objects and living beings share perceptual and cognitive faculties, they differ in three key ways: the affective responses they elicit, the interactive affordances they involve that inform social understanding, and the complexity and variability of uncertainty involved. Beyond theoretical discourses, this framework has been increasingly influential in social cognition research inspiring new experimental paradigms, particularly within second-person neuroscience (Redcay & Schilbach, 2019; Bolis et al., 2023), and was recently formulated in the context of active inference (Lehmann et a., 2023).

1.1.4. Predictive processing

A formal approach for describing how social inference is performed by the integration of multimodal sensory information to resolve uncertainty is provided by Predictive Processing (PP) and the Active Inference frameworks (Friston & Kiebel, 2009; Parr et al., 2022; Gallagher & Allen, 2018). Accounting for past experiences, top-down effects, and uncertainty estimations, this computational perspective explains how we generate inferential probabilistic predictions to explain the internal and external worlds. PP describes cognition as a process of prediction error minimisation, with the brain constructing probabilistic models to infer the causes of sensory input and actively anticipate future events. These generative models operate hierarchically, with low-level sensory data informing higher-level representations, which encode more abstract structures of information, such as intentions, emotions, and social norms, which, in turn, issue predictions to lower representation levels. When incoming sensory input deviates from prior expectations, prediction errors emerge, prompting the brain to either update its internal model (perceptual inference) or change the external world through agentive action (active inference).

This process is mathematically grounded in Bayesian inference, where prior beliefs (formed through learning and evolution) and sensory data are weighted according to their precision (reliability) to optimise predictions. Beliefs are represented as probability distributions, with variance (inverse precision) reflecting uncertainty. The brain dynamically assigns greater weight to more precise information while discounting unreliable inputs, allowing for adaptive updates to perception and behaviour. This dynamic weighting allows the brain to adjust its predictions and responses to minimise free energy, an information-theoretic measure of surprise or entropy, which is also an alternative measure of prediction error (Friston & Kiebel, 2009). Free energy minimization ensures that the organism operates within predictable and homeostatically accepted states, reducing the likelihood of encountering harmful or unexpected sensory inputs, thereby maintaining a balance between internal consistency and external adaptability (Allen & Friston, 2018).

PP thus provides a computational account of how we make sense of us and others (Isomura et al., 2019; Moutoussis et al., 2014), highlighting how uncertainty is managed through hierarchical belief updating (Clark, 2015; Hohwy, 2017). Since social interactions are inherently ambiguous, individuals attempt to resolve uncertainty by integrating sensory cues, contextual priors, and internal affective states (FeldmanHall & Shenhav, 2019; Hassin et al., 2013; Gerdon & Barrett, 2018). The extent to which priors or sensory input dominate perception depends on precision-weighting, which dynamically adjusts expectations and responses.

This framework can accommodate several top-down effects in perception and social cognition (Hassin et al., 2013; Melloni et al., 2014; Schurz et al., 2021), with evidence suggesting that such effects affect the perceptual content, even from low-level sensory processing stages (Adams & Kveraga, 2015; Marchi & Newen, 2015; Otten et al., 2017). Additionally, the variability in social inference reflects differences in prior expectations, learning history, and cognitive constraints. Social priors, formed through experience and cultural exposure, serve as anchors that reduce uncertainty but also introduce biases (FeldmanHall & Shenhav, 2019). Stereotyping, for instance, emerges when priors are over-relied upon, leading to rigid or distorted social predictions (McGovern & Otten, 2024). Importantly, these biases are even more pronounced under cognitively loaded conditions as neural and behavioural modulations are more restricted (Lieber et al., 2019; Wigboldus et al., 2004).

Interpersonal synchronisation and imitation can also be viewed as acts of active inference, where individuals entrain affective states to reduce uncertainty in dyadic or group interactions (Bolis et al.,2023; Quadt, 2017). Similarly, empathy can be described as an adaptive interpersonal emotion regulation process, where individuals predict the affective impact of their own actions on others and themselves, continuously updating responses based on social feedback (Schurz et al., 2021; Shamy-Tsoory & Hertz, 2022). In this context, affective precision-weighting modulates the prioritisation of social cues, ensuring that emotionally relevant stimuli receive greater processing resources. Such mechanisms illustrate how social cognition is fundamentally embodied, interactive and predictive rather than observation-driven simulation (Bolis et al.,2023; De Bruin & Michael, 2021; Gallagher & Allen, 2018).

Computational models of predictive processing and Bayesian learning have been applied to neurophysiological studies of social cognition, revealing neural populations involved in uncertainty estimation and social adaptation (Diaconescu et al., 2017; Keysers et al., 2024).

For example, the anterior cingulate cortex (ACC) has been found to play a crucial role in processing prediction errors and volatility in both reward-based and non-reward-based social learning (Behrens, et al., 2008; Iglesias et al., 2013). Furthermore, dopaminergic circuits, long associated with reward prediction errors, also respond to violations of social expectations, suggesting that reward-based learning and social belief updating share common computational mechanisms (Diaconescu et al., 2017; Joiner et al., 2017). These findings support the view that social learning is fundamentally a process of prediction error minimisation, where beliefs about others' emotions, intentions, and behaviours are continuously refined through hierarchical Bayesian updating.

PP offers a formal framework for understanding social cognition as an uncertainty-driven process, integrating top-down predictions, sensory evidence, and affective- bodily modulations (De Bruin & Michael, 2021; Lehmann et al., 2024). This way, individuals navigate complex social environments, dynamically adjusting perceptual and behavioural responses to minimize surprise.

1.2. Interoceptive inference and (social) cognition

Cognitive neuroscience has traditionally focused on how exteroceptive information, sensory signals from the external world, is sensed and integrated to shape perception and cognition, largely overlooking the critical role of ascending interoceptive signals from the internal bodily milieu. However, relatively recently, the emerging cognitive neuroscience of interoception highlights the different ways the internal state of the body influences our engagement with the world (Quigley et al., 2021; Tsakiris & Critchley, 2016). Despite the different perspectives and definitions, most accounts currently converge to describe interoception as the processes by which the nervous system senses, interprets, integrates and regulates signals originating from within the body, providing a real-time mapping of the body's internal states, encompassing conscious and unconscious processes (Desmedt et al., 2024). While initially interoception referred to perception of visceral processes (Ceunen et al., 2016), currently, interoceptive systems are now considered to extend beyond visceroception, encompassing autonomic, hormonal, visceral, and immune system signals, along with pain and affective touch (Berntson & Khalsa, 2021; Quigley et al., 2021).

Even though recently, interoception has been positioned as central to emotion, perception, social cognition, learning and bodily self-awareness, among others, these higher-order processes are thought to be modulated by low-level functions of interoception and specifically those related to energy and physiological regulation (Quigley et al., 2021). Interoception has key role in two complementary regulatory processes namely, homeostasis and allostasis. Homeostasis refers to the maintenance of stable internal conditions through reactive adjustments to restore equilibrium. On the other hand, allostasis relates to anticipatory process where the brain predicts future demands and modulates physiological states proactively (Sterling, 2012). This involves integrating interoceptive signals with prior experiences and environmental cues to prepare the body for potential challenges (Corcoran & Hohwy, 2017; Quigley et al., 2021). These processes underscore the importance of interoception not only in immediate physiological regulation but also in optimising resource distribution and energy efficiency, forming the basis of adaptive behaviour and long-term health (Barrett et al., 2016).

Interoceptive processing relies on a sophisticated neural architecture encompassing both ascending (afferent) and descending (efferent) pathways (Barrett & Simmons, 2015;

Berntson & Khalsa, 2021). The primary ascending pathways begin with sensory receptors in visceral organs (e.g., mechanoreceptors, chemoreceptors) that transmit information via the vagus nerve, glossopharyngeal nerve, and spinal nerves. These signals converge in the nucleus of the solitary tract (NTS) and the Parabrachial nucleus (PBN). From there the signals project to subcortical (e.g. hypothalamus, amygdala) areas and to noradrenergic and serotonergic nuclei in the brainstem. Next, through thalamus, interoceptive signals reach cortical areas, including the insular, anterior cingulate (ACC) and ventromedial prefrontal cortex (vmPFC). For the regulation of autonomic and endocrine responses, descending pathways involve projections from i) the hypothalamus that controls autonomic functions and endocrine responses via its connections with the brainstem and spinal cord, ii) the periaqueductal gray (PAG) and brainstem Nuclei that facilitate coordinated visceral-motor responses such as heart rate and breathing adjustments, and iii) the anterior insular cortex (AIC), OFC and AAC, i.e. the visceromotor areas, which provide top-down predictions from the top of the interoceptive hierarchy (Barrett & Simmons, 2015; Seth & Friston, 2016).

1.2.1. Interoceptive predictive processing

Recent formulations of PP have not only incorporated interoceptive processing but have position it as central in the emergence of the embodied self (Limanowski & Blankenburg, 2013; Seth, 2013; Seth & Tsakiris, 2018). Similarly to exteroceptive PP, the brain generates predictions about the body’s internal states, informed by prior experiences, current context, and anticipated demands. These predictions are compared against ascending interoceptive signals, with mismatches generating interoceptive prediction errors. These errors can then be resolved by updating the brain's generative models or by initiating regulatory responses, such as adjustments in autonomic, endocrine, or behavioural system, to align bodily states with predictions. This bidirectional interplay casts interoception as an active inference process, with descending predictions shaping the processing of ascending signals. The neural architecture of interoception, involving regions such as the AIC and ACC, supports this predictive model by integrating multimodal inputs and coordinating top-down and bottom-up communication (Barrett & Simmons, 2015). In this context, it has been suggested that interoceptive predictions allow the integration of somatic states with external sensory information via suppression of the sensory consequences of visceral activity for optimal behavioural adaptation (Seth & Friston, 2016).

1.2.2. Measurements of interoception

Before delving into the specific ways this reciprocal brain-body exchange shapes affect, cognition and behaviour, we will first overview the different interoceptive processing assessments. Empirical research in interception, has mainly focused on cardiac interoception, however, other visceroception processes have also gained attention more recently (Allen et al., 2023; Harrison et al., 2021; Holzer, 2017). Due to the relatively few years of systematic research, there is still no consensus on the terminology and methodology used to measure interoceptive processes. The following section outlines the main assessments of interoception spanning unconscious and conscious levels.

Interoceptive sensibility, according to Garfinkel and colleagues (2016) is a subjective measure of one's perceived sensitivity to internal bodily sensations, evaluated through self-report questionnaires such as the Body Perception Questionnaire (Porges, 1993) or the Multidimensional Assessment of Interoceptive Awareness (Mehling, 2018). Interoceptive accuracy (IAcc), conversely, provides an objective measure of one’s ability to detect internal bodily signals, and is usually assessed by two tasks: the heartbeat detection task (Schandry, 1981) and the heartbeat discrimination task (Whitehead et al., 1977). In the first task, participants silently count their perceived heartbeats within a given timeframe, which is then compared to the recorded heartbeats for accuracy. In the discrimination task, participants judge the synchrony between their heartbeats and external stimuli such as auditory tones. Notably, the partial dissociation between interoceptive accuracy, interoceptive sensibility, and interoceptive awareness (the correspondence between confidence and accuracy in interoceptive tasks) has been clearly emphasized as these dimensions reflect distinct underlying processes and abilities (Garfinkel et a., 2015).

At unconscious levels, cardiac interoceptive processing can be tracked by neural markers, such as the heart-Evoked-Potential (HEP). HEPs are cortical responses to cardiac activity, measured by time-locking EEG or MEG signals to individual heartbeats (Park & Blanke, 2014). From a PP perspective, HEPs reflect cortical processes that contrast cardiac inputs (e.g., baroreceptor signals) against descending interoceptive predictions (Ainley et al., 2016; Owens et al., 2018). When expected and actual cardiac signals misalign, ascending precision-weighted PE increase HEP amplitudes, likely reflecting activity in high-level interoceptive regions (AIC and ACC). HEP amplitudes can also be modulated by overt or covert attention towards interoceptive sensations, enhancing the gain of the related ascending signals (Ainley et al., 2016; Petzschner et al., 2019). Therefore, HEPs have been widely used in tasks measuring interoceptive attention, with recent meta-analysis demonstrating consistent associations with interoceptive accuracy, attention, and arousal (Coll et all, 2021).

1.2.3. Interoception and emotion

Major theories of emotion have long highlighted the role of internal bodily signals in shaping emotional experiences (Critchley & Garfinkel, 2017; Gendron & Barrett; 2009). PP accounts, aligning with evaluative and two-factor theories of emotion (Gendron & Barrett, 2018;

Schachter & Singer, 1962), frame emotions as arising from the brain's predictions about interoceptive states (Seth & Friston, 2016). In this view, emotions are the result of hierarchical generative models that integrate interoceptive signals with contextual information, minimising prediction errors through active inference. These contextualised interpretations and categorizations of physiological states are assigned a social-language sign, that is, a concept like anger or fear (Gendron & Barrett, 2018). There is now a considerable body of evidence demonstrating the close link between interoceptive and emotion experiences, as interoception has been suggested to underpin, among others, the perception of emotional stimuli (Garfinkel and Critchley, 2018), emotion regulation (Füstös et al., 2013) and the ability to describe and perceive emotional states (Brewer et al., 2016). This has vital implications for social cognition, as interoceptive predictions interact with exteroceptive and proprioceptive predictions to produce multimodal interpersonal predictions (Ondobaka et al., 2017). These predictions form the foundation of emotional responses to exteroceptive cues, including socially salient signals (Barrett & Simmons, 2015; Seth & Friston, 2016). In sum, under interoceptive PP, the emotional content of experience emerges through active inference, as the brain infers the likely internal and external causes of physiological changes (Seth, 2013), including social causes at various level of abstraction.

1.2.4. Interoception in memory, learning and decision-making

Learning and memory are also cognitive functions linked to interceptive processing. Associative learning paradigms, especially in classical and operant conditioning, show that internal bodily signals act as salient cues or reinforcers (Garcia et al., 1970; Razran, 1961; Werner & Schandry, 2024). In classical conditioning, interoceptive signals—such as changes in heart rate, gut sensations, or respiration—can function as conditioned (CS) or unconditioned stimuli (US), forming associations with external events. Compared to exteroceptive conditioning, interoceptive conditioning tends to be more implicit and persistent (Werner & Schandry, 2024) with a growing body of literature demonstrating how interoceptive signals relate to learning. For example, Katkin and colleagues (2001) showed that individuals with high IAcc, but not those with low IAcc, were able to form implicit associations between masked visual stimuli and aversive electrical stimulations. Interestingly, while in most studies the CS is external, Zaman and colleagues (2016), using a respiratory-based conditioning paradigm, showed that interoceptive signals can also serve as a CS, highlighting the bidirectional relationship of associative learning between external and internal states.

Interoceptive signals have also been implicated in memory processes, and specifically in the encoding and retrieval of emotionally salient information. For example, Werner and colleagues (2010) found that individuals with higher IAcc exhibited better implicit memory for emotionally charged words, likely due to more accurate physiological monitoring which enhanced emotional salience. Memory effects of cardiac timing further support this role. Garfinkel and colleagues (2013) found that words presented during systole (when baroreceptors are active) were less likely to be recalled compared to those presented during diastole. This impairment, however, was reduced in individuals with high IAcc, indicating that interoceptive precision influences attentional gating mechanisms. Similarly, in RL, feedback delivered during systole was found to enhance learning of fearful face-name pairs, especially in individuals with high IAcc, supporting the idea that interoceptive states influence predictive weighting mechanisms during memory formation (Pfeifer et al., 2017).

Further evidence for interoceptive involvement in learning and memory comes from studies stimulating the vagus nerve, one of the major pathways conveying afferent bodily signals to the brain. In fear conditioning paradigms, vagus nerve stimulation (VNS) enhances both acquisition and consolidation of conditioned responses (Alvarez-Dieppa et al, 2016; Childs et al., 2015), likely through its effects on noradrenergic and cholinergic systems, which regulate synaptic plasticity. It has also been shown to facilitate fear extinction, possibly by enhancing top-down regulation of amygdala activity via the vmPFC (Pena et al., 2014). In operant learning, VNS has been found to enhance reward sensitivity and motivation, potentially by modulating dopaminergic pathways (Neuser et al., 2020; Weber et al., 2021). However, findings have been inconsistent, with some studies reporting improvements in learning performance while others find no effect or even impairments (Kühnel et al., 2020; Carroll et al., 2024). This variability suggests that VNS effects may depend on task demands, individual differences, or specific learning mechanisms involved. While VNS has generally been proved to enhance learning and memory, with promising applications in clinical populations with atypical interoceptive integration and social behaviour (Jin & Kong, 2017; Yan et al., 2020), potential VNS effects in social learning have yet to be investigated.

Another domain where interoception influences cognition is in intuitive and motivational decision-making. These decision processes are postulated to involve the integration of various sources of information, including affective and bodily inputs. This integration informs complex value calculations for each outcome, thereby helping to reach an optimal decision. The importance of interoception in decision-making was emphasized by the influential Somatic Marker hypothesis (Damasio, 1996), according to which, emotional and physiological responses associated with past experiences (i.e. "somatic markers") function as implicit markers guiding decision-making under conditions of uncertainty and complexity. Based on this, several studies have demonstrated that Individuals with impaired interoceptive integration, like those with vmPFC lesions, exhibit deficits in emotionally informed decisions, lacking the formation of associated physiological reactions (Bechara et al., 1994; 1997).

Subsequent studies demonstrated how interoceptive abilities modulates the reliance on "somatic markers" during decision-making. For example, Werner and colleagues (2009) revealed that IAcc predicted the selection of advantageous option in a reward-learning task. However, Dunn and colleagues (2010) showed that interoception is not always helpful. In their study, they found that higher IAcc improved decision-making when bodily responses aligned with optimal choices, but also led to poorer choices when "somatic markers" favoured disadvantageous outcomes, suggesting that overreliance on interoceptive cues can sometimes be maladaptive. Another notable finding of this study is that individuals with higher interoceptive accuracy exhibited higher alignment between heart rate changes and subjective arousal in response to emotional stimuli.

Interoception is also implicated in social decision-making, especially in contexts of emotionally charged exchanges. Studies using the Ultimatum Game (Güth et al, 1982), where players decide whether to accept or reject unfair monetary offers, show that physiological arousal modulates rejection rates (Pillutla & Murnighan, 1996; Sanfey et al., 2003). Additionally, neuroimaging studies have linked right anterior insula activation to rejection behaviour, suggesting that heightened interoceptive awareness amplifies sensitivity to social fairness violations (Sanfey et al., 2003). Moreover, Dunn et al. (2012) found that individuals with higher physiological arousal were more likely to reject unfair offers, with IAcc moderating this effect. Interestingly, Lenggenhager and colleagues (2013) demonstrated that providing participants with real-time heartbeat feedback increased self-focused decision-making, indicating that heightened interoceptive attention can bias individuals toward self-interest in social exchanges.

To summarize, learning, memory, and decision-making are intertwined cognitive functions associated with affective and predictive processes. Interoception shapes associative learning, emotional memory, and adaptive decisions by regulating attention, salience, and precision weighting (Critchley & Carfinkel, 2018; Gu & FitzGerald, 2014). Understanding these interactions can provide insight into how bodily signals shape (social) cognition and how disruptions in interoception may contribute to maladaptive learning and decision patterns.

1.2.5. Interoception and social cognition

A developmental perspective

The first and foremost link between social cognition and interoception traces back to the first steps of ontogeny (Atzil et al., 2018; Fotopoulou & Tsakiris, 2017). Upon entering life, an infant, albeit equipped with some basic interoceptive functions through evolution, has its body regulation and survival in the hands of its primary caregivers. This vital social ontogenetic relationship, described as “allostatic-dependency”, has been supported to intertwine interoception with socio-affective processes (Atzil et al., 2018). A newborn’s unelaborate interoceptive model is initially cascaded by PE signals which are addressed either via learning (model updating) or actions like crying (active inference). By influencing their external social environment, through the embodied interactions and communication elicited, infants associate their bodily sensations with contextual social conditions and meanings. This contextualization of internal bodily feelings is performed by learning interoceptive-exteroceptive contingencies, via accurate and inaccurate predictions, thus serving as the basis for the emergence of the self as well as the self-other distinction (Fotopoulou & Tsakiris, 2017). This also suggests that moderate levels of ambiguity may be beneficial for adaptive self-other processing (de Bézenac et al., 2018). Lastly, the socially meditated formation of interoception and social brains outlines an internalization – externalization dialectic, where internal bodily states influence and are influenced by external social dynamics, in a continuous feedback loop that supports interpersonal regulation and prediction error minimization (Bolis et al., 2020; 2023).

The developmental origins of interoception and social cognition outlined highlight key processes that continue to contribute to the interconnection of these functions later in life, such as, multisensory integration, domain-general learning, and self-other distinction and attunement.

Interoceptive processes influencing social cognition

Several studies have investigated how bodily signals and affective states influence the perception of socio-affective stimuli using various methods. For example, studies manipulating stimulus timing relative to heartbeat have shown that afferent cardiac input affects the detection (Garfinkel et al., 2014), intensity (Gray et al., 2012) and trustworthiness of face stimuli (Azevedo et al., 2023), while another study found that the presentation of images of Black and White individuals at systole enhances racial stereotypes (Azevedo et al., 2017). These findings suggest a causal role for interoceptive signals in shaping first impressions and emotional judgements.

Similarly, it has been found that autonomic activity can impact on emotion recognition and empathy (Nitschke et al., 2023). For instance, acute psychosocial stress can enhance recognition of facial emotional expressions (Barel & Cohen, 2018; Beaurenaut et al, 2023).

Stress has also been found to affect self-other distinction (Tomova et al., 2014) and enhance emotional empathy and prosocial behaviour (Smeets et al., 2009), even if it can also impair cognitive appraisal, leading to inappropriate responses in some contexts (Nitschke et al., 2023). Other studies have linked higher HRV to better emotion processing in healthy and socially anxious individuals (Gaebler et al., 2013; Quintana et al., 2012). As HRV is regulated by the vagus nerve, higher HRV reflects better autonomic control, thereby influencing adaptive social behaviour (Porges, 2009). In sum, these findings suggest a complex relationship where physiological stress and arousal can both enhance and impair emotion perception and empathy, depending on the task, context and individual differences.

Bodily influences on social perception can also be studied more implicitly in the context of the mood-congruency effects, affect misattribution or emotion egocentricity biases. Specifically, various studies have manipulated participants' affective states using implicit emotion induction techniques to examine their impact on the perception of others' emotions (Anderson et al, 2012; Silani et al., 2013; Trilla et al, 2021). Some of these studies presented affective images using continuous flash suppression, rendering them invisible yet unconsciously perceived (Anderson et al, 2012; Siegel et al., 2018). It was found that unseen positive (smiling) or negative (scowling) affective stimuli biased the categorization of neutral faces, making them appear more or less pleasant, likable, trustworthy, attractive or socially warm. A more recent study employed the affective misattribution procedure, where a masked affective prime was followed by a neutral image, while the role of immune system activation (via vaccination) and interoceptive sensibility were also examined (Feldman et al., 2023). Results showed that higher inflammatory reactivity heightened sensitivity to affective primes, enhancing trustworthiness ratings for positive and neutral primes while reducing them for negative primes, with participants with greater interoceptive difficulty being more influenced by affective primes. These findings have been interpreted as examples of ‘affective realism’, which posits that perception is not purely an objective readout of external reality and that our affective states directly colour how we perceive the world (Anderson et al., 2012; Barrett & Bar, 2009).

Other studies have used different methods of emotion induction and perception to assess similar effects described as emotional egocentricity biases (EEB), i.e. the tendency to others’ affective states congruent with our own. For instance, one study developed an audio-visual paradigm where participants judged their own and others’ emotional experiences while listening to positive or negative sounds paired with visual stimuli during interpersonal congruent or incongruent conditions (Von Mohr et al., 2020). Significant EEB was observed as participants' judgments of others' emotions were influenced by their own emotional states during incongruent trials. A follow up study showed that EEB is modulated by individuals’ interoceptive abilities, particularly in conditions of heightened representation of autonomic activity (Von Mohr et al., 2021). Together, these findings highlight the interplay between physiological states, affective context and interoceptive traits in shaping social perception and self-other blurring or distinction.

However, an important limitation of some of these experimental designs is that they do not facilitate to establish a clear dissociation between self-projection perceptual processes and self-other distinction abilities. To isolate affective or interoceptive influences on perception related effects, it is critical to avoid the involvement of explicit self-referencing cues. Studies using implicit emotion induction techniques and psychophysical methods suggest that affective effects on social perception can be linked to actual perceptual processes rather than to post-perceptual biases (Trilla et al., 2021). Ambiguous stimuli may be particularly susceptible to these influences as their uncertainty leaves space for stronger reliance on top-down predictions to resolve uncertainty and shape the perceptual content (Hohwy, 2017; Otten et al., 2017).

Interoceptive traits and social cognition

As discussed, bodily and affective states do not directly affect emotion perception and social cognition; rather, individual differences in how these states are perceived and attended to play a key modulatory role. Evidence shows that IAcc and IS are associated with greater empathic concern and perspective-taking (Baiano et al., 2018; Mul et al., 2018). Higher IAcc has also been linked to increased empathy for pain, enhanced ability to infer others' intentions, and greater sensitivity in recognizing others' emotions (Terasawa et al., 2014). Interestingly, cardiac IAcc has been found to predict the perception of other interoceptive states (heart rate) as well (Arslanova et al., 2022 Moreover, IS correlates with greater sensitivity to ambiguous emotional expressions, suggesting a role in resolving uncertainty in social interactions (Hübner et al., 2021). These findings further highlight how interoceptive traits influence social perception and interaction. Although several facets of interoception have been examined in this context, other aspects, associated with the adaptive modulation of interoceptive processes, such as interoceptive attention, are yet unexplored.

Interoceptive processing in dynamic socio-affective environments

I have been discussing how internal bodily states affect (social) cognition, but I still have not addressed in detail the internalization processes of how the body reacts in dynamic socio-affective environments and the elicited brain-body interactions.

Evidence shows that interoceptive and bodily self-awareness are increased in social contexts (Baltazar et al., 2014; Hazem et al., 2018; Maister et al., 2017), illustrating the tight bidirectional link between social interactions and bodily functions. However, whether this increased interoceptive processing is beneficial depends crucially on the flexibility of attention allocation between one's own body and the external world. One way to explore how interoceptive processing dynamically interacts with sensorial and contextual factors is by measuring modulations in HEP amplitude. For instance, HEP amplitudes are known to be increased during the perception of emotion-inducing stimuli (Couto et al., 2015) while another study observed HEP modulation during empathic perception as a function of individual differences in empathic traits (Fukushima et al., 2011).

Interestingly, these neural markers of interoceptive processing have also been shown to be sensitive to changes in interoceptive predictions in dynamic socio-affective environments. Marshall and colleagues (2018) employed a repetition-suppression design to probe how repeated versus alternating presentations of facial expressions influenced interoceptive processing. Negative stimuli (angry and pained faces) reduced HEP and visual evoked potential (VEP) amplitudes, while sad faces elevated HEP amplitude compared to neutral and positive stimuli. Additionally, a significant correlation found between HEP and VEP suppression for angry faces suggests an attentional trade-off between interoceptive and exteroceptive domains. In a subsequent study (Gentsch et al., 2018), they controlled for low-level visual effects and added a probability manipulation to assess contextual expectations' effects on HEP amplitudes. In the high-probability repetition context, stimulus repetitions were more likely, whereas in the low-probability context, alternations were more likely. Expected angry faces reduced HEP amplitude compared to unexpected neutral faces, but this effect was absent for neutral expressions. Importantly, this modulation occurred only in the high-probability repetition condition, which suggests that the formation of top-down high-level predictions within a socio-affective context influences interoceptive processing, beyond low-level predictive signals from prior trials. Thus, these findings demonstrate that top-down affective probabilistic expectations and their interaction with emotional valence modulate interoceptive processing. However, this methodology has not yet been employed to examine interoceptive predictive processing during social interactions that involve probabilistic inference and learning.

1.2.6. Interoception and psychopathology

Given the vital role of interoception in homeostatic regulation and socio-affective functions, numerous studies and models support that various symptoms and underlying mechanisms of psychological and neurodevelopmental disorders are linked to atypical interoception (Quadt et al., 2018; Owens et al., 2018; Trevisan et al., 2021). Common features of aberrant interoception include negative interpretation biases towards bodily sensations, dysregulated attention (e.g., hypervigilance), distorted physiological sensitivity to bodily changes, reduced perceptual accuracy, and poor metacognitive insight (Paulus et al., 2019). Despite these shared features, the specific symptom and deficit clusters vary considerably within and across disorders.

For instance, several conditions, such as depression, alexithymia, depersonalization/derealization, and autism, often involve reduced IAcc, limited interoceptive integration, and maladaptive IS (Paulus et al., 2019; Quadt et al., 2018). In contrast, anxiety disorders exhibit a different profile of maladaptive interoception, often marked by higher IAcc, coupled with hypervigilance to bodily sensations (Domschke et al., 2010; Pollatos et al., 2009). According to PP accounts, such interoceptive atypicalities are often indicative of inflexible top-down interoceptive predictions, that is, disruptions in how the brain expects, interprets, and updates information about its own internal bodily state (Paulus et al., 2019). This has profound implications for embodied, affective subjectivity; that is, how people infer, "this is me, feeling like this, in this context." Next, we will focus on the main atypicalities of interoceptive inference in anxiety and autism.

According to PP models, anxiety disorders are characterised by rigid, negative expectations (hyperprecise priors) and difficulty updating these expectations to new contexts (Paulus & Stein, 2010). Consequently, even minor bodily fluctuations, like in heart rate or respiration, are magnified, triggering excessive reactivity (Domschke et al., 2010; Lapidus et al., 2020). This heightened focus on bodily signals perpetuates a self-reinforcing cycle: where the mismatch between expected and actual bodily signals remains unresolved, fuelling persistent worry and somatic discomfort, and in turn, hindering adaptation to changing environments and exacerbating anxiety-related symptoms (e.g., avoidance behaviours).

Autism is similarly postulated to arise from atypical interoceptive inference, where predictions about bodily states fail to align with incoming sensory signals (Proff et al., 2022). Unlike anxiety, these inflexible or imprecise predictions can lead to either underweighting or overemphasizing bodily cues, creating challenges in attuning to internal states. This contributes to difficulties in emotional awareness, such as identifying and describing bodily sensations and emotions (Kinnaird et al., 2019; Palser et al., 2021). This is why several studies have examined alexithymia as a mediating factor between interoceptive deficits and autistic symptomatology (Butera et al., 2023; Mul et al., 2018).

Moreover, faulty interoceptive processing in autism may exacerbate sensory sensitivities and emotion regulation difficulties and, in turn, impair the ability to form appropriate responses to social and emotional stimuli (Butera et al., 2023; Quattrocki & Friston, 2014). A proposed core mechanism underlying the autistic phenotype has been suggested to pertain to aberrant viscero-sensory integration, leading to inflexible contextualization of interoceptive signals and inability to minimize interoceptive PE by updating top-down predictions in new contexts (Noel et al., 2018; Proff et al., 2022). While the underlying mechanisms driving aberrant interoceptive processing in autism and anxiety might differ, they share overlapping symptomatology, including anxious feelings, heightened sensory sensitivity, emotional dysregulation, rigid behaviours, and social difficulties (Hwang et al., 2020; Palser et al., 2018).

These shared symptoms might be related to an inability to flexibly reduce interoceptive PE, leading to an exaggerated need for predictability and avoidance of novelty. Thus, interoceptive deficits and associated difficulties may be more apparent in environments of increased uncertainty, such as dynamic social interactions. Indeed, evidence suggests that individuals with anxiety and autism often struggle with uncertainty in social and non-social situations (Jenkinson et al., 2020; Pulcu et al., 2019; Sevgi et al., 2020). In social interactions, which are typically highly uncertain, accurate perception and optimal precision-weighting of interoceptive signals is considered key for adaptive behaviour (Ondobaka et al., 2017; Quattrocki & Friston, 2014). Therefore, the application of PP modelling along with interoception assessments can provide a promising formal framework to reveal underlying mechanisms and quantify how interoceptive priors and precision-weighted PE shape adaptive and maladaptive socio-affective processes under social uncertainty.

1.3. Bayesian modelling in psychology and psychiatry

1.3.1. General Framework

The field of computational psychiatry has emerged as a promising approach for providing mechanistic and functional insights into mental illness, moving beyond symptom-based classifications (Friston et al., 2014; Moutoussis et al., 2018; Stephan & Mathys, 2014). Bayesian modelling under the PP framework has propelled these advances in computational neuropsychology. As discussed, PP posits that the brain actively constructs and refines a generative model of the world to predict sensations, with perception, learning and action serving to minimize discrepancies between predictions and sensory data. Applications of this perspective attempt to reframe psychiatric and neurological disorders as manifestations of aberrant inference, wherein the brain's predictive and belief-updating mechanisms become dysregulated. To probe this, computational psychiatry aims to develop generative models that link observed behaviours to their hidden causes, namely, the psychophysiological processes and external factors driving them.

Specifically, generative models describe how observed data (y) is generated from latent variables (x) and parameters (θ) through a likelihood function p(y|x,θ) and priors p(x,θ), forming a joint probability distribution:

$$\mathrm{p}(\mathrm{y},\mathrm{x},\theta) = \mathrm{p}(\mathrm{y} \mid \mathrm{x},\theta)\mathrm{p}(\mathrm{x},\theta)$$

The likelihood function describes how the latent variables and parameters generate the observed data. For instance, in a perceptual decision-making task, the likelihood might represent how internal beliefs translate into observable choices and reaction times. The prior distribution represents beliefs about the latent variables before observing any data.

These priors, for example, can encode participants' expectations of environmental volatility or reward probabilities in a learning task (Mathys et al., 2014).

Using the Bayes' theorem, generative models invert this relationship, estimating the posterior probability of latent variables and parameters given the observed data:

$$p(x, \theta \mid y) = \frac{p(y)p(y \mid x, \theta)}{p(y)}$$

Here, p(y) represents the marginal likelihood, which is the evidence for the model and is computed by integrating over all possible values of x and θ. This evidence provides a metric for comparing different models.

Model inversion is thus the reverse process of generating data via a forward generative model (Stephan & Mathys, 2014). However, since directly computing the posterior is often intractable, approximation methods are used, such as the Variational Inference, that approximate the posterior with a simpler distribution that minimises the divergence from the true posterior or using Markov Chain Monte Carlo methods, which use sampling techniques to estimate the posterior distribution.

Model inversion allows researchers to interpret observed data in terms of the processes or mechanisms hypothesised by the models. The hierarchical structure of these models mirroring the postulated brain's architecture (Friston & Kiebel, 2009; Friston et al., 2014), adds to the biological plausibility of this computational framework, thus allowing for formalised hypotheses about the biological and cognitive mechanisms underlying human behaviour. In psychiatric contexts, generative models capture how cognitive processes go awry, and particularly in learning and decision-making, as considered central manifestations of maladaptive cognition (Moutoussis et al., 2011; Pulcu & Browning, 2019). Due to their hierarchical Bayesian architecture, PP models can capture learning and decision-making process beyond traditional modelling approaches (Friston et al., 2016; 2017). For example, in RL, model-based learning can incorporate explicit and implicit knowledge about the environment that drive structured, goal-directed probabilistic learning beyond simple PE computations. In this line, interoceptive inference has also provided a promising framework for computational psychiatry by modelling subjective states as Bayesian beliefs about the body's internal signals (Gu et al., 2019). However, despite recent advances, such as in computational psychosomatics (Petzschner et al., 2017), applications of Bayesian interoceptive inference to psychophysiological processes remain limited.

This Bayesian approach to both clinical and non-clinical applications has been cast as meta-Bayesian (Daunizeau et al., 2010), as it involves Bayesian inference of the experimenter on the Bayesian inference processes of the subject’s brain or mind. In Bayesian model comparison, experimenters evaluate different models (e.g., classical RL vs Bayesian learning) to determine which best explains the observed data (Stephan et al., 2014). At a meta-level, this framework aggregates evidence across individuals to assess model plausibility at the group level, using metrics like the exceedance probability, which quantifies the probability that one model outperforms others within a population. This approach enables uncertainty quantifications at multiple levels, making Bayesian inference ideally suited for handling uncertainty estimations related to probabilistic beliefs, parameters, models, and group-level heterogeneity (Daunizeau et al., 2010; Stephan & Mathys, 2014). Next, we will discuss an application of this approach: the Hierarchical Gaussian Filter (HGF).

1.3.2. Hierarchical Gaussian Filter

The Hierarchical Gaussian Filter (HGF), introduced by Mathys et al. (2011; 2014), integrates Bayesian inference principles within a hierarchical generative model of learning under uncertainty. It builds upon previous Bayesian approaches (Behrens et al., 2007; Daunizeau et al., 2010) while addressing limitations of reinforcement learning (RL) and classical Bayesian models. RL, albeit computationally simple, is heuristic, lacking probabilistic foundations and struggling in volatile contexts where flexible learning is required. On the other hand, ideal Bayesian models, though theoretically optimal, face challenges such as computational intractability, questionable biological implementation, and inability to account for inter-individual variability in learning (Mathys et al., 2011). The HGF provides a middle ground, explicitly modelling uncertainty across multiple hierarchical levels to allow for adaptive learning in dynamic environments.

In its main formulation, the HGF captures learning across three hierarchical representation levels, each encoding a different type of uncertainty. The first level represents perceptual or irreducible uncertainty, describing the inherent ambiguity in sensory outcomes; e.g., whether a facial expression perceived is friendly or hostile. When perceptual ambiguity is high, our interpretation of others' expressions is more heavily influenced by contextual information. The second level reflects estimation uncertainty, tracking beliefs about the hidden causes of observations, such as whether the other person tends to be friendly or hostile based on previous interactions. Over time, this form of uncertainty typically decreases as more evidence is accumulated, though it may fluctuate with new observations and especially in volatile environments. Finally, the third level encodes volatility, or the uncertainty around the stability of these hidden states over time—for instance, whether a person's behaviour is consistent or liable to change unpredictably. In contrast to simple RL models, this hierarchical structure allows belief updating to be dynamically modulated, with higher-level volatility driving adjustments in lower-level learning rates.

In short, such generative models formally describe how an agent's environment generates a sequence of inputs and how the agent optimally combines these inputs (u(1), u(2),…,u(k-1)) with prior information to predict the next input, u(k). The model can handle both continuous and discrete sensory data (Mathys et al., 2011). For instance, in a task based on our previous example, the input is binary as the other's expression can have two states, either hostile (u=0) or friendly (u=1). The first representation level, which captures this binary environmental state regarding the other's expression, also takes binary states: x1(k) ∈ {0,1}. This mapping from sensory input to the first level can be either deterministic or stochastic, depending on a fixed or variable parameter that reflects the amount of sensory noise. The higher representational levels are continuous and evolve as random walks with step size dependent on the state of the level above, enabling context-sensitivity of belief updating.

Specifically, belief updating occurs via precision-weighted prediction errors, with each level generating prediction errors that propagate to the level above, dynamically refining beliefs about the environment. Higher volatility estimates increase the influence of new observations, while greater stability assumptions reduce learning rates, making the system adaptive to both stable and volatile conditions. Importantly, the HGF includes model parameters that capture individual differences, influencing learning rates, belief trajectories, and the coupling strength between hierarchical levels independently of sensory input. Of note, the HGF, also referred to as the perceptual model, is combined with a response model, i.e., a function that describes the decision process of how beliefs are mapped to observed responses.

Lastly, Bayesian model inversion estimates posterior distributions by optimizing log-model evidence, which corresponds to negative surprise (the likelihood of observed data given the model). Because exact computation of log-evidence is intractable, variational inference approximates it by minimizing free energy, enabling biologically plausible, trial-by-trial belief updating through closed-form, single-step updates. This combination of computational efficiency, biological plausibility, and adaptability makes the HGF a powerful tool for studying both typical and pathological learning processes.

1.3.3. Modelling learning under uncertainty

Therefore, the HGF has been used in a wide range of studies to model behavioural and neurophysiological observations across social and non-social tasks. Notably, two studies have provided empirical support for theoretical accounts of how neuromodulators shape learning under uncertainty (Lawson et al., 2021; Marshall et al., 2016). Marshall and colleagues (2016) used pharmacological manipulations to show that noradrenaline (NA) enhances learning about environmental volatility, acetylcholine regulates uncertainty attribution within stable contexts, and dopamine modulates motor responses based on beliefs about volatility estimates. Similarly, Lawson and colleagues (2021) demonstrated that blocking NA with propranolol slows learning, particularly in volatile environments, by reducing sensitivity to prediction errors and volatility updates. Both studies suggest a specific role for these neuromodulators in adjusting precision-weighting of hierarchical belief updating, tuning behavioural responses to uncertainty.

Furthermore, the study by Becker and colleagues (2016) examined the relationship between uncertainty and acute stress responses. Their results showed that subjective and physiological stress responses were associated with irreducible uncertainty. Notably, the subjective and physiological sensitivity to uncertainty predicted task performance, suggesting an adaptive role for stress and interoceptive signals in navigating uncertain threatening environments. Other studies using similar methods have further investigated the impact of trait and state anxiety on adaptive learning. For example, Browning and colleagues (2015) found that high trait-anxious individuals failed to adjust their learning rate between stable and volatile conditions, treating both as similarly uncertain. In addition, the observed reduced pupil response to volatility suggested impaired noradrenergic signalling, further indicating a relation between neuromodulatory dysfunction and maladaptive learning under uncertainty.

The HGF has also been used to uncover differences in uncertainty processes due to neurodevelopmental factors. Lawson and colleagues' (2017) study showed that individuals with autism spectrum disorder (ASD) tend to overestimate environmental volatility in a non-social associative learning task. Another study examined how autistic traits in a neurotypical sample could affect probabilistic learning and decision-making, where participants had to infer reward contingencies from social (gaze direction) and non-social (card probabilities) cues. The findings revealed that higher AQ scores were associated with reduced weighting of social cues in belief updating, particularly in volatile conditions, leading to poor task performance. Importantly, this was not mainly attributed to an inability to process social information, but rather to weaker integration of social cues into decision-making. This study suggests that, apart from clinical autism, related traits could also affect volatility processing. Additionally, it further underscores the importance of using response models along with the HGF to examine how different sources of information are integrated during decision-making. These results support predictive coding theories of autism, suggesting that autistic traits are linked to atypical, context-sensitive precision-weighting and information integration in social and non-social learning contexts (Lawson et al., 2014; Palmer et al., 2017).

Beyond clinical application, these Bayesian learning approaches have also been used to investigate social learning under volatility in typical populations, offering significant insights into the neurocomputational mechanisms underlying adaptive social behaviour (Sevgi et al., 2020; Diaconescu et al., 2014). However, while theoretical and computational models have incorporated interoception into learning processes (Lehman et al., 2024; Smith et al., 2021), no experimental study has thus far investigated how interoceptive inference shapes dynamic social interactions, such as adaptive empathy under uncertainty.

1.4. The motivation for this thesis

The empirical evidence reviewed above highlights multiple connections between social cognition and interoceptive processing. Research evidence and theoretical models suggest that effectively navigating social interactions requires the continuous integration of external sensory cues and internal bodily states to infer others' emotions, intentions, and actions (Adams & Kveraga, 2015; Gallagher & Allen, 2018; Lehman et al., 2024). While these models emphasize the importance of uncertainty estimations in integrating multiple sources of information for adaptive social inference responses, few studies have examined the impact of uncertainty surrounding bodily functions and interpersonal dynamics. PP and IT accounts emphasize that the brain actively constructs social meaning based on past experiences and contextual information, rather than passively processing incoming data. (Barret et al., 2011; de Bruin & Michael, 2021; de Jaegher et al., 2010). Importantly, contextual information arises not only from external sources but also from within the body, influencing how individuals resonate affectively, perceive their own and others' emotions, and respond to others' needs. (Bolis et al., 2023; Ondobaka et al., 2017). On the other way around, social predictions and the related environmental uncertainty affect one's interoceptive predictions and allostatic regulation. Additionally, individual differences in interoception, emotion processing, and perceptions of uncertainty can further shape how individuals experience their bodies in dynamic social environments and their ability to respond adaptively. To shed light on these complex dynamics between bodily and social inference, a multimethod computational approach is required. Therefore, this thesis aims to enhance our understanding of how interoceptive processing influences social perception, learning, and decision-making under varying conditions of uncertainty by incorporating behavioural, psychophysiological, and computational methods. Specifically, through a series of experiments, this thesis attempts to understand how interoceptive and multimodal social predictions support and influence emotion perception and empathic learning. Do these factors have different impacts under conditions of perceptual ambiguity and environmental volatility? How are these socio-affective processes modulated by individual differences in interoception, autistic traits, alexithymia, and IU?

In the first empirical chapter (Chapter 2), we focused on the interplay between interoception, IU, anxiety and social cognition. Even though previous research has linked IS to alexithymia, anxiety, social inference, the underlying mechanisms remain clear (Butera et al., 2023; Trevisan et al., 2021; Nitschke & Bartz, 2023). While IU is associated with both interoception and anxiety (Carleton, 2012; Paulus & Stein, 2010), it has yet to be examined as a mediator in this relationship. Given IU's connections to social anxiety and aversion to uncertainty, we hypothesize that it may also impact affective empathy and emotion recognition in ambiguous social contexts. Moreover, although computational approaches have been employed to examine anxiety-related effects in emotion perception (Dillon et al., 2022; White et al., 2016), no study has investigated the perceptual, attentional, or decisional processes related to IU in interpreting ambiguous expressions using models like the Diffusion Drift Model (DDM). To address these gaps, we conducted two studies aiming at elucidating how IU and alexithymia interact with interoception to influence anxiety and social skills. We combined a path analysis approach with DDM analysis on an emotion recognition task to explore these relationships and the related mechanisms.

In Chapter 3, we examined how emotion perception, and particularly EEB, is influenced by interpersonal emotion regularities. While previous research has manipulated the congruency between self and others' emotional states (Trilla et al., 2021, Von Mohr et al., 2021), it is unclear how learned expectations about interpersonal emotional contingencies (IEC) shape these biases. Associative Learning and PP frameworks posit that social perception is shaped by probabilistic learning mechanisms rather than rigid simulation processes. However, this hypothesis has not been tested in a dynamic interpersonal context. To address this gap, this chapter introduced the HGF (see Chapter 1.3.2; Mathys et al., 2011) to model beliefs of temporal sources of uncertainty, i.e. social volatility, alongside perceptual ambiguity in a novel dual-task paradigm. The first experiment, conducted online, involved an emotion induction game followed by an emotion categorization task using ambiguous facial expressions, where participants implicitly learned probabilistic contingencies between induced emotional states and observed emotions in others. In addition, we included self-reported assessments of alexithymia and interoceptive sensibility. A subsequent lab-based replication study extends this approach by incorporating cardiac-related physiological arousal recordings. By integrating computational modelling, psychophysiological measures, and individual difference assessments, these two studies examined how IECs influence emotion categorization and how learning is modulated by individual differences in interoception, alexithymia and physiological responses.

Chapters 4 and 5 focused on more complex social cognition processes by investigating how adaptive empathy is underpinned by interoceptive processing. Traditional research on empathy has largely focused on static conditions, with limited studies examining how compassionate approaches can be learned in dynamic social interactions (Kozakevich Arbel et al., 2021). Additionally, few studies have investigated how interoceptive processing is influenced by socio-affective contextual information, and none has examined such effects in uncertain empathic learning environments. Moreover, while interoception has been linked to social inference and adaptive behaviour (Ondobaka et al., 2017; Quattrocki & Friston, 2014), no study has examined how modulations of interoceptive precision are linked to social learning and individual differences in empathic and autistic traits. The study of chapter 4 introduces an adaptation of an empathic learning study to examine how participants implicitly learn to respond adaptively to the empathic needs of the other based on social feedback and their 'gut feelings'. Unconscious interoceptive processing was probed using HEPs and linked to social predictions modelled by the HGF. Empathic and autistic traits were assessed with questionnaires. Thus, the study in Chapter 4 incorporated computational modelling, HEP recording, and self-reports to examine how interoceptive predictive processing underpins empathic learning under different volatility conditions and how interoceptive integration is linked to individual differences in socio-affective processing.

Chapter 5 presents a follow-up study based on the same paradigm but examining how afferent interoceptive signals influence adaptive empathic learning. Most studies linking interoception and social cognition have been correlational, without examining the causal role of afferent input in social interaction. Previous studies manipulating afferent cardiac signals or vagus activity in static social tasks, have demonstrated direct effects of interoceptive signals in emotion perception (Colzato et al., 2017; Selaro et al., 2018). VNS has also been used in associative learning in humans but with mixed results and with no application in social learning tasks (Burger et al., 2016; Kühnel et al., 2020). Moreover, transcutaneous auricular vagus nerve stimulation (taVNS) has been shown to increase interoceptive accuracy and enhance noradrenaline, which are essential for several functions, including learning, social attention, and emotion regulation. Building on past research, this last study investigates whether taVNS facilitates empathic learning, modelled by HGF, and how learning is modulated by individual differences in interoceptive sensibility, autistic traits, and affective empathy.

Chapter 2. From Interoception to Anxiety, Empathic Distress and Emotion Perception: The Role of Intolerance of Uncertainty and Alexithymia

2.1. Introduction

Navigating the world is influenced by the information by both the information we possess and the information we lack. In psychology research, this subjective lack of knowledge is described under the umbrella term uncertainty. Missing knowledge can stem from the indeterminacy and uncontrollability of future events or the ambiguity of current situations, and while usually refers to the unknowns of the external world it can also pertain to the uncertainty of one's internal bodily states and emotions. Uncertainty is typically associated with negative affect, as we feel more comfortable in situations we know and have control over - otherwise, anxiety and worry can be triggered (Anderson et al., 2019; Carleton, 2012; Gu et al., 2020). For example, people feel more stressed when 100% certain of receiving a shock than when anticipating it with a 50% probability (De Becker et al., 2016). However, uncertainty does not always cause negative emotions; for instance, not knowing what is going to happen next when reading fiction or watching a movie makes the experience more pleasurable. Importantly though, there are significant variations in how people perceive uncertainty. The aversive disposition towards uncertainty conditions have described and studied under the term Intolerance of Uncertainty (IU) (Carleton, 2012). Specifically, IU is defined as "an individual's dispositional incapacity to endure the aversive response triggered by the perceived absence of salient, key, or sufficient information, and sustained by the associated perception of uncertainty" (Carleton, 2016). This cognitive, emotional and behavioural aberrant reaction to uncertainty have been proven to be a reliable transdiagnostic risk factor presented across affective disorders, but more frequently found in anxiety disorders, e.g. generalized anxiety disorder (GAD), obsessive-compulsive disorder

(OCD) and panic disorder (Carleton, 2012; 2016). Individuals with increased IU exhibit, among others, hypersensitivity to even low levels of uncertainty, negative interpretations of unknown situations and either make intense attempts to control and reduce uncertainty or avoid any action, symptoms also found across anxiety disorders (Carleton, 2012; Gu et al., 2020). Although IU is a complex construct, it is primarily suggested to be associated with difficulties with emotion regulation, and indeed, a recent meta-analysis supports this thesis (Sahib et al., 2023).

Recently, several findings and theoretical accounts link anxiety symptomatology, among other disorders, with aberrant interoceptive processing (Palser et al., 2018; Paulus & Stein, 2010; Pollatos et al., 2007). Interoception refers to the sensing and interpretation of our internal bodily signals (Craig, 2003). Due the various operationalizations and measurement protocols of interoception, empirical evidence suggests that its relationship to anxiety is complicated (Adams et al., 2022; Domschke et al., 2010, Palser et al., 2018). Studies using questionnaires asking participants to report the intensity of their bodily sensations consistently find that increased scored on those measurement show high anxiety levels and along with increased somatization, hypervigilance and negative interpretation of their bodily signals (De Berardis et al., 2007; Olatunji et al., 2007). On the other hand, a recent study using network analysis and the subscales of the MAIA questionnaire which assesses adaptive interoceptive awareness and meta-awareness showed that increased trait anxiety to be negatively corelated to the less worry and higher trust towards ones' own bodily sensations (Slott et al., 2021). Critically, the different ways by which individuals attend to their bodies and use this information for self-regulation can be better discerned using the subscales of the MAIA questionnaire (Mehling et al., 2018). Moreover, the ability to accurate and adaptively process one's own bodily signals facilitates emotion regulation (Zamariola et al.,

2019) as well as decision-making under uncertainty (Dunn et al., 2010). The afferent bodily signals, however, are usually ambiguous and difficult to interpret. Thus, uncertainty arising from the physiological states can trigger negative interpretations and anxiety, along with sustained aversive responses, especially in individuals with high IU. Therefore, perceiving the body as a threating and unsafe place can fuel worry and increased arousal, or more generally, negative affect associated with anxiety. However, it has not yet been studied how IU can mediate the relationship between of interoceptive processing and anxiety.

Another construct related to emotion processing is alexithymia which refers to the difficulty in identifying and describing one's own emotions (Taylor, 2000). Alexithymia is associated with increased anxiety in both clinical (Berardis et al., 2008) and non-clinical populations (Karukiviki et al., 2010). In addition, alexithymia is also found to predict IU (Maisel et al., 2016; Ozsivadjian et al., 2021). Specifically, the study by Maisel and colleagues (2016) examining factors significant for the manifestation of anxiety in autistic individuals showed a close link between alexithymia and IU, as alexithymia could account for most of the IU variance when entered in the model. Another study demonstrated a mediating role of alexithymia between self-reported interoceptive sensibility and trait anxiety (Palser et al., 2018). Given that both alexithymia and IU play a role in emotion regulation and anxiety, it could be useful to better understand their relative contributions in mediating different aspects of interoceptive processing and trait anxiety.

In the social domain, the few findings we have indicate that incapacity tolerating uncertainty has also implications for social interactions (Boelen & Reijnjes, 2009; Carleton, 2010). Social encounters usually have high levels of ambiguity and unpredictability, which could be negatively interpreted, triggering in turn aversive responses, avoidance or maladaptive reactions. For instance, studies have found that IU in a central factor underlying social anxiety and negative interpretations of social events, even when controlling for other factors often correlated with social anxiety (Boelen & Reijnjes, 2009; Carleton, 2010). Even though, it appears that social interactions could be challenging for individuals with increased sensitivity to uncertainty, as happens for those with anxiety (Hirsch et al., 2004; Lamba et al., 2020), it has not been examined specifically how IU can impact social cognition aspects, such as affective empathy. Conversely, several studies link interoception to social cognition, where findings specifically suggest that individuals with better interoceptive abilities are more sensitive in perceiving others' emotions (Grynberg & Pollatos, 2015; Hübner et al., 2021; Shah et al., 2017). However, another study found negative correlations between MAIA subscales and self-reported affective empathy as measured with IRI (Stoica & Depue, 2020). Other findings have failed to confirm the link between social cognition and interoception, which can be attributed to either to the different assessments of interoception and social cognition or the implication of other factors (Ainley et al., 2015). Here, we attempted to examine IU as such a factor, since given the ambiguous nature of interoceptive signals, it could appear even more challenging for individuals with high IU to perceive an interpret them, especially when navigating social situations with increased uncertainty. Difficulties in self-regulation from poor interoceptive skills and incapacity to deal with uncertainty can, in turn, also pose challenges in regulating or processing others' emotions.

Computational approaches in cognitive psychology research have widely applied to reveal behavioural patterns not easily discerned with traditional methodologies (Gupta et al., 2022; Montague et al., 2012; White et al., 2010). One such model is the Diffusion Drift Model (DDM) which is used to examine decision-making in tasks two available option with noisy incoming information (White et al., 2010). Based on the reaction time distributions of the two alternative choices, it provides several parameters on the underlying decision-making processes. In short, it models the process of accumulating information until a response reaches a set boundary and this is performed around at least 4 basic parameters: 1) the drift rate, which relates to efficiency of processing each stimulus, 2) the boundary separation, which reflects the amount of evidence needed to reach a decision regardless of the response, 3) the bias, which reflects how easy it is to choose one response over the over, and 4) the non-decision time, which quantifies the time before and after making the decision (White et al., 2010). DDM approaches have recently been employed to elucidate differences in decision-making in clinical and non-clinical populations (Dietel et al., 2021; Gupta et al., 2022, Pedersen et al., 2017, White et al., 2010). For instance, a study (Dietel et al., 2021) examining decision-making among individuals with SAD, GAD, body dysmorphic disorder (BDD) and non-clinical controls found those with SAD and BDD exhibited a greater explicit negative response bias compared to the GAD and non-clinical groups in social and body-related decision-making. Further analysis with DDM showed that this difference was attributed to an implicit bias, reflected in drift rates and associated with faster rejection for negative values and slow endorsement of positive ones. However, similar modelling approaches have not yet been used to examine how IU could impact decision-making. Likewise social anxiety, increased IU can have several implications in perceptual processing, attention, decision-making and behaviour in a situation when there is need to categorise ambiguous social stimuli, for instance, the emotional expressions of another person (Anderson et al., 2019; Carleton, 2012). Thus, using DDM in an ambiguous emotion recognition task, we can examine whether social difficulties in individuals with IU relate to reduced attention and processing of specific emotional expressions (reflected on drift rates), behavioural inhibition (reflected on boundary separation) or negative response/perceptual bias (reflected on the bias parameter).

The first objective of this study was to examine how IU in tandem with alexithymia traits mediate the relationship between interoception awareness and anxiety, as well as how aspects of the interoceptive awareness relate to IU. The second objective was how IU influences social interactions either directly or as a mediating factor along with alexithymia between interoception and individual differences affective empathy and perception of others' emotions. To that aim, we collected online responses in questionnaires measuring interoceptive awareness, alexithymia, IU, trait anxiety and affective empathy. To probe social cognition, apart from self-reports, we used a 2 alternative forced choice emotion categorization task with combinations of ambiguous emotional expressions as stimuli and then modelled the RTs obtained using DDM. Last, we constructed different model based on two previous outlined objectives and performed path analysis to examine direct and indirect pathways among the measured traits and behavioural responses. Specifically, our main hypotheses are the following. Firstly, we expect that both IU and alexithymia mediate the relationship between aspects of interoceptive awareness and anxiety. Secondly, we expect that IU would predict social cognition, as measured in the affective empathy questions and in the emotion recognition task. Third, we expect that IU and alexithymia mediate the relationship between interoceptive awareness and affective empathy and emotion recognition.

2.2. Methods

2.2.1. Participants

In Study 1 we recruited 142 participants (aged 18-48, 120 females) and in Study 2, we recruited 153 participants (aged 18-39, 117 females). They were students at the School of Psychology and participated in exchange for credits in the research participation scheme of the University of Kent. Participants were "healthy volunteers" with no history of psychiatric or neurological disorder and provided written informed consent before the beginning of the experiments. Anonymity was ensured by using a unique code for each participant and with no reference to their real personal information. The studies were approved by the School of Psychology Ethics Committee at University of Kent.

The sample sizes were determined based on power analyses for multiple regression models with 7 predictors, assuming a moderate effect size ($f^2 = 0.15$). This analysis indicated that approximately 103 subjects would be sufficient to achieve 80% power at an alpha level of 0.05.

2.2.2. Questionnaires

Interoceptive sensibility

To measure interoceptive sensibility we used the Multidimensional Assessment of Interoceptive Awareness, Version 2 (MAIA-2, Mehling et al., 2018). This test includes 37 questions assessing the different ways people process and pay attention to their bodily sensations. Specifically, it consists of 8 subscales: noticing (awareness of uncomfortable, comfortable, and neutral body sensations), not-distracting (tendency to ignore or distract oneself from sensations of pain or discomfort), not-worrying (emotional distress or worry with sensations of pain or discomfort), attention regulation (ability to sustain and control attention to body sensation), emotional awareness (awareness of the connection between body sensations and emotional states), self-regulation (ability to regulate psychological distress by attention to body sensations), body listening (actively listens to the body for insight), and trust (experiences one's body as safe and trustworthy). Participants were asked to indicate how often each statement applied to them generally in daily life, in a 5-point Likert scale ranging from 0 (never) to 5 (always).

Alexithymia

To assess participants emotion processing traits, we used the 20-item Toronto Alexithymia Scale (TAS-20; Bagby et al. 1994). This test measures Alexithymia with questions like "I am often confused about what emotion I am feeling" where participants indicate their level of agreement in a 5-point scale.

Interpersonal Reactivity Index (IRI)

The IRI is a self-report questionnaire that measures difference in empathic social interaction and included 4 subscales assessing perspective taking, fantasy(imagination), empathic concern and personal distress. The first two subscales assess cognitive empathy while the last two affective empathy. It includes 28 items where participants respond in 5-point Likert-type scale from 0 (does not describes me well) to 4 (describes me very well).

Intolerance of uncertainty

To measure IU, we used the 27-item Intolerance of Uncertainty Scale developed Freeston and colleagues (1994). It comprises two subscales the first is described as "uncertainty has negative behavioral and self-referent implications" tapping mainly into self-related behavioural difficulties and the second as "uncertainty is unfair and spoils everything" pertaining to more general negative perceptions of uncertainty. One item, for example, of the first subscale is “Uncertainty stops me from having a firm opinion” and one item of the second subscale is “Uncertainty makes life intolerable”. Respondents report the degree to which each item applies to them on 5-point Likert-type scales ranging from 1 (not at all characteristic of me) to 5 (entirely characteristic of me).

Trait anxiety

To assess the general tendency to feel anxiety we used the State Trait Anxiety Inventory - Trait (STAI-T; Spielberger, 1985). The STAI-T includes 20 items where respondent report how they generally feel based on a 4-point Likert-like scale ranging from 1 (not at all) to 4 (very much so).

For the first study, we used the MAIA, IUS, TAS-20, and STAI-T. For the second study, we used the same questionnaires with the addition of IRI.

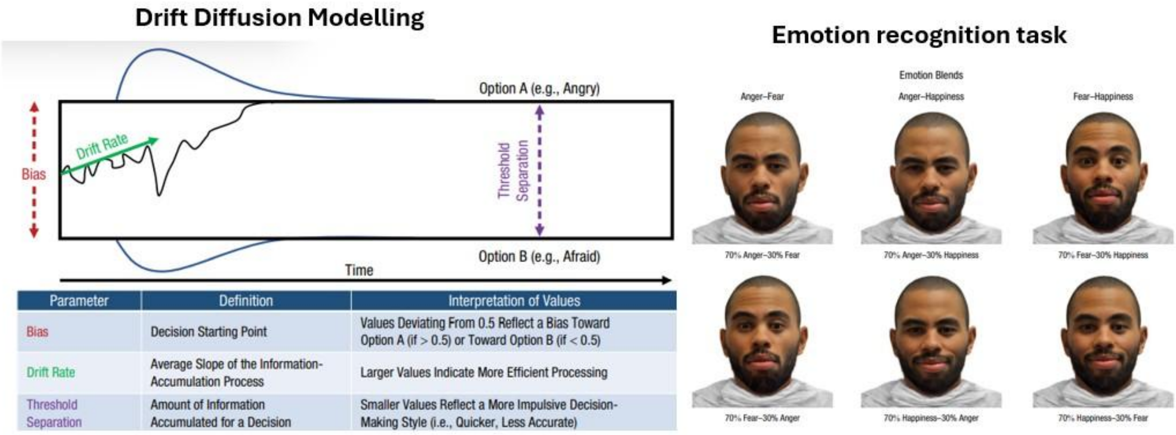
2.2.3. Emotion recognition task

The emotion recognition task was a 2-alternative forced-choice task where participants had to identify the presented emotion in a series of combinations of ambiguous emotional faces. The task was adapted, with minor changes, from the study of Brennan and Baskin-Sommers (2020), where a DDM approach was also used.

Stimuli

The stimuli were emotional face images from the dataset of the Racially Diverse Affective Expression (RADIATE) face stimulus set (available at http://fablab.yale .edu/page/assays-tools; Conley et al., 2018; Tottenham et al., 2009). We used images from 10 male models from 3 racial backgrounds: 4 Black, 3 Hispanic and 3 White. Stimuli were created by blending images of two emotional expressions at a 30%-70% level (Fig 2.1). This level of blending ensured that the categorization was of moderate difficulty, providing average accuracy performance appropriate for diffusion drift modelling (Ratcliff & Mckoon, 2008). We choose three emotions to generate the following three emotion blends: anger-happiness, anger-fear, and happiness-fear. For each blend type, one of the two emotions was the dominant one. For instance, for the anger-happiness blending, the two sets of stimuli were the 30% happiness and 70% anger morphing and the 30% anger and 70% happiness morphing.

The task included three blocks based on these combinations: the anger-happiness block (block 1), the anger-fear block (block 2), and the fear-happiness block (block 2). In each block only two emotions were presented and with the same frequency. Each block included 40 trials, so we had 120 trials in total, 80 for the blocks of interest.



Parameter	Definition	Interpretation of Values
Bias	Decision Starting Point	Values Deviating From 0.5 Reflect a Bias Toward Option A (if > 0.5) or Toward Option B (if < 0.5)
Drift Rate	Average Slope of the Information-Accumulation Process	Larger Values Indicate More Efficient Processing
Threshold Separation	Amount of Information Accumulated for a Decision	Smaller Values Reflect a More Impulsive Decision-Making Style (i.e., Quicker, Less Accurate)

Figure 2.1. Drift Diffusion Modelling (DDM) and Emotion Recognition Task. The DDM schematic represents decision-making dynamics, where the drift rate reflects the efficiency of evidence accumulation, bias represents a predisposition towards one response, and threshold (or boundary) separation determines the level of information required before committing to a decision. The emotion recognition task involved categorizing ambiguous facial expressions, with stimuli varying in emotional blends between anger-fear, happiness-anger and happiness-fear across three different blocks. Images taken from Brennan and Baskin-Sommers (2020) study.

Task procedure

This online emotion categorization task was created on Psychopy, and uploaded and presented on the Pavlovia platform. Participant were asked to respond in each block as fast and as accurately as possible and within 3 s from stimulus onset. After each response they had also to report their confidence for each decision in a scale ranging from 1 (not confident) to 5 (very confident). Before the presentation of the stimuli a fixation cross was shown for 500 ms. There was a variable intertrial interval lasting between 1.5 s and 2 s. In the beginning of the task participants had to complete 12 practice trials. Between the blocks we had two breaks for participants to rest. The total duration of the task was around 20 min.

2.2.4. Data analysis

Drift Diffusion Modelling

Diffusion models conceptual the decision-making process in 2AFC tasks as a process of continuous stochastic accumulation of information over time until evidence reaches one of the two boundaries that are associated with each response. This modelling approach is useful for the examination of implicit processes underlying decision-making when stimuli are ambiguous, time is limited, and decisions are difficult. DDM can account for the RT distributions and response decisions based on 4 parameters. First, the drift rate (v) models the speed of evidence accumulation towards one boundary. Drift rates can be affected by either endogenous cognitive factors such as attention or exogenous factors such as stimulus characteristics. Second, the boundary separation (a) defines the distance between the decision boundaries. It reflects the amount of evidence required to make a decision. Larger boundary separation means more evidence is needed, leading to slower but usually more accurate decisions. Third, the starting point or initial bias (z), this parameter indicates the initial position of the evidence accumulation process relative to the decision boundaries. It can reflect a bias towards one decision over another if it is not centred. Forth, the non-decision time (t-er), a parameter accounting for the time taken by processes other than evidence accumulation, such as sensory processing and motor response. It is the time before and after the decision process.

In this study, we modelled data from the first (angry-happiness) and third (fear-happiness) blocks for diffusion and regression analyses, while excluding block 2. Data from block 2 were analysed in a separate study. By combining data from blocks 1 and 3, we increased statistical power in the diffusion analysis and simplified the subsequent analyses. Specifically, we assumed a single decision process where participants identified whether a face was happy or unhappy (expressing fear or anger), effectively categorizing based on perceived valence (positive or negative facial expressions). Responses from the first and third blocks were combined, with happiness expressions presented in both blocks, anger in the first block, and fear in the third block. This categorization based on expressed valence for ambiguous stimuli is relevant to IU research, given the reported negative affective predictions for individuals with high IU in ambiguous situations (Carleton, 2016).

The decision process was modelled using the DDM, with the upper boundary representing negative expressions and the lower boundary representing positive expressions. We assumed different drift rates for each emotion type, as suggested in similar modelling procedures (Brennan & Baskin-Sommers, 2020; Myers et al., 2022; Ratcliff & McKoon, 2008). The other three DDM parameters—bias, boundary separation, and non-decision time—did not vary with stimulus characteristics. Thus, we used five parameters that varied across participants to model the decision process with DDM.

While there are other versions of DDM with additional parameters, we followed the most common diffusion modelling approach using only five parameters, adhering to guidelines from previous studies and tutorials (Brennan & Baskin-Sommers, 2020; Myers et al., 2022). Our data was modelled using the Rwiener package in R (Wabersich & Vandekerckhove,2014). Parameters were estimated using the Maximum Likelihood

Estimation (MLE) method, which is fast, accurate, and suitable for smaller trial numbers (<40) as in our study (Deng et al., 2018). To avoid settling on local maxima, we increased the number of iterations from the default 500 to 1000. We set neutral initial starting values for all estimates and ran diffusion analysis to obtain parameter estimates for each participant.

Data quality control

From the 154 participants that took part in the study, 30 were excluded from the main analyses as the failed to pass 2 out of 3 catch trials or did not respond to 20% of the trials. So, the final sample for analysis for the emotion recognition task included 124 participants. We excluded trials with RT<0.2 (less than 4% of total trials), as guidelines for diffusion analysis recommend (Myers et al., 2022), and those without in-time response.

Correlation analysis

As a preliminary step of our statistical analysis, we run zero-order correlations between along variables included in the two studies (Appendix A.2). These zero-order correlations were subsequently used to construct the models for path analyses. For the variable selection we considered which subscales of MAIA significantly correlate with the DVs as well as the mediators (IU and alexithymia).

Path analysis

Path analysis is a statistical technique used to describe the direct and indirect dependencies among a set of variables. It is a form of multiple regression analysis that allows to evaluate causal models by examining the relationships between a dependent variable and many independent variables, taking into account the covariance among all variables. This approach also allows to examine both direct and indirect dependencies between variables. Path analysis is similar to structural equation modelling with the main difference that it does not include latent variables. Model fitting was assessed with the following criteria: chi-square likelihood-ratio test p-value ≥ 0.05, root mean square error of approximation (RMSEA) ≤ 0.08 and comparative fit index (CFI) ≥ 0.95 (Hu & Bentler, 1999). This analysis was performed in R using the lavaan package (Rosseel, 2012).

The first models (*Model 1a*: Study 1; *Model 1b*: Study2) were created to test our first hypothesis examining the underlying factors linked to trait anxiety and the significant direct and indirect pathways. Thus, this model included at the lower level, as exogenous variables, the dimensions of the MAIA showing zero-order correlations with anxiety. Then, we included Alexithymia at the next level and, lastly, IU before the outcome measure trait anxiety (Fig 2.1). Variables from each level were linked with regression paths both directly and indirectly to trait anxiety. The second model (*Model 2 – IRI*), based on the second hypothesis, was similar to the one described above but instead of anxiety we used the two subscales of affective empathy of IRI, Empathic Concern and Personal Distress as outcome measures in different models. These variables should be considered separately, and not as a composite score, because the former is thought to reflect other-oriented concern for others and the latter self-oriented distress during empathic situations. Then, we created two additional models (Study 2) using again the MAIA subscales, alexithymia and IU as predictors, and instead of the IRI we used the behavioural performance in the emotion recognition task as DVs. Specifically, the accuracy scores for each emotion category (i.e. positive vs negative) were added as outcome variables in separate models, the MAIA subscales with significant zero-order correlations with the accuracy scores as exogenous predictors and IU and alexithymia as mediators (*Model 3 - Emotion Recognition – Accuracy).* Similarly, in the final models (*Model 4 - Emotion Recognition – DDM*), the parameters extracted from the DDM were used as outcome variables.

2.3. Results

Model 1a (Anxiety)

For the first model (Study 1), based on the first hypothesis, included the 5 MAIA subscales (Not Worring, Not Distracting, Emotional Awareness, Noticing, Trust) with significant zero-order correlations with anxiety, alexithymia and IU scores (Fig.2.1). This model had acceptable fit to the data (χ^2 (13)=23.594, p=0.035, CFI=0.976, and RMSEA=0.065) and explained 59.0% of the variance in anxiety (R^2=0.590), which corresponds to a large effect size (f^2=1.44).

The direct regressions showed that anxiety was negatively predicted by the MAIA subscales Not Worrying (β=-0.478, SE=0.153, p=0.002) and Trust (β=-1.466, SE=0.192, p<0.001) and positively predicted by Noticing (β=0.508, SE=0.212, p=0.017), alexithymia (β=0.232, SE=0.051, p<0.001) and IU (β=0.146, SE=0.031, p<0.001). IU was negatively predicted by Not Worrying (β=-1.760, SE=0.339, p<0.001) and Not Distracting (β=-0.527, SE=0.255, p=0.039), and positively by alexithymia (β=0.512, SE=0.114, p<0.001). Lastly, alexithymia was negatively associated with Not Worrying (β=-0.426, SE=0.214, p=0.046), Not Distracting (β=-0.431, SE=0.159, p=0.007) and Trust (β=-0.945, SE=0.280, p=0.001). The remaining associations were not significant (ps>0.05).

Regarding the indirect paths, we found that alexithymia mediated the relationship between Not Distracting and anxiety (β=-0.093, SE=0.041, p=0.022) as well as Trust and anxiety (β=-0.203, SE=0.077, p=0.008). In addition, IU was significant mediator between Not Worrying and anxiety (β=-0.945, SE=0.280, p=0.001). Lastly, we found two additional indirect paths from MAIA subscales to anxiety mediated by both alexithymia and IU. Specifically, the indirect path from Not Distracting to anxiety, passing through alexithymia and to IU, was significant (β=-0.034, SE=0.016, p=0.035) as well as the path from Trust to anxiety, passing though alexithymia and IU (β=-0.074, SE=0.031, p=0.017). Study 2 significantly replicated these results (see Appendix A.1.).

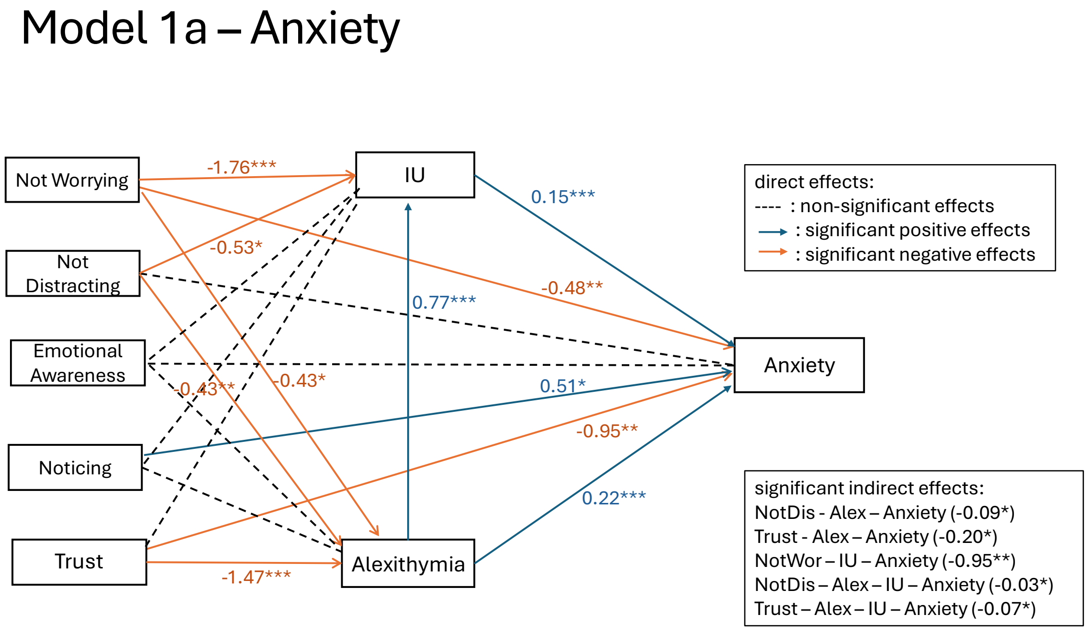

Model 1a – Anxiety
Not Worrying
Not Distracting
Emotional Awareness
Noticing
Trust
IU
Alexithymia
Anxiety
-1.76***
-0.53*
-0.43**
-0.43*
-1.47***
0.77***
0.15***
-0.48**
0.51*
-0.95**
0.22***
direct effects:
---- : non-significant effects
: significant positive effects
: significant negative effects
significant indirect effects:
NotDis - Alex – Anxiety (-0.09*)
Trust - Alex – Anxiety (-0.20*)
NotWor – IU – Anxiety (-0.95**)
NotDis – Alex – IU – Anxiety (-0.03*)
Trust – Alex – IU – Anxiety (-0.07*)

Figure 2.2. Path Analysis Results for Anxiety Model. This figure presents the path analysis examining the direct and indirect effects on trait anxiety (STAI). Model predictors are 5 dimensions of MAIA (Not Worrying, Not Distracting, Emotional Awareness, Noticing, Trust), alexithymia (TAS-20) and Intolerance of Uncertainty (IU scale). The negative and positive direct and indirect effects are accompanied by regression coefficients (β), with asterisks denoting the significance level (*p<0.05, **p<0.01, *p<0.001).

Model 2 (Affective Empathy)

For the second model, the zero-order correlations (Appendix A.2) revealed that Empathic Concern (EC) was positively correlated with Noticing, Not Distracting and Emotional Awareness of MAIA but with no correlations with IU or alexithymia. On the other hand, Personal Distress (PD) was negatively correlated with Not Worrying, Emotional Awareness, Attention Regulation, Self-Regulation, Body Listening and Trust, and positively with IU and TAS. Based on this correlation we created a model with PD as outcome variable and the MAIA dimensions correlating both with PD and either IU or alexithymia (Fig 2.2).

The model had good fit data (χ^2 (5)=5.222, p=0.389, CFI=0.999, and RMSEA=0.018) and explained 9.1% of variance in EC (R^2=0.091), indicating a small effect (f^2=0.10); in contrast, it explained 48.1% of variance in PD (R^2=0.481), corresponding to a large effect size (f^2=0.93). Path analysis revealed that PD was negatively predicted by Not Worrying (β=-0.1.08, SE=0.394, p=0.005), and positively by IU (β=0.070, SE=0.021, p=0.001) and alexithymia (β=0.105, SE=0.037, p=0.005).

Model 2 – Personal distress

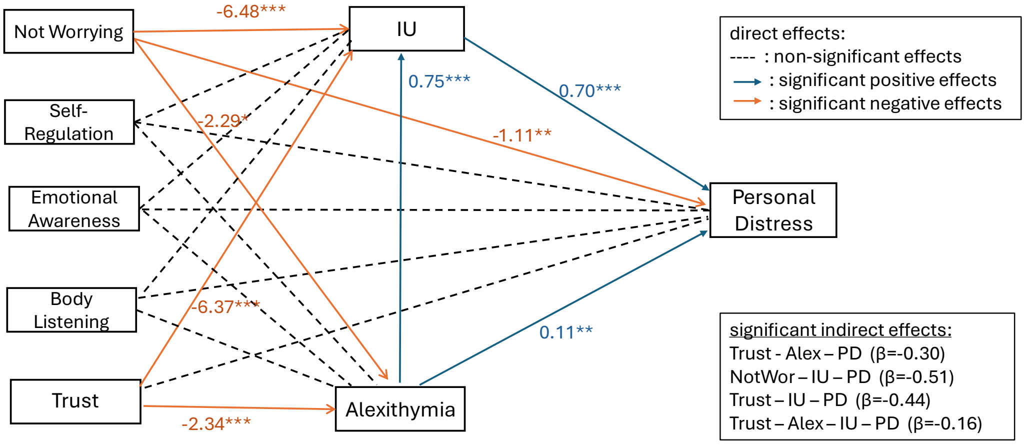

Not Worrying
Self-Regulation
Emotional Awareness
Body Listening
Trust
IU
Alexithymia
Personal Distress
-6.48***
-2.29*
-1.11**
-6.37***
-2.34***
0.75***
0.70***
0.11**
direct effects:
---- : non-significant effects
→ : significant positive effects
→ : significant negative effects
significant indirect effects:
Trust - Alex – PD (β=-0.30)
NotWor – IU – PD (β=-0.51)
Trust – IU – PD (β=-0.44)
Trust – Alex – IU – PD (β=-0.16)

Figure 2.3. Path Analysis Results for Affective Empathy Model. This figure presents the path analysis examining the direct and indirect effects on Personal Distress of IRI. Model predictors are 5 dimensions of MAIA (Not Worrying, Self-Regulation, Emotional Awareness, Body Listening, Trust), Alexithymia (TAS-20) and Intolerance of Uncertainty (IU scale). The negative and positive direct and indirect effects are accompanied by regression coefficients (β), with asterisks denoting the significance level (*p<0.05, **p<0.01, *p<0.001).

Regarding the indirect effects, we found that the mediation of alexithymia between Trust and PD was marginally not significant (β=-0.303, SE=0.155, p=0.050) while the mediations of IU between Not Worrying and PD (β=-0.502, SE=0.192, p=0.009) as well as between Trust and PD (β=-0.444, SE=0.180, p=0.014) were significant. In addition, similarly to what was observed in the previous models, we also found the path from Trust to alexithymia and IU to PD (β=-0.160, SE=0.081, p=0.048) to be significant.

Model 3 (Emotion Recognition – Accuracy)

Zero-order correlations revealed only a significant association between accuracy in the emotion of positive emotions and the Attention Regulation subscale of the MAIA and no correlation between accuracy and the mediators (IU, Alexithymia). Thus, the criteria set to build the previous models were not met here. Therefore, as exploratory analyses, we built the models using the subscales of MAIA closer to significance along with the mediators IU and TAS, keeping the same structure with the previous models. In this way, we could examine how path analysis might reveal significant associations when taking in account the covariances among the predictors. First, we structured two models predicting the accuracy score for negative (Acc-Neg) and positive (Acc-Pos) emotional expressions separately. The model predicting Acc-Neg was did not yield any significant direct or indirect effects, so we report here only the one with Acc-Pos as DV (Fig. 2.3).

This model had good fit to the data: χ^2 (5)=5.462, p=0.362, CFI=0.998, and RMSEA=0.027. The variance explained for Acc-Neg was 7.7% (R^2=0.055) with small effect size (f^2=0.08) and 13.9% (R^2=0.139) for Acc-Pos with moderate effect size (f^2=0.16). Regarding the direct effects, we found that alexithymia predicted accuracy negatively (β=-0.003, SE=0.001, p=0.014) while IU was positively associated with accuracy (β=0.001, SE=0.001, p=0.023). No MAIA subscales significantly predicted accuracy (p>0.86). Regarding the indirect effects, IU mediated the effect from Not Worring (β=-0.009, SE=0.005, p=0.046) and Trust (β=-0.009, SE=0.005, p=0.049) on accuracy.

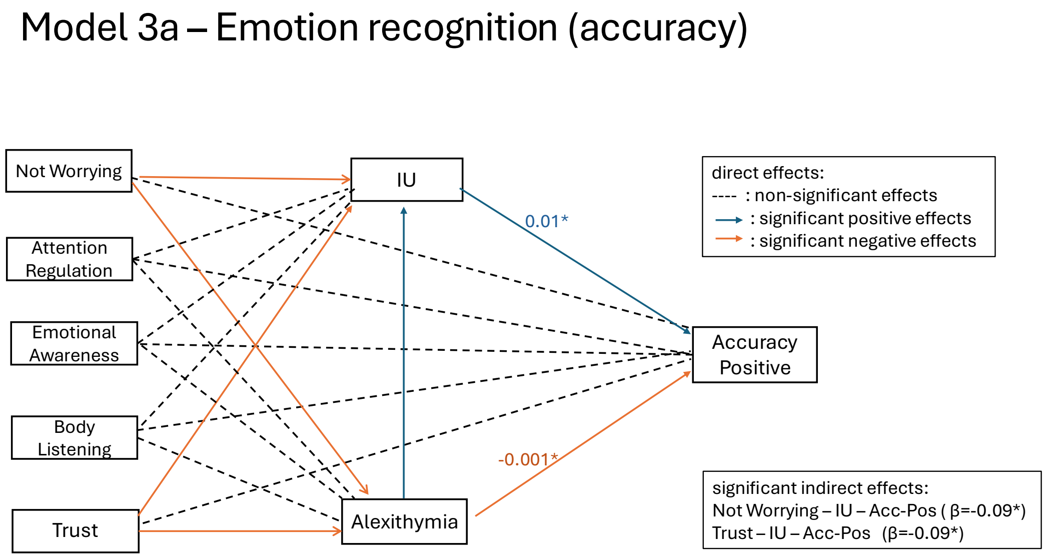

Model 3a – Emotion recognition (accuracy)
Not Worrying
Attention Regulation
Emotional Awareness
Body Listening
Trust
IU
Alexithymia
Accuracy Positive
0.01*
-0.001*
direct effects:
---- : non-significant effects
: significant positive effects
: significant negative effects
significant indirect effects:
Not Worrying – IU – Acc-Pos (β=-0.09*)
Trust – IU – Acc-Pos (β=-0.09*)

Figure 2.4. Path Analysis Results for the Emotion Recognition (accuracy) Model. This figure presents the path analysis examining the direct and indirect effects on emotion recognition accuracy for positive stimuli. Model predictors are 5 dimensions of MAIA (Not Worrying, Attention Regulation, Emotional Awareness, Body Listening, Trust), alexithymia (TAS-20) and Intolerance of Uncertainty (IU scale). The negative and positive direct and indirect effects are accompanied by regression coefficients (β), with asterisks denoting the significance level (*p<0.05, **p<0.01, *p<0.001).

Model 4 (Emotion Recognition – DDM)

A similar explorative approach was adopted here, using the 4 DDM (drift-neg, drift-pos, a, z) parameters as outcome variables in models with a similar structure to the previous ones. Of these 4 models only the one with the drift rate for the positive emotions (drift-pos) yielded significant results. The model predicting d-positive had good fit to the data: χ^2 (5)=5.462, p=0.362, CFI=0.998, and RMSEA=0.027. The variance explained for drift-neg was 5.5% (R^2=0.055) with small effect size (f^2=0.06) and 12.3% (R^2=0.123) for drift-pos with small-to-moderate effect size (f^2=0.14).

Regarding the direct effects, Attention Regulation was marginally not significant predictor of drift-pos (β=0.046, SE=0.023, p=0.051), while, likewise the accuracy model, alexithymia negatively predicted drift-pos (β=-0.008, SE=0.002, p=0.028) and IU positively (β=0.005, SE=0.002, p=0.024). With respect to the indirect effects, the path from Not Worring to IU to d-positive was significant (β=0.034, SE=0.017, p=0.014) and the path from Trust to IU to d-positive was marginally not significant (β=0.032, SE=0.016, p=0.050).

Model 3a – Emotion recognition (drift rate)

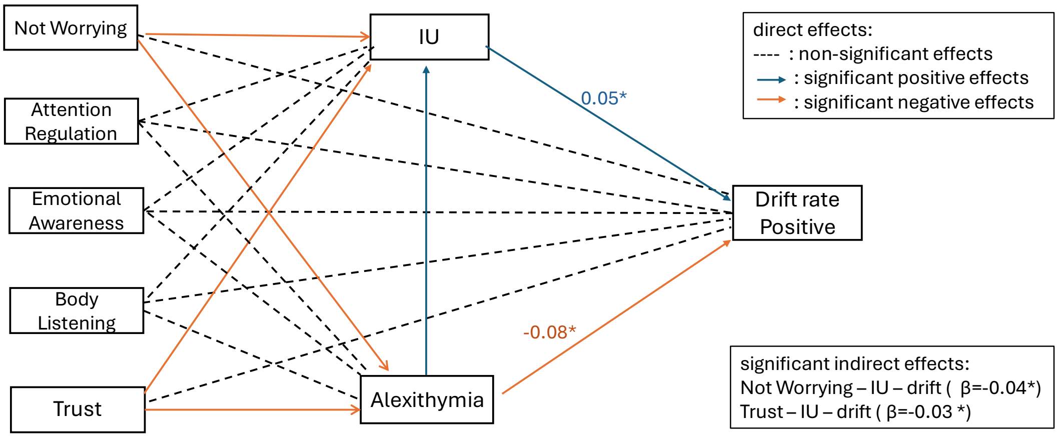

Not Worrying
Attention Regulation
Emotional Awareness
Body Listening
Trust
IU
Alexithymia
Drift rate Positive
0.05*
-0.08*
direct effects:
---- : non-significant effects
: significant positive effects
: significant negative effects
significant indirect effects:
Not Worrying – IU – drift (β=-0.04*)
Trust – IU – drift (β=-0.03 *)

Figure 2.5. Path Analysis Results for the Emotion Recognition (drift) Model. This figure depicts the path analysis examining the direct and indirect effects on emotion recognition drift-rate for positive stimuli. Predictors include 5 dimensions of MAIA (Not Worrying, Attention Regulation, Emotional Awareness, Body Listening, Trust), Alexithymia (TAS-20) and Intolerance of Uncertainty (IU scale). The negative and positive direct and indirect effects are accompanied by regression coefficients (β), with asterisks denoting the significance level (*p<0.05, **p<0.01, *p<0.001).

2.4. Discussion

Interoceptive abilities are a key factor in the processing of affective information, related either to the self or others, and play an important role in the development of clinical conditions like anxiety (Tsakiris & Critchley, 2016). It is believed that accessing and understanding own internal bodily states helps guiding the individual through their affective states and reactions during emotional and social situations. This may be particularly important in conditions of great uncertainty, when understanding own bodily and affective responses may provide cues to deal with the ambiguity. However, given that interoceptive states themselves may often feel illusive and ambiguous, incorrect or insecure appraisal of bodily reactions can, actually, contribute to increase the perception of uncertainty and associated feelings of anxiety. However, few studies, have yet examined how interoceptive and emotional processing are related to perceptions of, and attitudes towards, uncertain conditions. The main aim of the present study was to examine how IU and alexithymia mediate the relationship between different aspects of interoceptive sensibility and trait anxiety (Study 1) or social cognition (Study 2). Our findings expand previous studies associating different aspects of interoceptive processing with anxiety (e.g. Palser et al, 2018; Slotta et al, 2021; Vabba et al, 2023) by revealing the mediating role of alexithymia and IU. Interestingly, we also found that these two traits also mediate the relationship between interoception and social abilities, measured through self-reported questionnaires of empathy or by the emotional recognition of ambiguous facial expressions in a behavioural task.

Anxiety

Our findings corroborate previous studies using the MAIA questionnaire to probe the link between adaptive processing of interoceptive sensations and anxiety (Lee et al., 2024; Paulus & Stein, 2010; Slota et al., 2021). Here, we found that the MAIA dimensions reflecting less worrying and more trust on bodily sensations, as well as the ability to regulate emotions via intentional attention to the body, are associated with reduced anxiety. Conversely, increased perception of bodily signals, as measured by the Noticing and Emotional Awareness subscales, predicted higher anxiety levels. These findings highlight the importance of interoceptive appraisal, emotion regulation and anxiety. If the increased perception of bodily signals, especially in negative contexts, can enhance anxiety-related sensations (ref needed), possibly due to excessive focus on bodily states, being able to attend to, and experience, one's body as safe and trustworthy can help to reduce the ambiguity inherent to many affective and social situations and promote a coherent understanding of own affective responses.

This idea is supported by the observed roles of IU and alexithymia in the mediation between interoceptive processing and anxiety. Firstly, in the two studies performed, we demonstrate that individuals reporting higher trait anxiety also tend to exhibit lower emotional awareness and tolerance of uncertainty. Our results also confirm and extend previous findings, with different interoception measurements, showing that alexithymia mediates the relationship between interoceptive processing and anxiety (Palser et al., 2018). Specifically, here we found that alexithymia mediates the negative relationship between trust in bodily sensations and anxiety, suggesting that difficulties in experiencing one's body as trustworthy contributes to difficulties in understanding own emotions and anxiety.

Regarding IU, our findings are largely in line with a recent study examining the moderating role of IU in the relationship between several MAIA dimensions and body dissatisfaction (Bijsterbosch et al., 2023). In particular, while initial results showed several MAIA subscales to be correlated with both IU and anxiety, the path analysis model revealed that IU mediated the relationship between the dimensions Not Worrying, Trust and anxiety. The Not Worrying subscale assess the aversive emotional reactions to detected bodily signals due to the negative interpretations of them, while the Trust subscale assess the meta-awareness of bodily feelings, specifically, the sense of trust in one's body (Mehling et al., 2018). Thus, it is easy to consider how these may resonate with IU which is characterised by negative interpretations of the unknow and feelings of worry about the future (Carleton, 2016).

Individuals with high IU, struggling with uncertainty and may also worry excessively about ambiguous bodily sensations, interpreting them as signs of potential threats or health issues, which can in turn fuel feelings of anxiety Thus, our findings highlight specific affective and metacognitive paths by which IU and alexithymia mediate facets of interoceptive processing associated with trait anxiety.

Recent studies have implicated both IU and alexithymia with a range of body-related clinical conditions or symptoms, suggesting that these constructs may play a key role in the development or maintenance of a variety of conditions. For example, IU and Alexithymia were associated with somatic symptoms in autistic and neurotypical individuals (Larkin et al., 2023), body dissatisfaction (Bijsterbosch et al., 2023), anorexia nervosa (Abbate-Daga et., 2015) and externalizing and internalizing behaviour in autism (Ozsivadjian et al., 2021). Our findings expand on this by revealing the combined contribution of alexithymia and IU to anxiety, specifically through the indirect path from Trust to anxiety through alexithymia and IU (i.e. Trust → Alexithymia → IU → Anxiety). Individuals with alexithymia may struggle with IU because their emotional processing deficits hinder their ability to cope with uncertainty. This can create a feedback loop where the inability to identify and express emotions leads to heightened anxiety, particularly in uncertain situations. In other words, the lack of emotional clarity and trust on bodily sensations can exacerbate feelings of unease and anxiety when faced with ambiguity.

Affective empathy

The second objective of this study was to examine how different aspects of interoceptive processing affect social skills through feelings about uncertainty and emotional awareness. For that, we first explored how these traits predict the affective dimensions of empathy measured through the EC and PD scales of the IRI. Zero-order correlation analysis showed that EC correlated with some MAIA subscales (Noticing, Not Distracting, Emotional Awareness) but it did not correlate with either IU or alexithymia scores, and therefore was not included in path analysis. Conversely, PD showed moderate to high negative correlations with IU, Alexithymia and several MAIA subscales (Not Worrying, Self-Regulation, Emotional Awareness, Body Listening, Trust). However, only the Not Worrying subscale was a significant direct predictor of PD on the path model. These findings are in line with a previous study with respect to the direction and magnitude of the relationships between EC/PD and the MAIA subscales (Stoica & Depue, 2020). It should be noted that EC, which refers to feelings of compassion and concern for others, contrasts with PD, which involves feelings of anxiety and discomfort in response to others' distress. Thus, it is not surprising that IU and Alexithymia are strongly associated with the latter and not the former.

We also found that IU and alexithymia mediated, separately and together, the effect of interoceptive processing on PD. Specifically, IU mediated the relationships between the MAIA dimensions of Not Worrying and Trust and PD. In fact, IU and alexithymia were found to fully mediate the effect of Trust on PD, as Trust did not directly predict PD. Not trusting bodily responses my contribute feelings of distress in empathic situations, but only in individuals particularly aversive to uncertainty or with difficulties identifying own emotions. Conversely, trusting our bodily sensations might help mitigating feelings of worry in uncertain situations and enable better understanding of emotions, reducing, in turn, empathic distress. It is worth noting the high consistency in the results observed in studies 1 and 2 as, not only both PD and trait anxiety were found to be negatively predicted by the interoceptive dimensions Not Worrying and Trust, as in both studies alexithymia and IU showed the same mediating roles on these relationships. This is not surprising as PD refers to feelings of anxiety and discomfort in response to others' suffering and has been previous shown to be heightened in people with high trait anxiety (Nair et al., 2024). Individuals who have difficulties in identifying their feelings and worry about bodily sensations, tend to experience more anxiety in social interactions and empathic personal distress (Cosmoiu and Nedelcea, 2022). IU could mediate this relationship, as high IU could exacerbate worry and difficulty in understanding and managing bodily sensations and, in turn, further fuel feelings of discomfort and incompetence when managing others' distressing emotions.

Emotion recognition

The two models predicting the perception of ambiguous facial expressions, either based on accuracy or the DDM parameters, revealed a similar pattern of factors involved. Firstly, we did find any IS dimension predicting emotion recognition performance as it was revealed in a previous study using the MAIA and dynamic emotional facial expressions. However, our findings showing that higher alexithymia was associated with lower performance, accord with previous results indicating impaired social cognition and emotion recognition capacity in individuals with difficulties identifying and understanding own emotions (Cook et al., 2013; Di Tella et al., 2020; Sunahara et al., 2022). Contrary to expectations, individuals with higher IU showed better accuracy and better efficiency in processing happy, but not negative, facial expressions. That is, given its relationship with anxiety and negative affect, it could be reasonable to expect enhanced vigilance and sensitivity towards threat-signalling stimuli in individuals with high IU, but instead we found a facilitation in the identification of happy expressions. There was, however, no bias towards positive expressions (as indexed by the DDM parameter) that could reflect a tendency to identify or prefer happy emotions.

Instead, our results suggest the involvement of distinct perceptual processes beyond preferences or avoidance.

To the best or our knowledge, only one study so far tried to relate IU with emotion recognition abilities and found no correlation between the two (Monferrer et al, 2023). There are, however, several studies suggesting better emotion recognition in people with anxiety-related disorders, at least, under specific conditions. Most often, anxiety is related to enhanced processing of threat-related emotions (e.g. Cooper et al., 2008; Surcinelli et al., 2006) but several studies have also found greater emotion recognition ability irrespective of emotional valence (e.g. Gui et al, 2017; Kolassa et al, 2009). For example, Gui and colleagues (2017) found enhanced detection of emotional expressions in individuals with generalised anxiety disorder, compared to social anxiety and controls. Another study found stronger responses in the visual cortex in individuals with higher social anxiety disorder irrespective of stimulus emotional valence (Kolassa et al., 2009). It may be that the hypervigilance that characterises anxiety, and IU, and is typically associated only to stimuli with negative valence, could actually reflect increased attention towards social stimuli, possibly to resolve ambiguity and/or avoid threat. In support of this, IU has been associated with attentional processes, suggesting a more active alerting system, to help dealing with ambiguity and uncertainty (Fergus & Carleton, 2016). Attention, in turn, has been shown to enhance evidence accumulation and increase the drift rate (Nunez et al., 2017). Interestingly, a recent study using diffusion modelling to examine the influence of social anxiety disorder on decision-making in a probabilistic reward task with social feedback found faster accumulation of information for individuals with higher IU, as reflected by the drift rate parameter (Dillon et al., 2022).

Another possible explanation could be related to the known advantage in the recognition of happy expressions (Nummenmaa and Calvo, 2015), which are believed to have particularly distinctive features, making them less ambiguous as compared to other facial expressions, such as anger and fear (Du and Martinez, 2011; Johnston et al, 2003; Guo et al, 2019). It is possible that individual differences in the perception of emotional expressions as a function of IU may reflect more the ambiguity of the expressions and less their inherent threatening quality. This could mean that individuals with high intolerance of uncertainty are particularly good at processing the less ambiguous positive expressions. Interestingly, we found that IU mediated positive associations between the interoceptive dimensions of Trust and Not Worrying on accuracy and drift rate for happy expressions. This could mean that the advantage conferred by these interoceptive styles on emotion recognition (of happy expressions) could be partially dependent on how the individuals deal with uncertainty. However, we note that these interpretations for the associations between IU and emotion recognition are mostly speculative and future follow up studies are needed to shed further light on this.

Notably, our predictors were not significantly associated with any other DDM. We could expect, for example, that individuals with high IU would show more cautious behaviour and need to accumulate higher amount of evidence to make a perceptual decision, which would be reflected in larger parameter values of boundary separation (*a)*. Previous research has indicated more risk averse behaviour (e.g. higher *avoidance)* in individuals with increased IU in several tasks involving decision-making under uncertainty (Carleton et al., 2016; Ladouceur et al., 1997; Thibodeau et al., 2013). However, several other studies failed to demonstrate this link (Carleton et al., 2016; Jacoby et al., 2014), suggesting that this relationship may depend on the task being measured and affective context (Anderson et al., 2019).

Limitations – Future steps

There are several limitations in this study. Firstly, the sample size is primarily consisted of undergraduate female students, which limits the generalizability of the findings. The sample sizes for our path analyses can be considered adequate, ranging from 124 to 190 participants. The recommended number is 10-20 participants per factor included (Deng et al., 2018), while our a priori power analysis indicated a sample of 103 for moderate effect sizes. Our effect sizes varied between moderate to large effect sizes, which suggests that we had adequate power to detect this effect, even though our estimation did not include mediation effects which typically are weaker and required slightly larger samples. Regarding the number of trials in the ER task, our number of 40 trial per condition for the drift rate and 80 for the other 3 parameters is considered adequate for the DDM, but on the lower end for the drift rate (Lerche et al., 2016). Therefore, the statistical power could improve with more trials. The MLE method used for model fitting is considered the most appropriate for low trial numbers and, indeed, provided adequate convergence for all subjects in the DDM (Lerche et al., 2016). The study could also be improved by testing more models for more specific hypotheses regarding the interactions between our factors to indirectly predict our DVs, which was avoided here to maintain conciseness. While we found significant effects of IU in the emotion recognition task, we could improve the power and scope of this effects by implementing a different design that induces stronger affective engagement, such as creating a social environment where decisions have emotional implications. Such modifications could increase the importance of risk and uncertainty calculations associated with IU characteristics. In addition, a DDM on a wider range of specific positive and negative emotions could elucidate emotion processing differences related to IU. Lastly, future studies should also include a non-social perceptual decision-making task to demonstrate the specificity of IU effects on social perception.

Conclusion

This study demonstrated for the first time different ways in which IU and interoceptive sensibility are implicated in anxiety and social cognition. Specifically, we showed that IU, along with alexithymia, mediates the effects of specific facets of interoceptive sensibility on anxiety and social skills traits, and particularly those related to experienced personal distress. We also demonstrated that IU is associated with various aspects of social cognition, from personal distress to emotion recognition. Notably, we applied for the first time diffusion modelling to reveal that the better performance associated with higher IU for emotional expressions of positive valence is related to more efficient processing of social information. By revealing how interoceptive sensibility alexithymia and IU interact in this context, these findings highlight the potential for targeted interventions to improve social functioning and emotional well-being in individuals with high IU, alexithymia, empathetic anxiety. Future research involving more representative and larger samples, along with different ER task designs, would further elucidate how IU influences social perception and decision-making under uncertainty.

Chapter 3. Emotion Recognition and Interpersonal Emotion Expectations

3.1. Introduction

Understanding others' emotions is fundamental for effective social interactions and communication. However, emotion recognition is a complex process as we do not have direct access to others' emotional states and thus must infer them from available cues, such as facial or bodily expressions, speech prosody, contextual factors, or even one's own emotional and bodily states and previous experiences. Interestingly, it has been shown that our own affective experiences may bias these inferences, particularly when the interpersonal emotional states are incongruent. This phenomenon is described as Emotion Egocentricity Bias (EEB) (Steinbeis & Singer, 2014; von Mohr et al., 2019) and has been linked to mood-congruency effects (Forgas, 2017; Trilla et al., 2021). Research on EEB has used various methods of emotion induction and assessment of emotion perception to study its' properties and associated neural correlates (Riva et al., 2016; Steinbeis & Singer, 2014; Trilla et al., 2021), but the bodily and socioemotional contexts of such biases have not yet been identified.

One prominent mechanism proposed to explain how we understand others' emotional states is embodied simulation, where observing another's emotions automatically activates matching neural, emotional, and somatic states in the observer (Gallese et al., 2008; Singer & Lamm, 2009). This theory suggests that we, innately, use our own emotional experiences as a reference framework for interpreting others' emotions. However, alternative theories argue that affective and motoric mirroring arise not from innate or dedicated simulation mechanisms but through associative learning processes (Heyes, 2018), like Hebbian learning. According to the Learned Matching Hypothesis (LMH), repeated exposure to similar social situations links the observed exteroceptive cues (e.g., facial expressions) to specific bodily and emotional responses in the observer. Consequently, these learned associations allow observers to activate congruent affective states upon perceiving others' emotional cues. Interestingly, the reverse effect can also take place, an experienced affective state could be projected onto the other as the most probable hypothesis about the observed affective state. Since this bidirectional link between the emotional state of the observer and the ‘other’ is learned, it can also be sensitive to contextual factors and the dynamics of interpersonal emotion contingencies (Barsalou, 2013; Heyes, 2018; Kilner et al. 2007). Thus, EEB can be influenced by probabilistically shaped expectations of affective congruency between the observer and the other. For example, when feeling happy, one might expect others to display happiness as well, promoting EEB by perceiving ambiguous or neutral expressions as happy. However, what happens when this expectancy is not fixed or is unlikely based on our past experience? To the best of our knowledge, no study has so far investigated whether learning of interpersonal emotional contingencies (IEC) affects EEB.

Apart from the LMH, the predictive processing (PP) framework provides a complementary perspective on how we perceive and interpret social and emotional information (Ondobak et al., 2017; Seth & Friston, 2016). Unlike traditional views of perception as a passive reception of sensory inputs, PP posits that perception is an active, probabilistic inference process driven by prior expectations (Friston & Kiebel, 2009). According to this view, higher-order representational levels generate predictions about incoming sensory data, which are then updated based on prediction errors—discrepancies between expected and observed outcomes. These updates depend on the precision (or reliability) of the information at each representational level (Mathys et al., 2011). Thus, in this framework, learning is an integral function of cognition and behaviour. PP principles have been recently applied to model social learning by quantifying different sources of uncertainty in dynamic environments

(Diaconescu et al., 2014; Sevgi et al., 2020). A Bayesian learning model that captures associative learning between cues and outcomes in a subject-specific way is the Hierarchical Gaussian Filter (HGF) which accommodates three different levels of uncertainty (Mathy et al., 2014). The first level quantifies irreducible uncertainty, which is related to the inherent perceptual ambiguity of sensory input, and in the context of EEB reflects uncertainty regarding the perceived emotional state of the other in relation to the experienced emotional state. The second level, namely estimation uncertainty, describes the uncertainty relative to the probability of the outcome given a cue or a particular context. Regarding EEB, this level can capture the probability of perceiving an emotion congruent to the experienced emotion. At the third level, volatility estimates quantify the beliefs of how cue-outcome contingencies change over time and can thereby quantify how easily the expectations of interpersonal emotion congruency between the observer and the other can change. By using the HGF, we can examine how individuals learn IEC and how this learning process is linked to socio-emotion processing, body perception traits and physiological arousal.

The processing of internal bodily signals, i.e. interoception, is believed to play a critical role in learning and decision-making, particularly under conditions of uncertainty (Damasio, 1996; Dunn et al., 2010). Interoceptive abilities have been shown to influence key aspects of social cognition, like emotion perception, empathy, and self-other differentiation (Shah et al., 2017; Grynberg & Pollatos, 2015; Pamler & Tsakiris, 2018). In the context of EEB, a recent study by von Mohr and colleagues (2021) attempted to examine how interoception affects EEB by examining the role of stimulus presentation relative to the cardiac phase (systole/diastole). Results demonstrated an effect of ongoing cardiac activity on EEB in participants with high interoceptive accuracy, i.e. individuals good at identifying their heartbeats in a separate task. These findings indicate an important role for physiological activity and interoceptive processing in EEB which is largely underexplored.

A variety of different paradigms have been used to investigate EEB, typically involving the induction of affectively congruent or incongruent states between participants and observed targets, followed by emotion judgements of either the self or others (Steinbeis & Singer, 2014; von Mohr et al., 2019). The emotion manipulation methods usually include monetary games or synchronous visuo-tactile or audio-visual stimulation. However, these paradigms often explicitly require participants to differentiate between their own and others' emotions, as they involve successive ratings of both. This explicit distinction may encourage active disengagement from self-perspectives, potentially confounding the observed EEB effects. In other words, it is difficult to address whether EEB reflects self-other distinction difficulties or affective projection mechanisms. A recent study by Trilla and colleagues (2021) attempted to address this issue by employing a psychophysical task designed to examine how emotion perception is implicitly affected by participants' affective states that were manipulated using autobiographical recall and audio-visual clips to induce transient neutral, happy or sad emotions. The following emotion recognition task involved the categorization of ambiguous facial expressions (of varying ambiguity) as either happy or sad. The results showed that participants were more likely to judge faces as happy following happiness induction. While this approach, by avoiding explicit cues of self-other comparisons allows investigating EEB more implicitly, it still emphasizes participants' emotions during the emotion induction procedure. To further disentangle these effects, novel paradigms are needed that reduce the salience of self-referencing cues while assessing EEB more implicitly.

To address this, we developed a novel dual-task paradigm designed to examine EEB implicitly by minimizing the presence of explicit self-other distinction cues. In each trial, participants first engaged in an emotion induction task using a simplified roulette game, where successful bets are rewarded and unsuccessful ones penalized, to elicit positive and negative emotions, respectively. Immediately after the game's outcome, participants were presented with an ambiguous facial expression, i.e. morphings of sad and happy expressions, and asked to perform binary classifications. Crucially, the two tasks were presented as unrelated, reducing the likelihood that participants would explicitly associate their emotional states with their judgements of others' expressions, thus allowing for a more naturalistic assessment of EEB.

To examine the role of IEC in modulating EEB, across two studies, we created three block types differing on expected congruency (neutral, expected congruency and expected incongruency) by manipulating the probability of emotional congruence between participants' emotional states and observed expressions. Additionally, we examined the impact of perceptual uncertainty on EEB by introducing pixelated noise to the facial stimuli images. These manipulations allowed us to investigate how contextual factors, such as learned expectations and perceptual ambiguity, interact to shape EEB. Lastly, in Study 2 we also monitored participants' heart rate to assess how autonomic responses influence EEB. Building on evidence that interoceptive processing is critical for emotion perception and decision-making, we hypothesized that physiological arousal would modulate the participants' ability to adapt to changing IEC.

Firstly, we expected accuracy and reaction times to be affected by interpersonal expectancies and visual noise, with increasing EEB in the block of expected congruency and in trials with higher visual noise. In addition, we expected learning and response parameters to be affected by physiological reactions, with lower arousal levels facilitating learning and reduced EEB. Finally, we explored how individual differences in self-reported interoceptive sensibility, alexithymia, and empathy modulate perceptual and response parameters of the HGF as well as the degree to which these model parameters are linked to autonomic responses. Specifically, we predicted that participants with scores in these traits reflecting more adaptive socio-emotional abilities to adapt more efficiently to changes in IEC and perceptual noise.

3.2. Methods

3.2.1. **Participants**

Forty-five participants (aged 18-38 years, median age: 19, 33 females) were recruited for Study 1 and 44 (aged 18-34 years, median age: 19, 34 females) for Study 2. All participants were "healthy volunteers" with no history of psychiatric or neurological disorders. They participated in exchange for credits in the research participation scheme of the University of Kent, and participation in a lottery with the possibility of winning £20 worth prize vouchers. The sample size was calculated based on previous experiments using a similar experimental design (Lawson et al., 2017). All participants provided written informed consent before the beginning of the experiments. The studies were approved by the School of Psychology University of Kent Ethics Committee.

3.2.2. **Stimuli**

Stimuli in the emotion perception task consisted of photo morphings of sad and happy expressions at the average subjective perceptual level of 40-60% and 60-40%. The proportion of happy and sad features in each image resulted from subjective judgements of 15 participants that took part in a pilot study. For the pilot study we created morphings comprising sad and happy expressions at various proportions (i.e. 20-80%, 30-70%, 40-60%) of each emotion and asked participants to categorize them, in a 2-alternative forced choice task, as happy or sad. Then, we fitted the responses to psychometric curves to obtain the levels of morphing where each image was 60% (or 40%) likely to be categorised as sad or happy. This procedure allowed us to estimate levels of morphing for each image that, on average, are associated with 60% probability of being identified as a specific emotion, which would not be necessarily the case if the objective levels of morphings were used. The purpose was to create ambiguous images that, while more likely to be associated with a particular emotion, were also prone to be (miss) perceived as another emotion through the influence of own emotional states and learned contextual expectations. For the main task, we selected images with these levels of morphing from 10 different models, 5 females and 5 males, from various ages and ethnicities. These images were then pixelized to create another image version with higher perceptual noise. In total 40 different images (10 models x 2 morphing levels x 2 noise levels) were used in the main task. The original photos were taken from the FACES database (Ebner, Riediger, & Lindenberger, 2010).

The online version of the main task (Study 1) was created on Psychopy, and presented on the Pavlovia platform, while the lab version (Study 2) was created and presented on Maltab (version 2022b, Mathworks, Inc., MA, USA) using the Psychtoolbox (http://psychtoolbox.org/).

3.2.3. Experimental design and procedure

To manipulate IEC, we designed an implicit learning paradigm consisting of two, allegedly unrelated, tasks. In each trial, participants first played a simple betting roulette game followed by the perceptual categorization of ambiguous emotional facial expressions. The first task had the purpose of inducing positive and negative affect after win and lose trials, respectively, to test its influence on the emotion perception task. In the roulette game, participants had to make binary decisions (in some trials this would be black/red in others even/odd number) to bet on where the ball would stop. Then, they were shown whether they had won or lost. This feedback served as the emotion induction manipulation. After an inter-task interval of 3s they were presented with the ambiguous facial expression which they had to categorise, with a key press, as sad or happy as fast and as accurately as possible. The image was briefly presented for 500ms and then they had a time window of 2s to respond. No feedback was given for that response. Critically, to avoid the formation of explicit associations between the roulette's outcome and the emotion categorization task participants were told that the study was examining emotion perception under uncertainty while using an irrelevant, though demanding, secondary task to increase their cognitive load. This allowed us to study the subconscious effect of own emotional states when judging others' emotions.

The probability of winning or losing was fixed at 50% and was totally controlled in trial-by-trial fashion by the experiment's algorithm. However, participants were not aware of this. Instead, to increase emotional engagement with the task, they were told that when they win, they earn 3 points, otherwise they lose 2 points. These points were gathered throughout the task so that the total amount of points was linked to the probability of winning a voucher prize of £20 by collecting higher numbers of raffle tickets. Another addition to make the task more engaging and increase the relevance of each win/loss, was to include a special roulette game that participants played 3 times at specific points during the blocks. The special bet trial was same as the regular ones except they would get 5 times the usual reward (that is 15 points). Participants were told that they would have the opportunity to play the special trial only if they scored better from the average of the previous participants at specific points of the task.

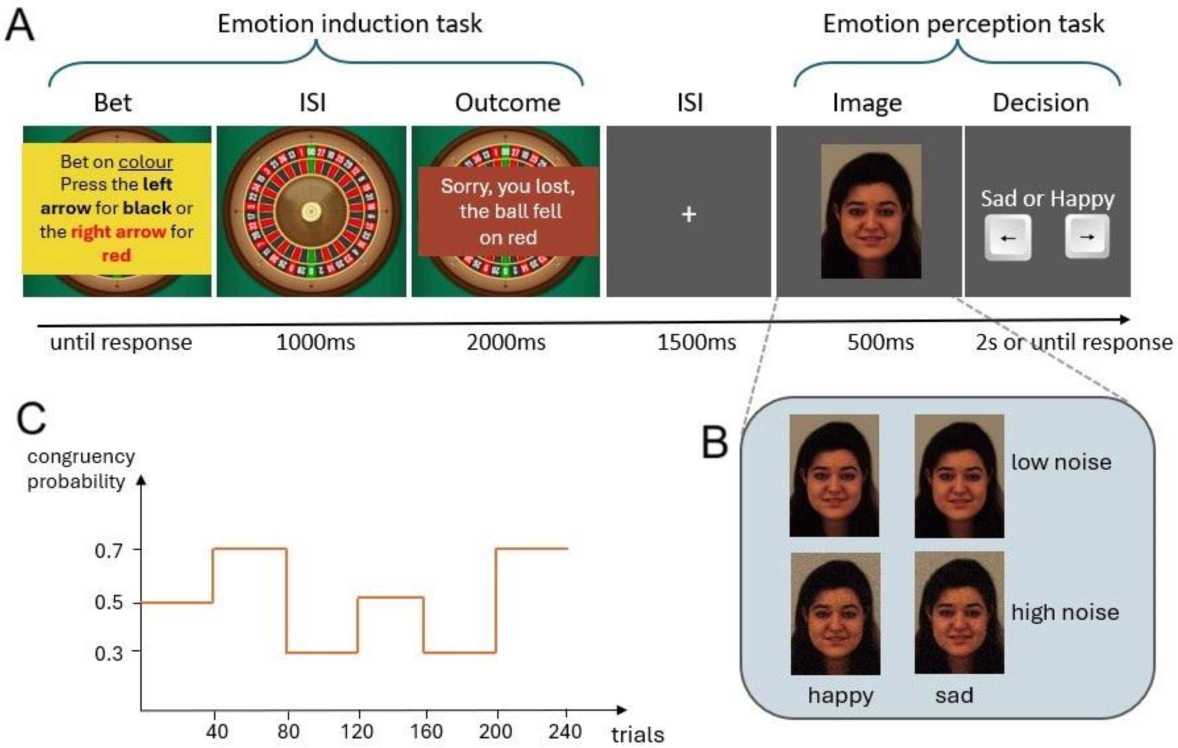


Figure 3.1. Experimental design. A) Example trial of the dual-task paradigm. In the emotion induction phase participants bet either on colour or on number and then see the outcome of the roulette game. Afterwards, they shortly see an ambiguous image and decide whether it was happy or sad. B) Four different conditions of images to be categorised. Two 60-40% morphings, one with happy as dominant emotional expression and one with sad as the dominant expression. C) The time schedule of the conditioned probability to see an emotional expression congruent to the induced emotional state. This manipulation resulted in 6 successive congruency blocks.

To manipulate the expectancy of IEC we created 3 different types of blocks: the expected congruency block (CB), the expected incongruency block (IB) and the neutral block (NB). In the expected congruency block, there was 70% probability to see a happy face after winning in the card game, and 70% probability of seeing a sad face when loosing. Conversely, in the expected incongruency block, there was 30% probability of seeing a sad (happy) face when loosing (winning). In the no expectancy block, there was 50% probability of seeing a sad (happy) face when loosing (winning). The total number of expectancy blocks were 6 with the following order: NB, CB, IB, NB, IB, CB. Each block included 40 trials, so there were 240 trials in total. In half of the trials of each block, the high noise (i.e. pixelized) images were presented in a randomised order, so we had two noise levels. In addition, to avoid classifications to be influenced by stimuli expectations the marginal probabilities of seeing a sad or happy expression were identical and constant across all blocks.

The task started with 4 practice trials. Participants had two breaks to rest for as long as they wanted. Each trial lasted around 7 seconds and the task's overall duration was around 35 minutes. At the end of the task participants were asked to answer 7 debriefing questions to assess what they felt during the task and whether they formed explicit associations between the two tasks, i.e. outcome in the roulette game and the following emotion categorization.

3.2.4. Questionnaires

Prior to the main task, participants were asked to complete online the following questionnaires.

Interoceptive sensibility

To measure interoceptive sensibility we used the Multidimensional Assessment of Interoceptive Awareness, Version 2 (MAIA-2, Mehling et al., 2018). This test includes 37 questions assessing the different ways people process and pay attention to their bodily sensations. See Chapter 2 methods for details on MAIA questionnaire.

Alexithymia

To assess participants emotion processing traits, we used the 20-item Toronto Alexithymia Scale (TAS-20; Bagby et al. 1994). This test measures Alexithymia with questions like "I am often confused about what emotion I am feeling" where participants indicate their level of agreement in a 5-point scale. Although, the TAS-20 has three subscales, we used only the total score here.

Empathy Quotient

To assess empathic traits or more specifically the sensitivity to others emotional experience we used the Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004). EQ was designed as a general, integrated measure of both affective and cognitive facets of empathy. This sixty-item questionnaire includes 40 questions like "I really enjoy caring for other people" and "It is hard for me to see why some things upset people so much" assessing empathic abilities and 20 filler questions, such as "I prefer animals to humans", to distract from constant attention to empathy. Participants indicated their level of agreement to the questions on a 4-point scale, from "strongly agree" to "strongly disagree". The score from all answers is summed to provide the total score. While the TAS-20 and MAIA-2 were administered in both the online and lab study, the EQ was administered only in the latter one.

3.2.5. Physiological measurements

We recorded skin conductance responses (SCR) and cardiac activity using the BIOPAC MP36 (https://www.biopac.com/). For the SCR we placed two electrodes on the middle and index finger of the participants' left hand (data not analysed). For the ECG recording, we used a lead II chest configuration, where two electrodes were placed under the left and right collarbone and one on the left lower back. ECG recordings were made with a sampling rate of 2000Hz and hardware filtering of 1000Hz. Changes in cardiac activity in response to the game's outcome, we calculated the interbeat interval (IBI; values were interpolated between heartbeats to obtain a continuous index of cardiac period) in the time window starting from

0.5s after the outcome's presentation and ending 4s later. As a measure of event-related autonomic arousal, we estimated the peak cardiac acceleration as the lowest IBI within this window, normalised by dividing this value with the baseline IBI (mean IBI 1s before the outcome).

3.2.6. Statistical analysis

Model free analysis

We used a novel associative learning task where participants formed implicit expectations on IEC that were assessed via reaction times (RT) and accuracy as behavioural measurements. To examine how the different conditions affect participants responses we first run a 3x2x2 within-subjects ANOVA on RTs and ER accuracy with Block Type (NB, CB, IB), Trial Congruency (congruent, incongruent) and noise (with, without) as our 3 factors. Reported post-hoc pairwise comparisons were corrected with Bonferroni method.

As we attempted to manipulate expectancies of interpersonal IEC in each block, we further divided our trials into expected (ET) and unexpected (UT) based on the assumed learned ground-truth of each block. Then, we examined RT differences as an index of surprise across the Congruent and Incongruent Blocks for expected and unexpected trials. Specifically, the unexpected trials for the congruent block were the incongruent trials, whereas the unexpected trials for the incongruent block were the congruent trials. We then ran an 2x2 within-subjects ANOVA on RTs for Block (CB, IB) and Expectancy (ET, UT). We expected an interaction effect, with higher surprise levels (i.e. slower RTs) for unexpected trials in the congruent block compared to the unexpected trials in the incongruent block, reflecting a tendency for EEB.

HGF – model-based analysis

We used the HGF to assess the learning processes of IEC (Mathys et al., 2010; 2014). The HGF is a generative model that approximates a Bayesian observer that dynamically estimates the hidden states of its environment, in a framework described as 'observing the observer' (Daunizeau et al., 2010). The HFG models trial-by-trial belief updating at different representational hierarchies. The states of the second and third level evolve in time as Gaussian random walks, with step size (variance) determined from the level above. This approach utilizes subject-specific priors and parameters that capture individual differences in learning, beyond statistical optimality. Inversion of the perceptual and response models is performed by mapping sensory input to observed responses.

More specifically, the perceptual model generates the sensory inputs from the n hierarchical levels: $x_1^{(k),}$ $x_2^{(k)}$,..., $x_n^{(k)}$ (where k is the trial number). Our input was binary, conditionally coded as $u^k = 1$ for congruent trials, that is when the roulette's outcome was matching the emotion of the image to be categorized (win/happy or lose/sad), and $u^k = 0$ for incongruent trials accordingly (win/sad or lose/happy). The first level of the model (x_1) represents the binary belief about whether a congruent or incongruent facial expression to the emotional stated induced by the game's outcome is presented in each trial. Thus, this level represents irreducible, perceptual uncertainty of interpersonal emotional states. The mapping of input to first level beliefs can be deterministic in the absence of perceptual noise or, where perceptual noise is captured as a fixed parameter alpha (α) representing the variance of the noise on the input. The second level (x_2) represents the probability to see an emotionally congruent facial expression to the outcome of the roulette game. The level is linked to the first one with a Bernoulli distribution: $p(x_1|x_2)$ = Bernoulli (x_1; $s(x_2)$), where $s(x)$ is a sigmoid function $s(x) = 1/(1 + \exp(-x))$. The step size of the Gaussian random walk of x_2 is determined by the higher level x_3 according to the following relationship: $x_2^k \sim N(x_2^{k-1}, \exp(\kappa x_3 + \omega_2))$.

The second level in this way represents the so-called estimation (or informational) uncertainty. The third level reflects how fast/easily the contingencies of level 2 change over time, thereby representing the beliefs about volatility uncertainty. The states of this third level evolve with a step size $x_3^k \sim N(x_3^{k-1}, \exp(\omega_3))$, i.e. solely determined by parameter ω_3. The 3 variables κ, ω_2, ω_3 are fixed parameters that capture individual differences in belief updating. The parameter κ reflects the degree to which level 3 affects belief updating of level 2, in this it affects the phasic component of volatility at the second level. Here, to simplify the model, κ was fixed at value 1. Parameter ω_2 is the tonic volatility of level 2 and represents participants' belief on how easily the IEC can change. Lastly, ω3 captures the belief about the social environment's volatility, or 'meta-volatility', how fast changes the way level 2 contingencies change. Individuals with higher ω_3 have higher rate of updating x_3 as they believe in a less stable world.

As mentioned before, belief updating is performed in a trial-by-trial fashion under the assumption that participants apply an approximation to the ideal Bayesian inference. Under the mean-field approximation the resulted update equations have the following form that simply describes that beliefs (posterior means) at each level of the hierarchy (i) are proportional to the prediction error (δ_{i-1}) and to a precision ratio:

$$\Delta\mu_i^{(k)} = \mu_i^{(k)} - \mu_i^{(k-1)} \propto \frac{\widehat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)}$$

The prediction error describes the difference between the lower-level prediction $\hat{\mu}_{i-1}^{(k)}$ before seeing the input and the posterior expectation $\mu_{i-1}^{(k)}$ afterwards. This term is weighted by a ratio of the precision of the prediction of the lower level $\pi_{i-1}^{(k)}$ before seeing the input divided by the precision of the current belief $\pi_{i-1}^{(k)}$, where precision is defined as the inverse variance of the posterior expectation: $\pi_i^{(k)}=1/\sigma_i^{(k)}$. This precision ratio can be considered as the adaptive learning rate at each level. At the second level the learning rate takes the following form: $\mu_2^{(k)}= \mu_2^{(k-1)} – \sigma_2^{(k)}\delta_2^{(k)}$, with $\sigma_2^{(k)}=1/(1/ \hat{\sigma}_2^{(k)} + \hat{\sigma}_1^{(k)}$.

Our design does not include a manipulation of the volatility of IEC, though beliefs about how IEC change over time could be volatile as well. Here, apart from a 3-level HGF with volatility representation level, we also included a version (HGF2) without the third level to examine if a less complex model (without volatility estimates) could better account for our data. For that purpose, for the HGF2, we fixed the parameter κ to 0 (prior κ=0, variance κ=0).

We also tested the widely used reinforcement learning model of Rescorla Wagner (RW) where, similarly to HGF, PEs drive belief updating but with a fixed learning rate (Rescorla and Wagner, 1972). We can thereby examine how a simpler model with steady learning rate or a more complex one with adaptive learning rate fit our data. Details on all priors used in the HGF, response, and RW models are provided in the Appendix (B.3).

Response models

Unit-square sigmoid response model

For the response model we used the unit-square sigmoid function for binary responses (Iglesias et al, 2013). This function maps beliefs to observed responses in the following way:

$$f(y = 1) = \frac{\hat{\mu}_1^{\zeta}}{\mu_1^{\zeta}+(1-\hat{\mu}_1\ )^{\zeta}}$$

that is, by transforming the predicted probability μ_1 (seeing a congruent facial expression) into the probability of choosing the congruent p(y(k) = 1) and p(y(k) = 0) or the incongruent response. The parameter ζ captures the exploration tendency and decision noise, where higher values meaning more deterministic choices and lower values reflecting more random choices.

To assess how participants use the available information in their responses we also created two alternative models of the unit-square sigmoid response mode. In this task, participants can decide whether the image they see is happy or sad by using the (ambiguous) visual information of the image or/and by the shaped expectations from the previous trials under the statistical manipulation of IEC of each block. Therefore, to examine how participants use these two sources of information in their decision we used one version of this models that takes as input the predictions of the first level of the perceptual model ($\hat{\mu}_1^{(k)}$) and a second version that takes as input the posterior beliefs of the first level ($\mu_1^{(k)}$).

Response model with weighting factor

To assess how participants use the available information in their responses we also created an alternative model which combines the beliefs related to the priors and posteriors mentioned above. In this way, we quantified how participants weight these two information sources in their decisions by adding a weighting factor β, so that the integrated belief is given by the following equation:

$b^{(k)}=\beta*\hat{\mu}_1^{(k)} + (1-\beta)*\mu_1^{(k)}$, with $\beta \epsilon [0,1]$.

Model inversion

We performed an approximate Bayesian inversion for each model by calculating the maximum-a-posteriori (MAP) estimates of model parameters. This calculation was executed using the HGF toolbox version 2.1 after the prior for all parameters and given the input sequence with the congruency and noise condition for each trial. We used the tapas_hgf_binary_pu function for our main perceptual model from the TAPAS toolbox and the quasi-Newton optimization algorithm as our optimization function (Frässle et al., 2021). The objective function for maximization was the log-joint posterior density over all perceptual and observation parameters, given the data and the generative model.

Model comparison

The formal model comparison was conducted in two steps, where firstly we tested the alternative perceptual models and secondly the different response models. So, we first compared 4 perceptual/learning models, i.e. the HGF2, HG3, RW and a non-learning model (RW0). The RWO had a fixed learning rate at 0, treating all input as noise, thereby simulating a model without learning behaviour. This non-learning model served as a baseline to assess whether participants could learn the IEC. After identifying the winning perceptual model, we combined it with three response models from the unit sigmoid function family where we parametrise the input of the perceptual model they get: (1) unit_sgm with posteriors, (2) unit_sgm with predictions, and (3) unit_sgm with a weighting factor. This resulted in a total of seven models tested in two steps. Bayesian model selection (BMS) was conducted using the spm_bms function of the SPM toolbox to perform random-effects BMS. For each model, we extracted expected posterior probabilities and exceedance probabilities based on participants' log model evidence (LME) to assess which models best accounts for the observed data. The expected posterior probability is a measure of the probability of each model at the individual level, averaged across subjects. Exceedance probability provides an estimate of which model is most likely to be the best overall, compared to the alternative models tested.

Regression analysis for individual differences

With the estimates of the winning model that we extracted after model fitting and BMS, we run separate linear regressions. That is, the estimates of the perceptual and response models of the HGF were entered as DVs and the dimensions of MAIA, TAS20 and EQ (study 2 only) as predictors. Specifically, to examine learning effects we used as DVs the learning rate parameters ω2 and ω3 of the second and third level, respectively. In addition, we also examined how the learning rate change at the second level (ΔLR2) between expected congruency blocks (Congruent vs Incongruent block) is related to questionnaires' scores by calculating the average learning rate in these blocks and then taking their difference as our DV. Lastly, we also run a similar regression on the response model parameter ζ.

Physiological data analysis

We also examined whether trial-wise physiological arousal (Study 2), indexed by evoked heart rate acceleration (HRA), was associated with learning of interpersonal emotional contingencies (IEC) and how this relationship varied with individual differences. HRA was calculated as the inverse of inter-beat intervals (IBIs), such that higher values reflect increased heart rate acceleration. Mixed-effects models were used to test whether heart rate changes in response to roulette outcomes tracked trial-by-trial fluctuations in learning rates and belief updating. HRA served as the dependent variable, with participant ID included as a random intercept. In one model, trial-wise belief trajectories from the second and third HGF levels (i.e., μ_2 and μ_3), representing the estimated probability of congruent outcomes and its volatility, were entered as fixed effects. In a second model, trial-wise learning rates at the second and third levels (LR_2 and LR_3), which quantify the amount of belief updating in response to prediction errors at each level, were used as fixed predictors to examine whether arousal dynamically tracked precision-weighted learning.

Additionally, to further explore how individual differences contribute to behaviour in this task, we ran participant-wise linear regressions predicting HRA from the two learning rates. The resulting regression coefficients for each participant were then used as dependent variables in subsequent linear regressions, with MAIA subscales, TAS-20, and (in Study 2) EQ scores as predictors. This allowed us to assess whether trait measures predicted the degree to which physiological arousal was coupled to learning dynamics.

3.3. Results

3.3.1. Model-free analysis

Study 1 - Accuracy

We run the same analyses for the online and lab study on the ER accuracy and RT data. The 3x2x2 (Block x Trial x Noise) within-subjects ANOVA on accuracy for the online study revealed a significant main effect of Block (F(2,84)=6,291, p=0.003, η^2_p=0.124) with participants being less accurate in the incongruent block compared to the neutral one as shown in post-hoc pairwise comparisons (t(1,42)=4.010, p<0.001),. No differences were observed in the other block comparisons (ps>0.05). We found a significant effect of Trial (F(1,42)=9.147, p=0.004, η^2_p=0.156) with incongruent trials producing more errors than the congruent ones. The main effect of Noise was also significant (F(1,42)=15.395, p<0.001, η^2_p=0.272) with participants, surprisingly, being more accurate in the high noise trials.

Our 3-way interaction between Block x Trial x Noise interaction was found to be significant (F(2,84)=11.106, p<0.001, η^2_p=0.219) as well as the Block and Trial interaction (F(1,42)=29.706, p<0.001, η^2_p=0.424). Pairwise comparisons for congruent vs incongruent trials within each block showed higher accuracy for congruent trials (t(85)=3.930, p=0.048, Cohen’s d= 0.343) in the neutral block, revealing of a small baseline EEB behaviour, with small to medium effect size. That is, in the absence of contextual constrains (manipulation of expected congruency) participants tended to perceive others’ emotions as congruent to their own. Interestingly, this difference was particularly pronounced in the expected congruency block (t(85)=8.120, p<0.001, Cohen’s d= 0.868) with a large effect size, mainly due to a considerable drop in the accuracy of the incongruent trials (Figure 1a). Conversely, in the expected incongruency block accuracy was higher for the incongruent trials compared to the congruent ones, with an accuracy difference close to significance (t(85)=-1.742, p=0.054, Cohen’s d= 0.246). The higher tendency to provide egocentrically biased responses in the expected congruency block (large effect size vs small effect size in the Neutral block) and a tendency towards the opposite behaviour in the expected incongruency block reveal contextual adaptation, i.e. implicit learning of contextual expectancies.

The Block x Noise interaction (F(2,84)=14.767, p<0.001, η^2_p=0.259) was also significant (Fig.1b) as accuracy increased with higher noise only in the expected congruency block (t(85)=5.718, p<0.001, Cohen’s d=0.767), while no difference in accuracy between noise levels was observed in the incongruent (t(85)=0.973, p=0.333, Cohen’s d=0.117) and neutral block (t(85)=0.380, p=0.704, Cohen’s d=0.051). Lastly, the Trial x Noise interaction did not reach significance (F(1,42)=0.781,p=0.382, η^2_p=0.024).

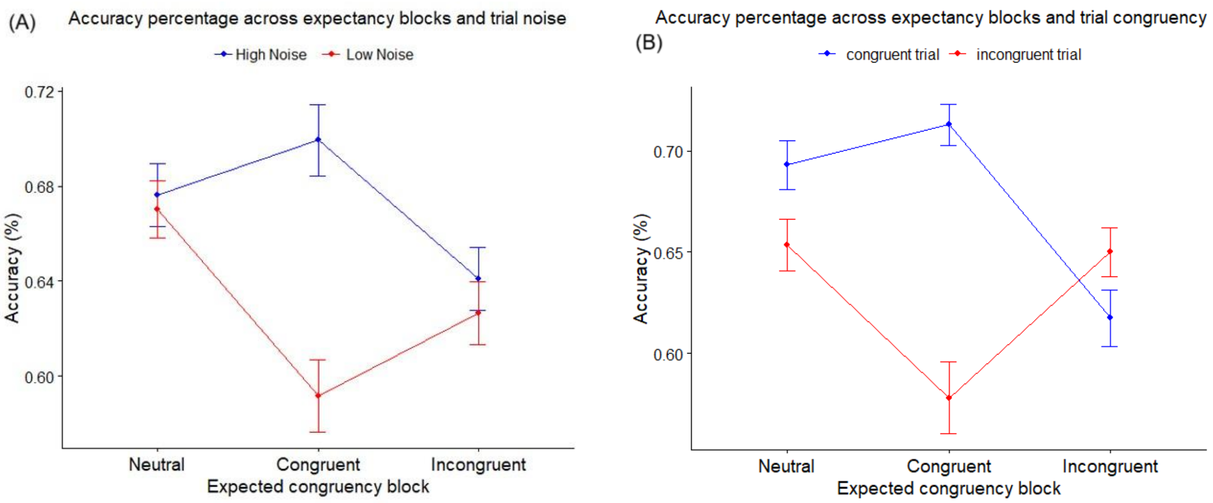


Figure 3.2. Study 1 behavioural results. Mean accuracy scores across the three expected congruency blocks for the two perceptual noise conditions (A) and the two trial congruency conditions (B).

Study 2 - Accuracy

The 3x2x2 (Block x Trial x Noise) within-subjects ANOVA on accuracy for the lab study was significant (F(2,86)=19.672, p<0.001, η^2_p=0.281) with results largely in accord with those of Study 1. There was a significant main effect of Trial (F(1,43)=15.796, p<0.001, η^2_p=0.245) with incongruent trials producing more errors than the congruent ones. The main effect of Noise (F(1,43)=36.049, p<0.001, η^2_p=0.405) was also significant, with higher ER in the high noise condition than the low noise one. The main effect of Block approached significance (F(2,86)=2.879, p=0.060).

As in study 1, the Trial x Block interaction (F(2,86)=21.790, p<0.001, η^2_p=0.321) is attributed to the increased difference (i.e. larger effect size) in accuracy between congruent and incongruent trials in the expected congruency block (t(89)=3.151, p<0.001, Cohen's d=0.982) compared to the Neutral block (t(89)=6.589, p=0.002, Cohen's d=0.518), with no difference in the expected incongruency block (t(89)=0.247, p=0.805, Cohen's d=0.039) (Fig. 2A). Similar to study 1 findings, the Block x Noise interaction (F(2,84)=14.767, p<0.001,

η^2_p=0.259) was significant (Fig.2B) as accuracy increased with higher noise only in the expected congruency block (t(89)=6.195, p<0.001, Cohen's d=0.762), while no difference in accuracy between noise levels was observed in the other two blocks (p>0.665).

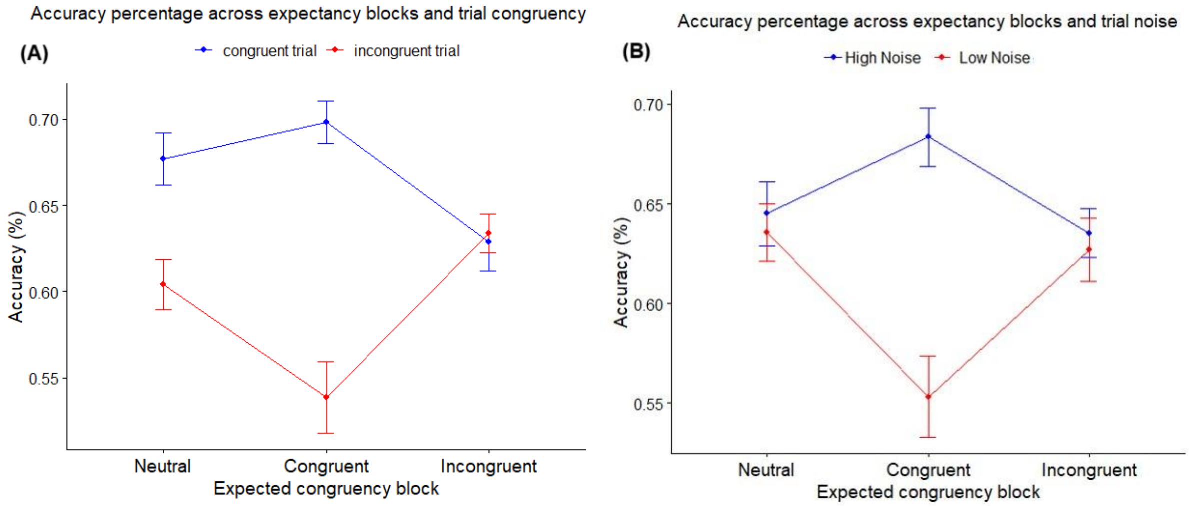


Figure 3.3. Study 2 behavioural results. Mean accuracy scores across the three expected congruency blocks for the two perceptual noise conditions (A) and the two trial congruency conditions (B).

3.3.2. Model-based analysis (HGF)

Bayesian Model Selection

Study 1

For the comparison of perceptual models, we found that the 3-level HGF with 0.77 expected posterior probability and 1.0 exceedance probability outperformed the 2-level HGF and the RW models (Fig 3A). Thus, the winning model with volatility, albeit the most complex among the three, explained better the behavioural data, suggesting that participants took the change in environmental volatility into account to adjust the rate of belief updating of IEC. In addition, since the zero-learning model was the worst among all learning models, with expected posterior probability of 0.03, provides clear evidence that participants learned the IEC.

Then, we combined the winning HGF3 model with three different response models. The results of BMS showed that the model with posteriors as input in the response model outperformed, with 0.61 expected posterior probability and 1.0 exceedance probability, the models with priors or combined input of priors and posteriors, which had 0.12 and exceedance probability respectively 0.27 (Fig. 3B). Even though, some participants appear to use a combination of posteriors and priors, model selection showed that the model with only the posterior beliefs better explained learning behaviour with a simpler response model compared to the weighting factor response model.

Study 2

Similarly to study 1 findings, the HG3 with expected posterior probability of 0.64 and exceedance probability of 0.96 outperformed the HGF2, RW and zero-learning model with expected posterior probability of 0.15 and 0.19 respectively (fig. 3C). Consistent with study 1, Bayesian model comparison of the response models showed that the HGF3 combined with response model with posteriors only, which had expected posterior probability of 0.64 and exceedance probability of 0.98, was better than the models with priors or weighting factor, with expected posterior probability of 0.07 and 0.28, respectively (fig. 3D).

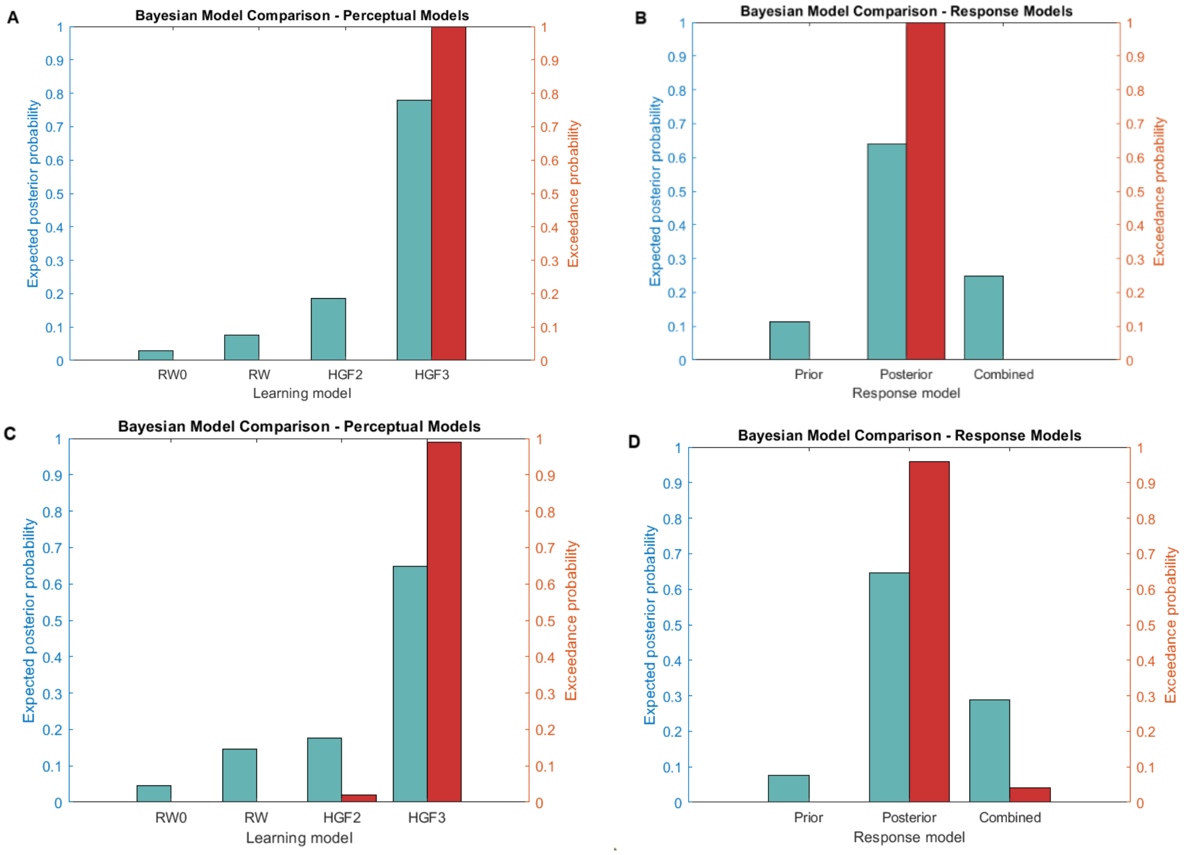


Figure 3.4. Bayesian model comparison for the learning and response models tested in study 1 (A, B) and 2 (C, D). RW0 is the zero-learning model, RW is a Rescorla-Wagner model with stable learning rate, HGF-2 does not include volatility representation, HGF3 is the typical HGF model with 3 representation levels of uncertainty. All perceptual models were combined with the default unit sigmoid model with prior beliefs as inputs. Then for the response model comparison, the three response models were combined with the winning HGF3 perceptual/learning model.

3.3.3. Regression analysis for individual differences

Study 1

To investigate how individual differences in EBB learning relate to individual traits we run a series of linear regressions with model parameters as DV and questionnaire scores (MAIA subscales, TAS20, EQ) as predictors. The LM on the learning rate of IEC (ω2) showed that Noticing (β=-1.40, t=-2.450, p=0.024), and Emotional-Awareness (β=1.95, t=3.570, p<0.015) were significant predictors. These findings suggest that people with reduced overt attention to their bodily sensations and those with increased awareness of the relationship between bodily and emotional states track better the IEC. The LM on meta-volatility of interpersonal emotion contingencies (ω3) showed that Noticing (β=-0.66, t=-2.585, p=0.014), Attention Regulation (β=0.64, t=2.088, p=0.044), and Emotional-Awareness (β=1.00, t=4.524, p<0.001) were significant predictors, which similarly indicates that varying beliefs in the volatility of IEC are associated with different interoceptive awareness traits.

The LM on the learning rate difference between the congruent and incongruent blocks (ΔLR2) revealed Emotional-Awareness (β=0.02, t=2.123, p=0.003) as significant predictor of learning rate change which indicates faster learning in the congruent block compared to the incongruent one for those with higher emotional awareness. The LM on the response model parameter ζ reveals Emotional-Awareness (β=0.096, t=1.029, p=0.037) as the only significant predictor, indicating that those with more awareness of the role of the bodily sensations in their emotional experience tend to use more consistently their trial-wise updated beliefs about IEC during emotion categorization.

Study 2

For the lab study data, we run the same regressions with the addition of EQ to our predictors. The LM on tonic volatility of IEC (ω2) showed that only Noticing (β=-1.60, t=-1.869, p=0.042) was a significant predictor. In contrast to study 1 findings the LM on tonic volatility at the third level (ω3) showed only Attention Regulation (β=0.94, t=1.377, p=0.035) as significant predictor.

In contrast to study 1 findings, the second level learning rate difference ΔLR2 was not predicted by any of the three questionnaires scores. Similarly, no significant predictors were found for the regression on the response model parameter ζ (ps>0.203).

	ω2	**ω3**	**ΔLR2**	**ζ**
Study 1	**Noticing** (β=-1.40, p=0.024), **Emotional Awareness** (β=1.95, p=0.015	**Noticing** (β=-0.66, p=0.014), **Attention Regulation** (β=0.64, p=0.044), **Emotional Awareness** (β=1.00, p<0.001)	**Emotional Awareness** (β=0.02, p=0.003)	**Emotional Awareness** (0.096, p=0.037)
Study 2	**Noticing** (β=-1.60, p=0.042)	**Attention Regulation** (β=0.94, p=0.035)	None	None

Table 3.1. Regression results for individual differences on HGF model parameters. The table presents the significant predictors of the questionnaires including all IS dimensions (MAIA) in the five regressions on model states and parameters (second-level learning (ω2), meta-volatility learning rate (ω3), learning rate difference between the congruent and incongruent block (ΔLR2), and decision parameter (ζ)) across the two studies. Alexithymia and trait Empathy did not reach significance in any regression.

3.3.4. HGF and physiology

To examine whether the changes in heart rate in response to the outcome of the roulette game associate with trial-wise fluctuations in the learning rates, we conducted mixed-model analysis with the evoked HRA (heart rate acceleration) as DV and the level 2 and 3 learning rates as fixed-factor effects. The results show level-2 learning rate as a significant predictor (β=-0.014, t=-2.330, p=0.020), indicating increased learning rate of IEC was associated with lower physiological arousal. Level-3 learning rate did not reach significance (β=-0.042, t=-1.393, p=0.167).

Then, in a similar mixed-model with the same DV we enter the belief trajectories of the second (x_2^k) and third (x_3^k) level of the HGF model as fixed factors. The results revealed that level-3 beliefs was as a significant predictor (β=-0.03, t=-2.088, p=0.037), suggesting that belief of lower volatility was associated with stronger sympathetic responses.

Finally, we explored how individual differences predict the degree to which physiological arousal was associated with learning of IEC in this task. The regression on LR2 coefficients revealed Body Listening (β=0.769, t=2.425, p=0.022) as significant predictor. Trust was also close to significance (β=0.435, t=2.026, p=0.052) while all other predictors were not (ps>0.170). Similarly, the regression on LR3 coefficients revealed Body Listening (β=13.561, t=2.588, p=0.015) as significant predictor, whereas the rest of the scores did not predict the relationship between psychophysiological activity and volatility learning rate (ps>0.200). Together, these results suggest that individuals who, as a trait, tend to listen to, and maybe trust, their bodies more showed a closer relationship between physiological arousal and learning of IEC.

3.4. Discussion

Past research has consistently shown that judging others' emotions is influenced by our own emotional and bodily states, a phenomenon often named Emotional Egocentricity Bias (EEB; Folz et al., 2022; Trilla et al., 2021; Van Mohr et al., 2021). Recent evidence suggests that such egocentric biases are shaped by the precision of self-related predictions (Sevi et al., 2022) and can be modulated by targeted stimulation of cortical regions involved in perspective-taking (Weigand et al., 2021). Yet, the relative roles of contextual, affective and perceptual factors underlying EEB remain unclear. Across two studies, we used a roulette-based, implicit emotion-induction task paired with the classification of ambiguous facial expressions to demonstrate that EEB is jointly shaped by learned IEC and sensory uncertainty. To our knowledge, this is the first evidence that EEB is modulated by implicit learning about socio-emotional context in interaction with perceptual ambiguity. Moreover, individual learning trajectories were estimated using hierarchical Bayesian models to reveal that interoceptive sensibility and physiological responses to the emotion induction contributed to the interindividual variability in adaptive socio-emotional learning.

In our task, the interaction between block type and trial congruency revealed several notable behavioural patterns related to EEB. In the neutral block, performance was better for congruent than incongruent trails, indicating a baseline EEB likely rooted in lifelong social priors. In the expected congruency block, this effect was amplified due to significantly poorer accuracy in incongruent trials. This demonstrates that manipulating congruency expectations increased participants' tendency to categorise ambiguous expressions as matching their own induced emotional state. Conversely, when incongruency was expected (incongruent block), accuracy declined for congruent trials and improved for incongruent trials, reflecting reduced EEB and implicit learning of IEC. These results indicate that expected contextual congruency influenced emotion recognition and EEB, while also suggesting successful implicit IEC learning. Bayesian model comparison further confirmed participants' IEC learning, as the zero-learning model was outperformed by all models incorporating learning parameters. This also indirectly supports the effectiveness of the roulette game as an emotion induction method. Although we lacked trial-by-trial affective assessment, post-task engagement ratings suggested at least moderate emotional impact (Appendix B.1). Previous EEB studies typically required explicit self–other distinctions (Trilla et al., 2020; Sevi et al., 2020; von Mohr et al., 2019), leaving unclear whether observed biases reflected perceptual projection or explicit contextual influences. A key strength of our study is that we attempted to make the self and other distinction as implicit as possible, minimizing the salience of self-referencing cues. This approach likely fostered a more natural expression of EEB, driven by automatic top-down projection mechanisms rather than strategic processes.

When comparing response models, we found that in both studies participants primarily relied on posterior beliefs (the congruency belief updated after perceiving the facial expression) rather than priors alone. While this may seem to suggest that prior expectations of IEC had little influence on decision-making, it is important to note that posteriors inherently integrate both prior beliefs and sensory input through Bayesian updating. Thus, the superior fit of the posterior-based model underscores that participants' emotion categorisation relied predominantly on updated, context-sensitive expectations rather than on prior expectations or sensory input alone.

Interestingly, adding visual noise improved performance, as faces with higher pixel noise were classified more accurately and quickly (Appendix B.2).   Although this contrasts with previous studies (e.g., Lawson et al., 2017), it might reflect unique characteristics of the morphings or noise pattern used, potentially highlighting features (e.g., spatial frequencies) that facilitated emotion recognition (Kumar and Srinivasan, 2011; Vuilleumier et al, 2003). Further research is needed, however, to understand the reason for this pattern.

Critically, noise interacted with contextual and trial congruency as accuracy was substantially lower for low-noise incongruent trials only in the congruent block, suggesting that strong contextual expectations may increase EEB in conditions of higher visual quality. This may seem counterintuitive as it could be reasonable to expect perceptual ambiguity to amplify the reliance on contextual expectations. Instead, it is possible that, in conditions of strong contextual priors such as in the congruent block, reduced weighting is given to sensory evidence due to diminished need to sample the environment to explain sensory information.

Under sensory degradation though, additional weight might be allocated to sensory information to resolve arising prediction errors, thus narrowing the precision gap and attenuating the bias. This effect was absent in neutral and incongruency blocks, implying that noise-driven perceptual biases may not arise when congruency priors are weak or reversed.

Cognitive, perceptual and contextual constrains are considered critical in EEB and social cognition more generally. For instance, Steinbeis and Singer (2014) showed that EEB can be enhanced by time constrains, while proponents of the interaction theory of social cognition argue that simulation of others' mental states could be more useful in the absence of clear or meaningful interpersonal interaction (Gallagher & Varga, 2014). Our findings support such perspectives, highlighting the importance of different forms of interpersonal constrains and uncertainty.

These findings further align with PP, which formally describes how different sources of information are integrated based on their respective reliability (inverse uncertainty) during perceptual inference, capturing the interplay between top-down and bottom-up information streams (Friston & Kiebel, 2009). Cognitive penetrability and top-down perceptual influences have long been reported in various perceptual tasks and explained within the context of PP (Hohwy, 2017). Our findings contribute to this body of research by demonstrating that short-term social priors, formed through the manipulation of IEC, significantly influence emotion recognition. This suggests that EEB, traditionally considered a largely fixed egocentric projection, may be better understood as dynamic and context-sensitive, emerging from interactions between probabilistically learned interpersonal priors and moment-to-moment sensory reliability.

Our results also resonate with the Learned Matching Hypothesis (Heyes, 2018), which proposes that empathic and mirroring responses are shaped by associative learning of sensorimotor contingencies. Previous studies have demonstrated flexible mirroring effects modulated by expectations, sometimes reversing neural activation patterns (Catmur et al., 2011). In the context of EEB, our findings suggest that such learned associations can be dynamically adjusted to reflect interpersonal contingencies and minimize prediction errors, thus also aligning with PP theories of adaptive social behaviour (Ondobaka et al., 2017). Incorporating the Learned Matching Hypothesis into a PP framework provides a unified understanding of how contextual expectations influence emotion recognition. Our hierarchical Bayesian modelling further indicates that participants adjusted learning rates based on perceived volatility of IEC, suggesting that adaptive emotion recognition relies on estimating environmental uncertainty - consistent with previous PP studies on social inference (Diaconescu et al., 2014). Extending PP models to interpersonal emotion perception underscores their broad applicability within social cognition.

Another key aspect of our study is clarifying how interoceptive signals shape social-learning dynamics, with interoception widely implicated in social cognition (Shah et al., 2017; Grynberg & Pollatos, 2015), learning (Werner & Schandry, 2024) and adaptive decision-making (Dunn et al., 2010). Trial-by-trial analyses showed that larger heart-rate accelerations (higher HRA) co-occurred with lower estimates of environmental volatility and slower belief updating, suggesting heightened autonomic sensitivity to unexpected outcomes in perceived stable contexts. Since higher surprise in game outcomes entails greater prediction errors, excessive autonomic responses could interfere with the implicit tracking and updating of IEC, resulting in reduced learning rates.   This heightened arousal could reflect elevated autonomic reactivity to emotional context, as in the case of anxiety, which has been linked to reduced flexibility in updating expectations and greater reliance on rigid priors in state and trait anxiety (Browning et al., 2015; Hein et al., 2021; Paulus & Stein, 2010). One alternative interpretation is that elevated HRA may reflect increased motivation or attentional capture following emotionally salient outcomes, rather than heightened surprise or prediction error. However, this would predict enhanced learning, not the reduced learning rates we observed. Importantly, individual differences further moderated this coupling as participants who relied more on bodily signals (Body Listening subscale of MAIA-II) showed stronger coupling between physiological arousal and learning dynamics, supporting recent theoretical perspectives emphasizing that interoceptive integration modulates the calibration of precision in PP models, driving adaptive learning and behaviour (Allen et al., 2020; Bidell et al., 2024).

Other dimensions of interoceptive sensibility were also associated with adaptive learning. Individuals with greater emotional and body awareness, as well as ability to self-regulate by attending to their bodily sensations, exhibited higher sensitivity to changes of IEC. Conversely, the self-reported conscious awareness of bodily sensations (Noticing) had opposite effects in learning. This aligns with previous studies reporting mixed effects of interoceptive sensibility subscales in social cognition (Stoica & Depue, 2017), with findings linking scores on the Noticing subscale to maladaptive behaviour (Vabba et al., 2023). In our study, interoceptive sensibility primarily influenced learning parameters at the second and, to a lesser extent, third level of the HGF. In contrast, alexithymia and empathic traits did not significantly predict individual differences in learning trajectories. Notably, some individual differences were not completely replicated across the two studies, likely due to small effects and limited sample sizes for these regression analyses. Nonetheless, our findings align with previous research linking interoception to social cognition and EEB (Shah et al., 2017; von

Mohr et al., 2020), highlighting the role of bodily sensations in emotion perception and social interactions.

Despite the contributions of our study, several questions remain, especially regarding how the examined processes operate in real-world social situations. For instance, the current study focused on a relatively controlled environment with experimentally induced interpersonal contingencies. Are such biases more robust in interactive naturalistic contexts or in more passive judgements without actual engagement? Future research could explore whether similar biases occur in more naturalistic interactions, where multiple contextual and interpersonal cues compete for attention. Additionally, while our study revealed sensitivity to environmental volatility, we did not explicitly manipulate volatility conditions. Including conditions of high and low volatility in future designs could provide a more nuanced understanding of how uncertainty at different representational levels affects EEB and adaptive learning.

Another area for improvement would be enhancing participants' emotional engagement with the game to induce emotional changes more effectively. This could be achieved through various modifications, such as adding an indication of monetary gains and losses or incorporating a greater variety of monetary games. To better evaluate the affective impact of emotion induction manipulations, participants could provide related responses every few trials, offering direct insights into their emotional states. In addition, future studies could explore more socially embedded or interactive emotion induction methods, such as modulating interpersonal attunement in real-time interactions (Bolis et al., 2023). Brain activity measurements can also be incorporated to monitor underlying emotional processes during emotion induction, perception and decision-making. Another important modification could be differentiating the facial expression stimuli used for shaping IEC and emotion categorization. Using ambiguous facial expressions balanced at a 50/50 mixture of sad and happy emotions could provide greater precision in assessing EEB. Finally, despite our efforts to implicitly manipulate IEC, it is possible that some participants became aware of the link between their emotional states and the categorization task during the experiment. Post-task debriefing suggested none or very limited awareness across participants, but future studies could incorporate additional measures to further reduce explicit associations (Appendix B.1).

In conclusion, this study advances our understanding of EEB by demonstrating how implicitly learned interpersonal emotional expectations and perceptual ambiguity jointly modulate emotion recognition. By linking computational modelling parameters to physiological arousal and interoceptive sensibility, we provide evidence of how both conscious and pre-conscious aspects of interoception drive adaptive socio-emotional learning. These findings underscore the critical role of predictive and interoceptive processes in emotion recognition and social cognition. Future research can build on this paradigm to explore the neural, affective, and social underpinnings of EEB with improved experimental designs.

Chapter 4. Interoceptive processing and adaptive empathy

4.1. **Introduction**

Empathy is multifaceted social process that allows us to understand and share others emotional states, involving neurobiological, affective and cognitive components (Decety and Jackson, 2004). It is a construct that has been intensively studied over the past few decades, but empathic responses or tendencies are typically studied under relatively static conditions, such as, for example, when participants consider isolated events or engage in “one-shot” interactions with others, which do not reflect the complexity and dynamics of most social interactions. Real-world empathic interactions often occur in dynamic contexts where the emotional states of others are uncertain and change over time. This necessitates the empathizer to continuously adapt their responses. Thus, it could require learning what is the most effective empathic strategy for that person or situation and also to overcome or regulate one's own affective states and preferred empathic strategies (Shamay-Tsoory & Hertz, 2021). Based on this dynamic view of empathy, Kozakevich Ardel and colleagues (2021) developed a task examining adaptive empathy as a learning process. Specifically, their study involved a reinforcement learning task where participants learn over multiple exposures to distress scenarios the preferred emotion regulation strategy (reappraisal or distraction) of two virtual characters. It was found that, through the emotional feedback given by the virtual characters, participants could associate their actions with outcomes and understand which emotion regulation strategy had higher probability to relieve the distress of each character.

Thus, understanding our own and others' emotions is fundamental for effective empathic behaviour. Interoception, the processing and awareness of internal bodily sensations, has been linked to several dimensions of emotion processing and social cognition (Arnold et al., 2019; Gao et al 2019). For instance, stronger interoceptive abilities are associated with better emotion recognition and mindreading of others' intentions (Fukusima et al., 2011; Grynberg & Pollatos, 2015; Shah et al., 2017), and self-other distinction (Engelen et al., 2023; Palmer & Tsakiris, 2018), skills essential for navigating empathic interactions. Additionally, interoception has also been linked to better learning (Pfeizer et al., 2017), intuitive decision-making under risk and uncertainty (Dunn et al., 2010; Kandasamy et al., 2016; Werner et al., 2009), and hippocampal function (Stevenson et al., 2018), suggesting that being attuned to one's bodily signals can enhance learning in dynamic environments. Thus, interoception may also play a pivotal role in empathic learning and decision-making under uncertainty, where perceiving fluctuations of internal bodily states can help detect subtle changes in the social environment and guide adaptive responses.

Interoception's relationship with learning has been conceptualized within Prediction Processing (PP) frameworks (Barrett & Simmons, 2015; Seth, 2013). In this view, the brain does not merely passively process bodily and sensory signals but actively shapes perception based on prior expectations. The brain's primary function, according to PP frameworks, is to minimize prediction errors (PEs), i.e., discrepancies between expected and observed sensory input, by continually updating its expectations through learning. However, not all PEs errors are treated equally; their impact depends on the precision, or reliability, assigned to the sensory data (e.g. interoceptive and social cues) relative to prior expectations, generating precision weighted PEs (pwPEs). In this framework of probabilistic inference, precision and uncertainty are directly linked, and both refer to properties of probability distributions that represent our beliefs about the possible causes of our sensations, with uncertainty quantified as the variance of probability distributions and precision as the inverse quantity. Thus, uncertainty estimations are critical; when uncertainty is high – such as in volatile social interactions - the brain may place more weight on new sensory information (including interoceptive signals) to update its beliefs (priors). In this case, higher precision of PE (i.e. high pwPE) drive learning more rapidly. Conversely, in stable environments, more precision is assigned to expectations and learning is slower (with low pwPE). Another way to frame this is in terms of expected and unexpected uncertainty (Soltani & Izquierdo, 2019; Yu & Dayan, 2005). The low-volatility phase primarily captures "expected uncertainty", variability participants can anticipate and refine their predictions around. The high-volatility phase, in contrast, represents “unexpected uncertainty,” where sudden changes necessitate more radical revisions of one’s internal model.

The fundamental computational mechanism to optimize precision or salience between and within modalities (between priors and PE in the latter case) is attention (Feldman & Friston, 2010; Yu & Dayan, 2005). When the brain detects uncertainty in sensory information, it increases the precision of sensory signals deemed important for resolving that uncertainty. By allocating more attentional resources to these signals, the brain boosts their impact on the PE minimization process, allowing for more accurate updates of internal models. In addition to this bottom-up driven process, precision can also be adjusted by top-down endogenous attention, amplifying the processing of selected signals. Importantly, the precision of priors (the confidence regarding the knowledge of the internal and external world) increases over time through learning as predictions become more accurate.

The optimization of precision across interoceptive and exteroceptive modalities is particularly critical when navigating dynamic social interactions as it facilitates adaptive learning, attention and decision-making (Ondobaka et al., 2017). In social contexts, environmental uncertainty informs higher-order multi-modal predictions (including precision expectations), shaping how unexpected changes in others' behaviour are interpreted, i.e., as meaningful shifts of preferences or noise. For example, during empathic interactions, learning may be driven by the target’s emotional feedback triggering affective responses and related interoceptive signals (i.e., PEs) in the empathiser. Prior expectations and uncertainty estimates (including the distinction between expected and unexpected uncertainty), in turn, can modulate the precision and salience assigned to these bodily responses, influencing how we adjust our reactions in a dynamic social landscape. While multi-modal PP has been linked to various affective processes (Seth, 2013; Tsakiris & Critchley, 2016), an account of interoceptive integration in social learning must consider not only how attention fine-tunes precision under expected fluctuations, but also how abrupt changes in social contingencies demand substantial model revisions, thereby identifying how individuals balance reliance on prior expectations with responsiveness to new cues requiring deeper learning.

Hierarchical Bayesian models have been developed under the PP framework to quantify individual differences in learning and decision-making under uncertainty (Becker et al., 2016; Lawson et al.,2017; Sevgi et al., 2020). One such formulation, the Hierarchical Gaussian Filter (HGF), represents uncertainty across three levels (Mathys et al., 2011; 2014). At the first level, irreducible uncertainty refers to the baseline uncertainty inherent in any situation. In an empathic social interaction, this could reflect the empathizer's uncertainty regarding how best to respond and the expected outcome of their actions. The second level, estimation uncertainty, captures unknown information about the probabilistic relationship between actions and outcomes, such as between an empathic response and its effectiveness. The third level, volatility uncertainty, represents how this relationship changes over time. HGF has been applied to study how individuals with conditions like anxiety or autism process uncertainty in social or other learning tasks (de Berker et al., 2016; Lawson et al.,2017; Sevgi et al., 2020).

People with autistic traits or anxiety often face challenges with emotion regulation (Cisler et al., 2010; Mazefsky et al., 2013) and processing uncertainty in dynamic (social) situations (Browning et al., 2015; Hodgson et al., 2017; Nicholson et al., 2019; Sevgi et al., 2020; Van de Gruys et al., 2014). Importantly, these individuals also exhibit aberrant interoceptive processing (Füstös et al., 2013; Garfinkel et al., 2016; Pollatos & Traut-Mattausch, 2009;

Zamariola et al., 2019). Recent evidence and theoretical accounts suggest that these problems may stem from inflexible and maladaptive expectations, interpretations and attention regarding internal bodily states, leading to impaired self-regulation (Owens et al., 2018; Paulus et al., 2019; Smith et al. 2018; 2020). Given that such difficulties are magnified under conditions of increased uncertainty, HGF modelling provides a framework to explore how inflexible interoceptive predictions contribute to challenges in regulating emotions - both for oneself and others. Interestingly, a study using HGF found that the extent to which fluctuations in autonomic arousal follow the changes in uncertainty is related to learning performance (de Berker et al., 2016), indicating the importance of adaptive interoceptive processing in navigating dynamic environments.

One way to examine interoceptive PP is through the analysis of the EEG signal time-locked to cardiac activity, an ERP component usually referred to as heartbeat-evoked potential (HEP) (Park & Blanke, 2019). HEPs are thought to reflect the cortical processing of cardiac activity and its amplitude has been shown to be modulated by several relevant factors, such as overt attention to interoceptive states (Petzschner et al., 2019) and interoceptive abilities (Pollatos & Shandry, 2004). HEPs have also been recently linked to affective and interoceptive PP (Ainely et al, 2016; Gentsch et al., 2018). Despite these findings and related theories, only one recent study has examined how HEPs are related to PP and learning, and specifically PEs, within a computational framework (Fouragnan et al., 2024). In this study, researchers used a reward-based learning task and modelled their behavioural data with the Rescorla-Wagner model to observe that absolute PEs were associated HEP amplitude during outcome presentation. However, because the Rescorla-Wagner model uses a fixed learning rate, this approach does not capture the adaptability of interoceptive predictions under varying conditions of uncertainty. Furthermore, more flexible models can reveal individual differences in interceptive prediction adjustments in dynamic affective environments.

Importantly, HEPs have been associated not only with interoception but also with socio-affective processes. For instance, HEP amplitudes are associated with higher sensitivity in recognizing others' emotions (Terasawa et al., 2014), affective judgments, and self-reported empathy scores (Fukushima et al., 2011). This suggests that HEPs could be a valuable tool to examining interoceptive predictions within a social context, such as in an empathic reinforcement learning task.

In the present study, we examine how interoceptive processing underpins learning and decision-making in an adaptive empathy task by using the HEP as an index of interoceptive prediction and attention. The main task of the study is an adaptation of Kozakevich Ardel and colleagues' (2021) task where participants need to choose among two options - representing distraction and reappraisal strategies (Sheppes et al., 2011) - to alleviate the distress of a virtual character facing a different distressful situation in each trial. To probe implicit learning of the characters' preferred strategy, participants are not informed that the two available options correspond to two specific emotion regulation strategies (ERS). One main difference from the original task is that there is only one empathic target and instead of two. Another key difference is the contingency schedule between emotion regulation strategies and outcomes, where we introduce two volatility phases: a low-volatility and a high-volatility phase. To assess participants' own preferred strategy and potential egocentric biases in relation to empathic learning and interoceptive processing, we asked participants, in a prior online session, to indicate which response they would prefer to receive if they were in the same distress scenarios presented. Additionally, to investigate whether individual differences in socio-affective processing are linked to social learning and decision-making differences as well as to distinct interoceptive processing profiles, we included three relevant questionnaires. We also incorporated the Heartbeat Attention Task (Petzschner et al., 2019) to examine how intentional adjustments of interoceptive attention may modulate empathic learning and decision-making. To capture individual differences in learning, we modelled the behavioural data using the HGF, linking pwPEs to HEPs during the feedback phase of task. Additionally, we examined whether HEP amplitudes during decision-making influence their empathic responses. This allowed us to investigate whether the interoceptive processing tracks pwPEs related to social feedback and decision-making processes, as well as its underlying cognitive and affective predictions at both group and individual levels.

We predicted that HEP amplitudes during social feedback will vary based on the magnitude of absolute pwPEs (Fouragnan et al., 2024) generated by the HGF. Additionally, we predicted that the degree of adaptability in HEPs to parallel pwPEs of the empathic learning task will be associated with individual differences in affective empathy and autistic traits. Furthermore, we hypothesized that these traits would predict HGF's learning and decision-making parameters. Lastly, we expected that the last the two relationships described will be moderated by individual differences in interceptive attention (or precision) changes, as reflected in the HEP analysis of the Heartbeat Attention task.

4.2. **Methods**

4.2.1. **Participants**

A total of 48 participants, students in the School of Psychology of the University of Kent, were recruited to take part in all conditions of our tasks, including the online questionnaires.

However, we excluded those with incomplete data (N = 3) or with noisy EEG data due to different technical issues (N = 6). Thus, the final dataset comprised 39 participants (29 women, 1 non-binary, 9 men; age range: 18-53; mean age 19.8 with normal or corrected to normal vision. In addition, they were healthy with no history of psychiatric or neurological disorders. The sample size was calculated based on previous experiments using similar computational modelling methods with healthy, neurotypical participants (de Becker et al., 2016; Sevgi et al., 2020). All participants provided written informed consent before the begging of the experiments. The experiments were approved by the University of Kent Ethics Committee.

4.2.3. **Adaptive empathy task (AET)**

Procedure

In the adaptive empathy task, participants are first shown a description of a distress scenario for 3 s accompanied by an image of the virtual character “Amy”, displaying a stressed-fearful facial expression (Fig. 1a). Two empathic responses options are then presented in left and right boxes on either side of the image. One response represents the emotion regulation strategy of reappraisal while the other one represents distraction (Sheppes et al., 2011). The positions of these responses are randomized and counterbalanced to avoid an association between response type and box side. Examples of responses include: ‘We can go on a weekend trip away from the city. I have a few good places to suggest’ (distraction) and ‘The most important thing is that you were not there at the time and weren’t hurt’ (reappraisal). Participants had 12 s to read the responses and make a decision, but they were only allowed to respond after the first 6 s. Following their selection, feedback appeared on the screen for 3 s, showing Amy’s face turning happy if the correct response was chosen and sad if it was incorrect. A 2s inter-trial-interval (ITI) followed. Participants were seated at a distance of 60-70 cm from the monitor where stimuli were presented.

Design

Sixty different everyday distress scenarios of medium severity were selected, each evaluated by 20 people in a pilot study. Each scenario was presented twice (with the same empathic responses), resulting in a total of 120 trials. Each trial lasted approximately 16 s, making the entire task about 35 min in duration, including two breaks for participants to rest. The task began with 4 practice trials. The main task started with a stable condition block in which one empathic response strategy type was correct with a probability of ~73% (order counterbalanced across participants). This was followed by a volatile phase consisting of 5 mini-blocks, where the probability of one strategy being correct was 80% and alternated between the two strategies every 10 trials. Finally, the second stable block followed, but with the alternative empathic response now correct in ~73% of trials. In total, 80 trials were categorized as stable and 60 as volatile.

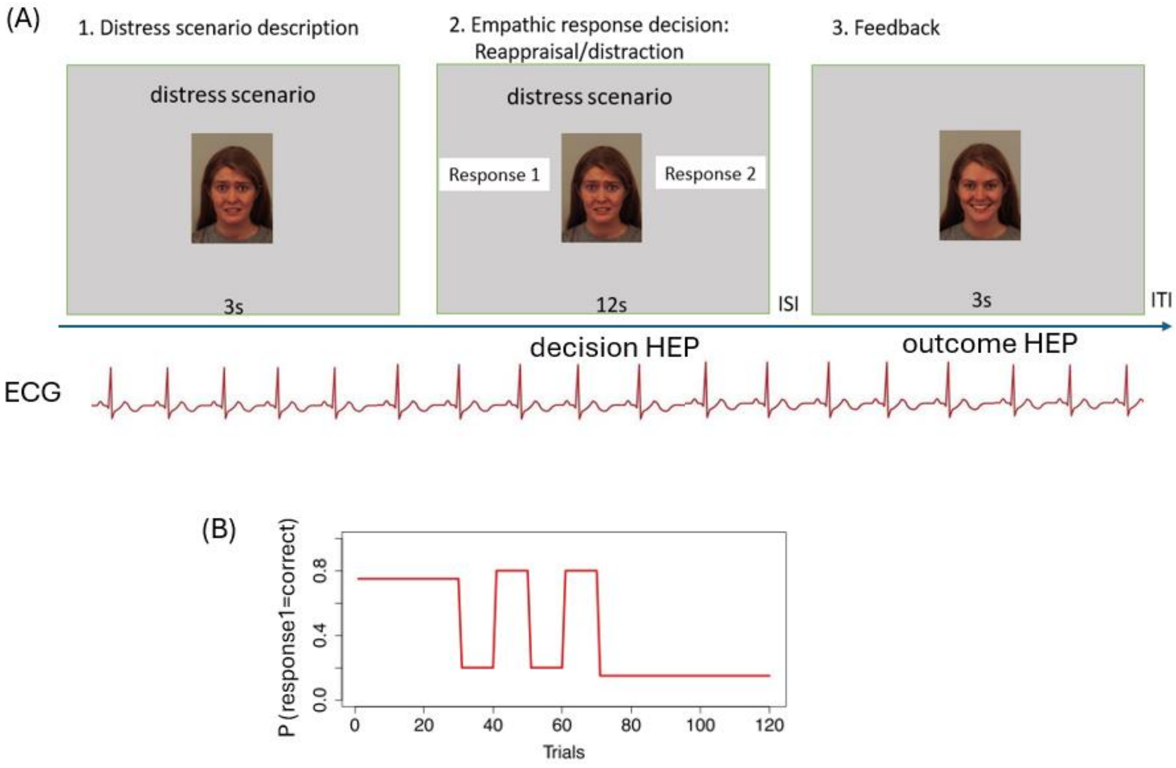


Figure 4.1. Adaptive Empathy Task and probability schedule. (A) Experimental design. The adaptive empathy task comprises three phases: 1. scenario presentation; 2. emotion regulations option and decision; 3. feedback/outcome presentation. Participants first read the distress scenario and then select among two ER options (reflecting reappraisal and distraction strategies) presented randomly and counterbalanced in a left and right box. If the correct option was selected, a happy face is presented as feedback and if the wrong option is chosen a sad one is displayed. ER strategies are not labelled to avoid participants being aware of the existence of these 2 response categories. HEP are measured at specific time points of the decision and feedback phase. (B) Probability schedule. Probability of one of the two ER strategy responses being the correct one; the alternative ER strategy follows the opposite probability schedule.

Stimuli and paradigm

Our tasks were programmed and presented in Matlab (Mathworks, version: 2022b) using the Psychtoolbox (http://psychtoolbox.org/). The stimuli presented in this task were taken from the FACES database (Ebner et al., 2010). We used only happy, sad and fearful expressions of one white female model.

The empathy task we designed is an adaptation of the adaptive empathy task developed by Kozakevich Arbel and colleagues (2021). In their paradigm, there were two virtual characters (instead of one in ours) with stable ER preferences throughout the 20 trials of task, such that one ER strategy was correct for each character at 80% of times. In contrast, our task focused on a single character and incorporated volatility, requiring participants to learn the character's ER strategy and detect when it changed due to unknown contextual factors. The distress scenarios used by Kozakevich Ardel and colleagues were based on "everyday life situations related to relationships, work, daily routines, and the like". The use of the two ER strategies was decided based on the related literature suggesting that reappraisal and distraction (or expressive suppression) to be the most commonly used ER approaches. Within the emotion regulation model of Gross (2001), cognitive reappraisal is a strategy focusing on the (re)interpretation of the event in a way that alleviates its emotional impact. Whereas suppression is a response-focused strategy to distract attention away from the distressing situation and diminish the negative emotional experience.

For our task, we used the 20 distress scenarios from Kozakevich Ardel's and colleagues (2021) study and added 40 new scenarios that were tested in a pilot study. We aimed for scenarios of moderate severity that allowed both ER strategies to be perceived as plausible responses. Therefore, we created 50 distress scenarios, each paired with 50 different reappraisal and distraction responses and asked the participants of the pilot study to rate each scenario's severity and the likelihood of choosing each response. Responses were given in 10-point Likert scale from 1 ("I would always choose distraction") to 10 ("I would always choose reappraisal"). Thus, we excluded scenarios where the two ER strategies appeared less equiprobable, specifically when appraisal was selected more than ~65% on average (usually the most preferred option was appraisal).

4.2.4. **Questionnaires**

All questionnaires were created and presented online in Qualtrics platform. Participants completed the following questionnaires prior to the lab tasks.

Emotion Regulation Preference questions (ERPQ)

First, participants read the 60 distress scenarios and were instructed to report their preferred emotion regulation response if they found themselves in these scenarios. They responded on a scale from 1 to 10, with the 1 being always prefer the first option (distraction) and 10 being always prefer the second option (reappraisal). Next, participants evaluated 10 of these scenarios under two different imagined conditions: experiencing scarcity and having a family member facing a severe health issue. These additional contexts were included to illustrate how contextual parameters might change emotion regulation preferences. This was done to demonstrate the potential variability in ER preferences, making the volatility of the virtual character's ER preferences in the subsequent AET more realistic and reasonable.

Autism-Quotient (AQ)

The Autism-Quotient is a self-report questionnaire developed by Baron-Cohen and colleagues (2001) to measure autistic traits in normal adult populations. It included 50 statements where participants have to report their agreement in 4-point Likert scale from ‘definitely agree’ to ‘definitely disagree’.

Interpersonal Reactivity Index (IRI)

The IRI is a self-report questionnaire that measures difference in empathic social interaction and included 4 subscales assessing perspective taking, fantasy (imagination), empathic concern and personal distress. The first two subscales assess cognitive empathy while the last two affective empathy; here we focused only on affective empathy dimensions. It includes 28 items where participants respond in 5-point Likert-type scale from 0 (does not describes me well) to 4 (describes me very well).

4.2.5. Heartbeat attention task

The HbAttention task (Petzschner et al., 2019) comprises an interoceptive attention block (where participants had to focus on their heart rate) and an exteroceptive attention block (where participants had to focus on a sound). The two alternating blocks had a duration of 20s each and were separated by a resting period of 9s and an ITI of varying duration (5s-15s).

During the HEART condition, a heart image was shown on screen. In this block, participants were instructed to focus their attention on their heart, without measuring their pulse though. Throughout the SOUND block, a headphone symbol was shown while participants were instructed to focus their attention on a white noise played from two speakers and attend to any potential changes in the sound. The sound was played during both blocks to ensure that any change in brain responses was related to attention shifts and not to sensory input changes.

Additionally, for both conditions, participants were asked to respond to questions associated with the previous block. In these questions, they rated aspects of their perception like ‘How well were you able to concentrate on the white noise in the last block?’ or ‘How much would you associate your perceived heartbeat in the previous block with the colour red?’ and had to respond in 10-point scale. The purpose of these questions was solely to provide an additional motivation to focus their attention on the task throughout its duration.

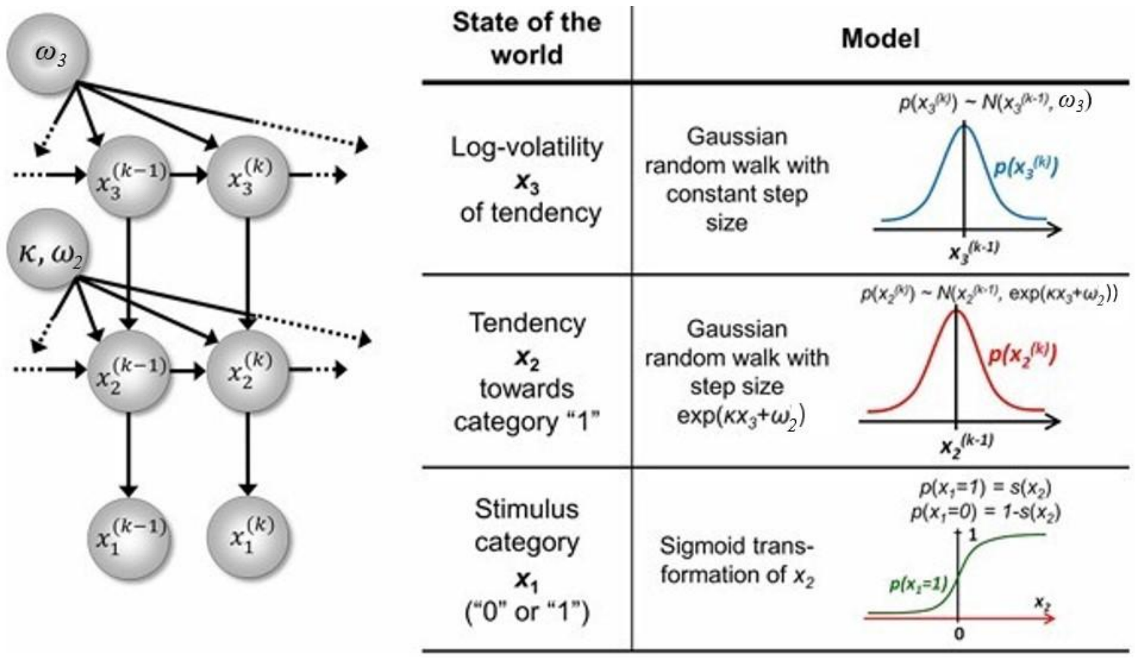
4.2.6. Computational modelling

In this study, we observed how participants learned to recognize the preferred emotion regulation (ER) strategy of a virtual character, using clear and immediate feedback to associate their actions with outcomes. However, the relationship between empathic actions and emotional responses was probabilistic and changed throughout the task, presenting participants with limited information to interpret social cues under varying conditions of uncertainty. By accumulating available information, participants could infer the character's unobserved mental states (i.e., emotion regulation preferences). To examine how participants use the available social information to learn and makes decisions in our adaptive empathy task, we used a hierarchical Bayesian learning model named HGF. Such Bayesian models provide a formal framework for modelling the evolution of an observer's beliefs about another's mental states over time. Since each belief state depends on the previous one, this approach can be viewed as a partially observable Markov decision process (POMDP), which defines the relationship between observed environmental states and the unobserved mental states of the other (Baker et al., 2009).

The HGF is a generic Bayesian learning model that quantifies belief states and their associated uncertainties across multiple levels (Fig.2). Each belief state is represented as a gaussian distribution, where the mean value represents the current belief state, while the variance reflects the uncertainty around this belief, with precision ($\pi_i^{(k)}=1/\sigma_i^{(k)}$) being the inverse variance. In our 3-level HGF model, each level is denoted x_1^k, x_2^k,..., x_n^k, where k is the trial number. The model's lowest level x_1 represents outcome uncertainty (or irreducible uncertainty) of the effectiveness of the ER responses coded as binary categories (1 for reappraisal being the correct response, 0 for distraction). In the HGF version implemented here, these beliefs are stochastically mapped to corresponding binary inputs with reappraisal coded as $u^k = 1$ and distraction as $u^k = 0$. The second level x_2 represents the character's tendency to prefer one ER response over the other, which fluctuates over the task according to a predefined schedule (estimation uncertainty). The third level x_3 represents the volatility of these preferences, capturing how frequently the character's ER preferences change over time (volatility uncertainty). The model's three levels capture different forms of uncertainty to produce predictions for the specific social environment. The three representation levels are hierarchically coupled so that each one is influenced by the one above it. Specifically, the first level is coupled to the second according to the sigmoid function of Equation 1:

$$\hat{x}_1^{(k)} = s\left(x_2^{(k-1)}\right) = \frac{1}{1+exp(-x_2^{(k-1)})} \quad \text{(Equation 1)}$$

At the second and third level, beliefs evolve as Gaussian random walks in the following way. The second level's step size is controlled by the third level through the equation: $x_2^k \sim N(x_2^{k-1}, \exp(\kappa x_3 + \omega_2))$, and similarly, the third level's step size is determined by $x_3^k \sim N(x_3^{k-1}, \exp(\omega_3))$. The three parameters κ, ω_2, and ω_3 capture individual differences in belief updating within this social context: κ represents the influence of volatility at the third level on the belief updating rate at the second level; ω_2 is the tonic volatility at level two, reflecting beliefs about how easily ER preferences may change; and ω_3, or 'meta-volatility', captures beliefs about environmental volatility, with higher values indicating a greater readiness to update x_3 in response to perceived instability. Here, the parameter κ is fixed to 1, defining an invariable relationship between the second and third level across participants.



ω_3
$x_3^{(k-1)}$
$x_3^{(k)}$
κ, ω_2
$x_2^{(k-1)}$
$x_2^{(k)}$
$x_1^{(k-1)}$
$x_1^{(k)}$
State of the world
Model
Log-volatility x_3 of tendency
Gaussian random walk with constant step size
$p(x_3^{(k)}) \sim N(x_3^{(k-1)}, \omega_3)$
$p(x_3^{(k)})$
$x_3^{(k-1)}$
Tendency x_2 towards category "1"
Gaussian random walk with step size $\exp(\kappa x_3+\omega_2)$
$p(x_2^{(k)}) \sim N(x_2^{(k-1)}, \exp(\kappa x_3+\omega_2))$
$p(x_2^{(k)})$
$x_2^{(k-1)}$
Stimulus category x_1 ("0" or "1")
Sigmoid trans-formation of x_2
$p(x_1=1) = s(x_2)$
$p(x_1=0) = 1-s(x_2)$
1
$p(x_1=1)$
x_2
0

Figure 4.2. The 3-level HGF generative model. In our task the lowest representation level x1 represents outcome uncertainty with the two ER categories coded as 1 (reappraisal) and 0 (distraction), x2 represents the (fluctuating) preference of the virtual character towards ER towards reappraisal, and x_3 represents the volatility of this preference/tendency. $x_1^{(k)}$, $x_3^{(k)}$, $x_3^{(k)}$, are the hidden states of the social environment at each time point (k). Inputs $u^{(k)}$ values are generated for each time points based on the values of the previous time point (k-1) of the HGF levels, and the learning parameters ω_2 and $\omega_{3.}$ The figure has been adopted from Mathys et al. (2014).

In this framework, individuals perform trial-by-trial belief updating by calculating posterior probabilities, applying an approximation of ideal Bayesian inference. Here, a mean-field approximation is applied, yielding the following update equations:

$$\Delta x_i^{(k)} = x_i^{(k)} - x_i^{(k-1)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \delta_{i-1}^{(k)} \qquad \text{(Equation 2)}$$

According to Eq. 2, beliefs (posterior means) at each level (i) are proportional to the PE (δ_{i-1}) weighted by a precision ratio. The PE term quantifies the difference between the lower-level prediction $\hat{\mu}_{i-1}^{(k)}$ before the input and the posterior expectation $\mu_{i-1}^{(k)}$ afterwards. This error term is weighted by the ratio of the precision of the prediction of the lower level $\hat{\pi}_{i-1}^{(k)}$ before the input to the precision of the current belief $\pi_i^{(k)}$. Precision here refers to the inverse variance of the posterior probability distribution. This ratio of precision estimates represents the adaptive learning rate at each level, where the second level's learning rate can be transformed to reflect that of the first level.

The HGF serves as the perceptual model in our study and is used in conjunction with a response model. Response models are functions for mapping beliefs generated by the perceptual model to the observed responses. Here, we will used as a response model the unit-square sigmoid function for binary responses (Ingesias et al, 2013), which has the following form:

$$f(y=1)=\frac{\hat{x}_1^{\zeta}}{\hat{x}_1^{\zeta}+(1-\hat{x}_1\ )^{\zeta}} \quad \text{(equation 3)}$$

f(y=1) represents the probability of choosing a particular response (reappraisal) based on the belief state x_1, with values closer to 1 indicating a higher probability of selecting reappraisal. The sensitivity parameter ζ, reflects exploration tendencies and decision noise. Higher values of ζ make choices more deterministic, while lower values indicate increased randomness or exploration. This response model enables us to investigate how participants' inferred beliefs about the correct response translate into action probabilities, with the sensitivity parameter ζ adjusting the influence of these beliefs in dynamic and uncertain contexts.

Model inversion

The priors of our model parameters were selected based on the experimental design and the questionnaires regarding ER preferences. We opted for relatively uninformative priors, setting large variances values to allow for adjustments based on individual differences in empathic learning and decision-making. Similarly, for parameters bounded between 0 and

1, we set the prior mean at 0.5. Details on all priors used in the HGF, response, and RW models are provided in the Appendix (B.3).

Parameters and states were estimated in unbounded spaces, thus confined parameters (between 0-1) were log-transformed. Using the TAPAS toolbox (Frässle et al., 2021), we estimated the maximum-a-posteriori (MAP) of model parameters, given the priors provided and input sequence for each participant. Optimization was performed with a quasi-Newton algorithm, maximizing the log-joint posterior density over all perceptual and observation parameters based on the data and generative model. Finally, we calculated the log model evidence (LME) for each model and participant, which measures the likelihood of observed data under a given model and penalizes for model complexity. These LME values were used for model comparison and selection.

Model space and model comparison

To test complex theoretical hypotheses, we formulated competing cognitive models and compared them to identify which best explains the data. Our primary model for individual differences in empathic learning is the 3-level HGF (HGF3). To examine whether participants take into account environmental volatility, we included a 2-level HGF model without a third level. Lastly, we tested a Rescorla-Wagner (RW) model to determine whether participants adopt a stable learning rate. All learning models where combined with the same sigmoid response model. Thus, our model space comprised four models, evaluated using a random-effects Bayesian model selection (BMS) procedure. More specifically, using the spm_BMS in SPM12, we performed model selection by entering the LME values obtained for each participant. This analysis provided: (a) the expected posterior probabilities, indicating each model's prevalence in the general population, and (b) the exceedance probability, which reflects the likelihood that a given model outperforms all others.

4.2.7. EEG and ECG data acquisition

For the EEG/ECG recording we used 32 Ag/AgCl active electrodes mounted on an elastic electrode cap (Easycap with sizes 54, 56 and 58 cm head circumference). The continuous EEG signal was recording using BrainAmp amplifiers (BrainProducts, Munich, Germany; 0.1 μV 1 analog-to-digital conversion resolution; 1000 Hz sampling rate; 0.01-100 Hz online cut-off filters). The ground electrode was placed at FCz and the reference at AFz. All electrodes were referenced off-line to the arithmetic average of all electrodes. Additionally, two electrodes (TP7, TP8) were used to record eye movements and blinks (EOG). The one EOG electrode was placed on the zygomatic bone under the right eye to capture vertical eye movements and the other one was place on the corner of the left eye to capture horizontal eye movements. An ECG electrode (Oz) was placed under the left collarbone to record cardiac activity.

4.2.8. EEG preprocessing

To carry out all steps of electrophysiological data preprocessing we used the functions provided by fieldtrip (https://www.fieldtriptoolbox.org/). First, the raw data was band-pass filtered between 0.5-30 Hz and downsampled to 500 Hz. After visual inspection of the signal noisy channels were detected and interpolated using the average of the neighbouring channels. Around 1-2 channels were interpolated for each participant in this way. Then, the detection of R-peak was performed using the Pan-Tompkins algorithm in MATLAB. The R-wave of the QRS complex of cardiac contraction determines the first timepoint of the HEP. Next, the continuous data was segmented into intervals time-locked to either the onset of the feedback, the R-peaks onset of the heartbeats occurring during the feedback period, or the onset of the decision phase. Thus, the time interval for the feedback-related analysis lasted 3.5s, starting 0.5s before the feedback onset. The intervals time-locked to the decision phase lasted 3.5s, starting 2.5 before the response timepoint. We identified the heartbeats happening within these periods and created intervals time-locked to each R-wave with 0.8s duration, starting 0.2 before the R-wave. This segmentation was done separately for the correct and incorrect trials and the volatile and stable phases.

Then, using the Fieldtrip toolbox we performed automatic artifact rejection where trials and channels whose variance (in z scores) surpassed a specific threshold (20 µV) were excluded. This was the first step to remove muscle and eye-related artifacts. Next, the segmented signal was entered in an ICA (independent component analysis) procedure (RUNICA, logistic infomax algorithm). The components that resembled the timing and spatial distribution of physiological artifacts, like saccades, eye-blinks and the volume-conducted cardiac-field artifact were removed. Thus, around 3 components per participant were excluded. The ECG and EOG channels were excluded from subsequent analyses and the signal was re-referenced to the arithmetic average of all EEG electrodes.

The resulting cleaned segments were baseline-corrected using an interval from -0.9 to -0.1 for the segments time-locked to the feedback, an interval from -0.15s to -0.05 for the segments time-locked to the R-wave, and an interval from -0.3 to -0.1 for the segments time-locked to the response.

4.2.9. HEP analysis

The typical topography of the HEP has a frontal-to-parietal distribution over the scalp (Petzschner et al., 2019; Pollatos & Schandry, 2004), most often over the right hemisphere

(Pollatos & Schandry, 2004; Schulz et al., 2015). Like previous analysis detecting HEP effects due to learning (Fouragnan et al., 2024), we used Fieldtrip functions to analyse subject-wise time-series as described in detail by Moris and Oostenveld (2007). Specifically, we used a non-parametric cluster-based permutation test. This test is a popular approach to deal with the high dimensionality of the EEG data that causes the multiple comparison problem making difficult to control the so-called type 1 error using standard inferential statistical methods. T-values from adjacent temporal and frequency points with p-values less than 0.05 were clustered by summing their t-values. This cumulative statistic was used for inferential statistics at the cluster level. The procedure, involving calculating t-values at each temporal point and clustering adjacent t-values, was repeated 5000 times with randomized swapping and resampling of subject-specific time-frequency activity. Thus, this Monte Carlo method yields a nonparametric estimate of the p-value denoting the statistical significance of each cluster.

In our HEP analysis, we included all electrode sites and the entire time window (0.2-0.6 seconds after the ECG's R-peak) where the HEP typically occurs. We grouped all scalp electrodes into 8 a-priori regions of interest (ROIs – clusters: fronto-central, right and left frontal, right and left centro-parietal, right and left temporo-parietal, and occipital). The electrodes of each cluster are presented in the Appendix (C.1). Then, we applied the cluster-based permutation approach from FieldTrip to assess whether the HEP varied based on learning dimensions and response outcomes during the feedback and decision phases of the task.

Given that this method allows for the comparison of only two conditions, we categorized the trials into two groups. For the feedback phase, we computed average signals by aggregating, on a subject-level, trials: a) with high absolute pwPEs versus low absolute pwPEs for the second and; b) third HGF level; c) as well as trials with correct versus incorrect outcomes. For the decision phase, we aggregated trials based on: d) response outcome (correct vs incorrect); e) as well as high and low second and f) third level learning rate. For the a, b, e and f comparison we had around 35-40 trials per comparison condition and for the c and d around 55-60 trials per condition. Subsequently, at the group level, we run the 6 comparisons on the averaged heart-evoked potential (HEP) data based on the above conditions for the feedback and decision phase. We employed a within-subject two-tailed cluster-based permutation analysis for these 6 contrasts over the 8 predetermined ROIs and the time-window of interested after the R-peak. This approach effectively addresses the multiple comparison problem, avoiding biases associated with pre-selecting time-windows and minimizing type I error inflation. To prevent spurious findings, effects shorter than 15 milliseconds were excluded from further analysis. In addition, p-values were adjusted for multiple comparisons using the Bonferroni-Holm correction method.

To extract HEPs for the interoceptive attention task we followed the same preprocessing procedure as described above. Firstly, we segmented the 20s period of each block (Heart and Sound condition) based on the R-peaks and averaged them across the two attention conditions. Then, we compared the averaged signals for each condition in the cluster permutation analysis across the same 8 ROIs in the time window of 200-600ms after the R-peak.

4.2.10. Regression analysis for individual differences

To explore how individual differences in social interaction, autistic traits and interoceptive attention relate to interoceptive predictions and empathic learning and decision-making, we run a series of linear regressions with the questionnaires scores and the results of the HbAttention task as predictors. To assess individual differences in the extent to which empathic predictions are reflected to interoceptive predictions, we estimated per individual the peak HEP amplitude across the trials with high and low absolute pwPEs over the electrodes and the time window where the most significant cluster is detected via the permutation analysis for the comparison of these two conditions (high and low absolute pwPEs). Then, the difference calculated between peak HEP amplitudes for those two sets of trials per individual (ΔHEPpe) was the DV of the first regression with predictors the dimension of affective empathy of the IRI (i.e., Empathic Concern and Personal Distress) and the AQ score, along with the ΔHEPatt as an interaction term and the ERP as a control variable. ΔHEPatt was the corresponding difference in peak HEP amplitude between the interoceptive (Heart) and exteroceptive (Sound) attention condition in the second significant cluster detected (424-470ms). So, the regression has the following equation:

$$\Delta HEPpe \sim (IRI\_EC + IRI\_PD + AQ) * \Delta HEPatt + ERPQ \qquad \text{(Equation 4)}$$

Then we run regressions on 5 HGF parameters with the same predictors and interactions terms as shown in eq. 4. These parameters were the learning rate parameters of the second and third level (ω2, ω3), the absolute difference between the average learning rate of the stable and volatile phase across in the second and third level (ΔLR2, ΔLR3), as well as the parameter ζ of the response model.

4.3. Results

To examine the learning performance, we ran a 2x2 mixed ANOVA with volatility Phase (stable, volatile) as within-subjects factor and ER Order (reappraisal first, distraction first) as between-subjects factor, with mean accuracy as DV. The results revealed (Figure 2A) a significant main effect of Phase (F(1,38)=11,236, p=0.003, $\eta^2{}_p$=0.124) with participant being more accurate in the stable (M=59.2%, SD=9.12) than the volatile phase (M=53.0, SD=7.87). There was no effect of ER Order or Order x Phase interaction (p>0.219).

Bayesian Model Comparison

To test which learning model best fits to our behavioural data we performed Bayesian model comparison among 2 HGF models and the RW model. Model comparison (Figure 2B) based on individual LME showed that the HGF3, i.e. the model with volatility estimates, with exceedance probability of 0.91 (posterior probability = 2.03) outperformed the HGF2 the model without volatility and the Rescorla-Wagner model, with exceedance probability of 0.08 (posterior probability = 10.43) and 0 accordingly (posterior probability = 17.52). This model comparison strongly indicates that the Bayesian hierarchical model that accounts for belief updating of volatility expectations can better explain participants' empathic learning progress.

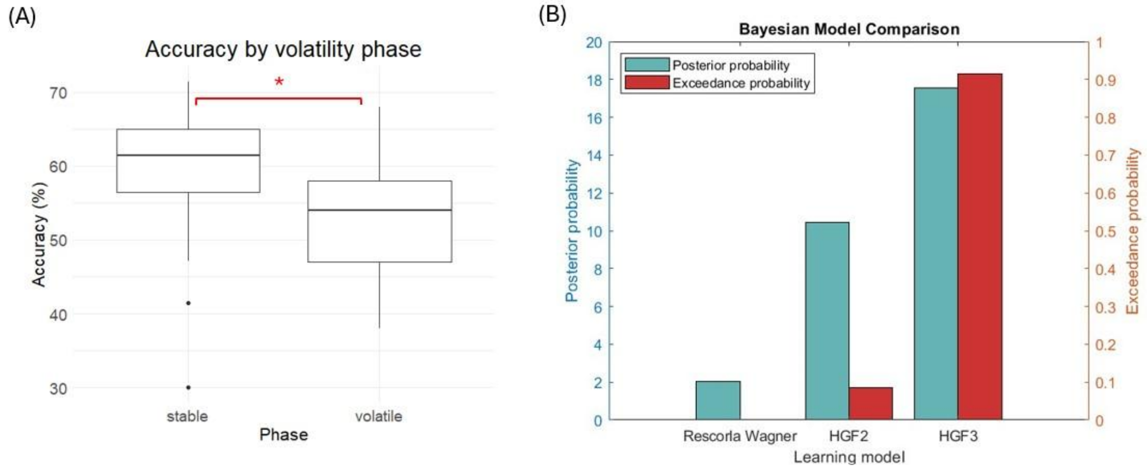


Figure 4.3. Mean accuracy and model comparison. (A) Mean accuracy score for the two volatility phases (B) Bayesian model comparison showed that the HGF with volatility level outperformed the other two learning models.

HEPs adaptive empathy

The results of the cluster-based permutation analysis for the feedback HEPs revealed 3 significant clusters where HEPs differed between trials with high absolute and low absolute pwPEs. There was a significant cluster over the left fronto-central (FP1, F3, F7, FC5) electrodes (p=0.029) in the time window from 461 to 568 ms after the R-peak. The second cluster was over the right fronto-central (FP2, F4, F8, FC6) electrodes (p=0.001) in the time window 482 – 590 ms. The third cluster was over the left centro-parietal (CP1, CP5, C3, P5) sites (p=0.021) in between 474 and 515 ms after the R-peak. The HEP activity demonstrated opposite polarity over the left-to-right axis across the sculp, i.e., in the left clusters high pwPE HEP exhibited higher negativity compared to the low pwPE HEPs, while in right clusters high pwPE HEP exhibited higher positivity. The HEP analysis for the pwPEs of the third HGF level as well as for the response outcome did not reveal significant clusters (p>0.082). Time-course activity and topographic distribution of feedback HEPs are depicted in figure 3.

The analysis for the decision phase HEPs revealed a significant cluster when comparing trials based on response outcome (Fig. 4) indicating that HEP fluctuations can predict accuracy of empathic decision-making. The cluster was detected over right temporo-parietal electrodes (TP10, T8, P8) in the time-window 466-500 ms (p=0.042), with HEPs preceding correct trials showing larger negativity than those preceding incorrect trials. Conversely, no significant clusters were found when comparing HEPs for trials with high and low second level learning rates.

HEPs interoceptive attention

Regarding the attention-related modulations of HEPs, permutation analysis revealed a significant cluster over the frontocentral electrodes (FC1, FC2, Fz, Cz) in the time window between 231 and 287 ms and between 424 and 470 ms. Thus, interoceptive attention changes had a significant effect on HEP amplitude over 4 central and frontal electrodes within an early and a late time-window.

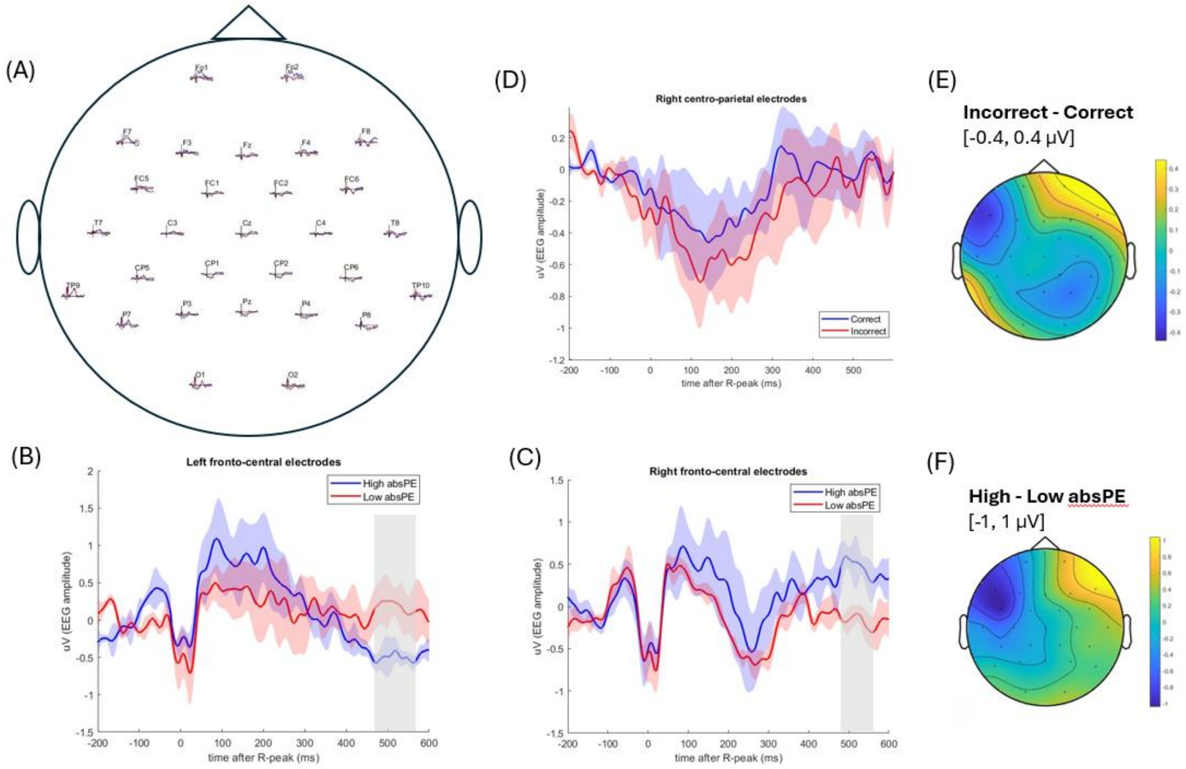


Figure 4.4. Feedback HEPs. (A) Grand average EEG activity time-locked to the R-wave across all electrodes during the feedback phase for the comparison between trials with high and low absolute pwPEs for the learning of emotion regulation preferences (second HGF level). (B-D) Averaged HEPs after feedback. Clusters with significant HEP differentiation were found for absolute pwPEs but no significant cluster for response outcome. Shaded areas represent the time-windows where significant difference was detected between conditions. (E-F) Topographic scalp distribution of the μV difference for the two comparison conditions averaged over the time window 460-560 ms after the R-wave.

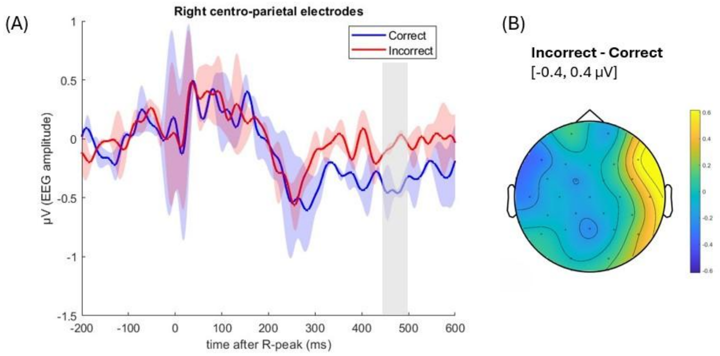


Figure 4.5. Decision HEPs. (A) Averaged HEPs during the decision phase for the significant cluster detected for the comparison of trials with correct and incorrect response. (B) Topographic scalp distribution of the μV difference for the two comparison conditions.

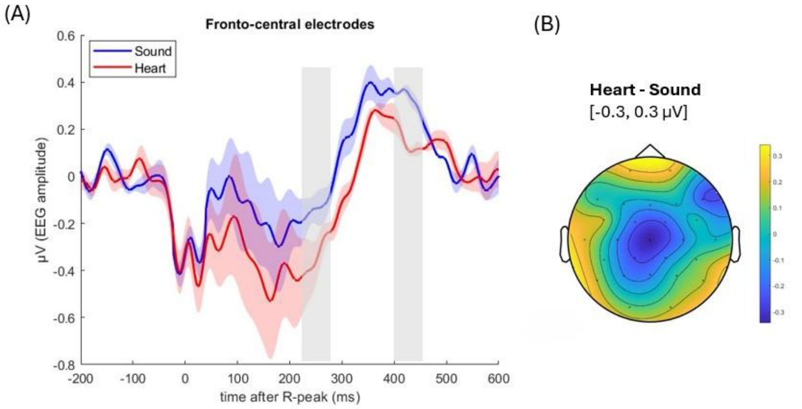


Figure 4.6. Attention HEPs. (A) Averaged HEPs for the two conditions of the HbAttention task (attention to the heart vs sound) over the electrodes where significant clusters were detected. (B) Topographic scalp distribution of the μV difference for the two comparison conditions averaged over the time window of the first significant cluster (231-287 ms).

Individual differences

The first regression analysis on HEP difference between the trials with high vs low second level pwPEs revealed as significant predictor the AQ score (β=-0.694, t=-2.549, p=0.016), with EC of IRI relatively close to significance as well (β=-0.445, t=-1.885, p=0.070), meaning that individuals with higher reported autistic traits showed less modulation of HEPs between the trials with high and low absolute pwPEs (ΔHEPpe).

Then, we examined whether questionnaires scores predict HGF parameters with HEP difference in the interoceptive attention task (ΔHEPatt) as interaction term in these regressions. The first regression showed that ω2 was negatively predicted by ΔHEPatt (β=-11.304, t=-3.273, p=0.003). Empathic Concern (β=1.199, t=2.005, p=0.055) and Emotion Regulation Preference were close to significance (β=-0.602, t=-1.921, p=0.064), while all other predictors and interactions were not significant. Therefore, individuals with lower modulation of interoceptive attention had higher learning rate of the changes regarding the effectiveness of empathic regulation strategies. In addition, higher empathic concern and stronger preference for distraction, relative to reappraisal, as emotion regulation approach for themselves might also contribute to increased values of this learning rate (ω2).

DV	Predictor	β	t-value	p-value	interpretation
ΔHEPpe	ASD	-0.694	-2.549	**0.016**	Higher AQ → Less ΔHEPpe modulation
ω2	ΔHEPatt	-11.304	-3.273	**0.003**	Lower ΔHEPatt → Higher 2nd level learning rate
	Empathic Concern	1.199	2.005	0.055	Higher EC → Higher 2nd level learning rate
ω3	Personal Distress	-1.187	-2.112	**0.044**	Higher PD → Lower 3rd learning rate
ΔLR2	ΔHEPatt	-0.359	-1.977	0.058	ΔHEPatt ~ ΔLR1
ΔLR3	None				No significant predictors
ζ	ASD	-2.979	-2.308	**0.028**	Lower ASD → More consistent responses
	ΔHEPatt	11.125	3.793	**<0.001**	Higher ΔHEPatt → More consistent responses

Table 4.1. Regression Results on Individual Differences. This table presents the significant (and near significance) regression results. Predictors are individual traits (i.e., empathic concern, personal distress, autistic traits) and interoceptive attention modulation (ΔHEPatt), while dependent variables are learning and decision parameters as well as HEP modulation based on feedback pwPEs (ΔHEPpe). ω2, ω3 = second-level and third-level learning rate; ζ = decision parameter of the response model; DLR2, DLR3 = learning rate differences between stable and volatile phases at the second and third hierarchical levels, respectively.

The second regression showed that ω3 was negatively predicted by Personal Distress (β=-1.187, t=-2.112, p=0.044), indicating that individuals with higher personal distress during empathic learning have lower estimations of social volatility in this context. In the third regression on ΔLR1, we did not find any significant predictor with only ΔHEPatt approaching significance (β=-0.359, t=-1.977, p=0.058). Similarly, the regression on ΔLR2 did not reveal any significant predictor (p>0.351). The last regression on the decision parameter ζ revealed

ASD (β=-2.979, t=-2.308, p=0.028) and ΔHEPatt as significant predictors (β=11.125, t=3.793, p<0.001), that is, individuals with lower autistic traits and increased HEP modulation during overt interoceptive attention responded in a way more consistent with their trial-by-trial probabilistic beliefs.

4.4. Discussion

Empathy is essential for meaningful social interactions, enabling individuals to understand and respond to the needs and emotions of others. However, real-world social interactions are complex and dynamic, posing challenges for adaptive empathic responses. Given interoception's central role in shaping the current emotional experience (Craig, 2009; Seth, 2013), recent evidence and PP accounts suggest that it also plays a key role in guiding social behaviour (Allen et al., 2020; Ondobaka et al., 2017; Shah, et al., 2017; Stoica & Depue, 2020). This study examined how interoceptive processing underpins empathic learning and decision-making in a volatile social context by assessing whether HEPs, a neural marker of interoceptive processing, are modulated by social predictions during an empathy-related reinforcement learning task.

At each trial, participants had to select between two emotion regulation (ER) strategies to alleviate the distress of a virtual character. Unbeknownst to them, the target's preferred ER strategy fluctuated across the task according to a predefined probability schedule with various levels of uncertainty to test adaptive learning. Behavioural data were modelled using the HGF to extract individual learning trajectories and related parameters. Our findings reveal that HEP amplitudes were significantly differentiated by the magnitude of absolute pwPEs during the feedback phase, particularly over frontal and centro-parietal electrodes, probably indicating sensitivity of interoceptive predictions to social feedback. In addition, HEPs during decision-making were found to predict empathic response accuracy. Individual differences in autistic traits and empathic concern were associated with reduced differentiation of feedback-related HEPs as a function of pwPE fluctuations, suggesting less adaptive interoceptive processing. Interestingly, neural indices of cortical processing in a separate interoceptive attention task correlated with learning rates in the empathic decision-making task, further highlighting the importance of interoceptive processing modulation for adjusting empathic responses under uncertainty.

Adaptive empathy

In this study, we developed an empathic learning task where participants had to infer the preferred emotion regulation strategy of a virtual character across various distressing situations, based on feedback. Our model comparison showed that the HGF with three representation levels outperformed alternative models, namely, a two-level HGF and the Rescorla-Wagner model. This suggests that participants accounted for environmental volatility during learning, adjusting their learning rate in response to volatility fluctuations, which critically involves implicit differentiation between expected an unexpected uncertainty (Soltani & Izquierdo, 2019; Yu & Dayan, 2005). These findings indicate that individuals not only learn from social feedback to improve empathic responses, as shown in previous research (Kozakevich Ardel et al., 2021) but also optimize their responses by tracking changes in the statistical regularities of the effectiveness of emotion regulation strategies over time. Interestingly, part of this learning may have occurred implicitly, i.e. without participants explicit awareness that they had to learn preferences over two different types of ER (as the debriefing answers also suggest). Additionally, the observed association between learning rates and affective empathy dimensions suggests that this social learning task taps into underlying empathic traits and affective processes. Whereas in Kozakevich Ardel and colleagues (2021) study learning performance was correlated only with cognitive empathy, our suggest that empathic concern dispositions are linked to greater attunement to evolving social contingencies.

HEPs in AET

Even though interoceptive (predictive) processing has long been considered critical in affective, cognitive, and social processes, including learning and decision-making (Damasio, 1994; Gu & FitzGerald, 2014; Seth & Friston, 2016), only recently studies have started to provide empirical evidence for the involvement of interoception on predictive cognition and behaviour. Relevant research has mainly examined how exteroceptive and interoceptive expectations are intertwined measuring HEPs (Gentsch et al., 2018; Pfeiffer & De Lucia, 2017), with only one study (Fournagan et al., 2024), so far, linking HEPs to PE during reinforcement learning. Here, using a hierarchical Bayesian learning model, we found that social predictions were reflected in HEP fluctuations during the adaptive empathy task, specifically at the feedback phase. Cluster-based permutation analyses revealed that HEP amplitudes were significantly associated with pwPEs over fronto-central regions, with opposite polarity between the left and right clusters. Interestingly, HEPs did not differ significantly between feedback outcomes (correct vs. incorrect), suggesting that HEPs are more closely linked to the violation of one's socio-affective predictions than to simple outcome valence. That is, these results suggest that during uncertain social interactions we may attend more to our interoceptive sensations in response to surprising outcomes than to the negative outcomes of our actions. This is in line with Fouragnan and colleagues' (2024) study, who found that while HEPs were modulated by PEs during reward feedback in a non-social task they were not associated with correct vs incorrect outcomes. However, in that study researchers used a learning model with fixed learning rate, and thus, they could not assess how fluctuations in uncertainty can influence the evaluation of PEs. Our use of a hierarchical Bayesian learning model allows modelling adaptive learning in response to volatility changes via precision adjustments of PEs and learning rates. In more uncertain environments, assigning higher precision and salience to relevant interoceptive signals may help individuals to adapt more rapidly to changing social contingencies. By linking pwPEs during social feedback to HEP modulation, we present new evidence on how interoceptive prediction mechanisms extend to complex social learning tasks, where the brain must navigate changing social demands, while tracking changes in environment's volatility.

We also found that HEPs in the decision phase predicted response accuracy, highlighting the significance of interoceptive processing in decision-making process and extending previous studies showing that high interoceptive accuracy is associated with better intuitive decision-making (Dunn et al., 2010; Kandasamy et al., 2016; Katkin et al, 2001). These brain-body interactions could be linked to arousal effects in decision-making under uncertainty (FeldmanHall et al., 2016; Morgado et al., 2015; Wichary et al., 2016), as HEP is directly associated with fluctuations in autonomic arousal (Coll et al., 2021; Luft & Bhattacharya, 2015). Another explanation could be related to effective downregulation of interoceptive precision during action execution as active inference accounts suggest (Seth, 2013; Seth & Friston, 2016).

Besides cardiac interoception, recent fMRI evidence also indicates that several brain areas are linked to interoceptive predictions, with the anterior insula, in particular, being associated with both reported prediction certainty and PE in a breathing learning task (Harrisson et al., 2021). Notably, the anterior insula, is known to be a primary source of HEPs (Park & Blanke, 2019; Seeley, 2019) and to be involved in uncertainty estimation and multimodal integration (Allen, 2020; Graig, 2009, Preuschoff et al., 2008). Specifically, interoceptive PP models suggest that the insula, through its involvement in generating high order multimodal predictions, prioritizes the processing of incoming signals (exteroceptive and interoceptive) by optimizing their precision and salience based on their relevance to social and affective needs, with the overarching goal of minimizing uncertainty over time (Gu et al., 2013; Singer et al., 2009; Seth & Friston, 2016). In empathic interactions, such optimizations of precision are critical in various ways. Minimizing interoceptive PE with accurate interpretation and contextualization of bodily signals can be key to understand the socio-affective context, perceive subtle social cues, attend to other's needs and being able to engage in the complex process of interpersonal allostatic regulation (Bolis et al., 2023). Even if we did not perform source localization analyses in the present study, we argue that it is possible that the anterior insula was involved in in the generation of context-sensitive interoceptive predictions during volatile social interactions that helped to guide behaviour (Gu et al., 2013; Seth & Friston, 2016).

HEP and attention

Empirical findings and theoretical accounts of PP suggest that attention is a fundamental computational brain mechanism to optimize precision-weighting. Previous studies have considered precision-related modulations of interoceptive signals, particularly HEPs, in relation to overt attention (Banellis & Cruse, 2020; Petzschner et al., 2019) and individual differences in interoceptive accuracy (Ainley et al., 2016). For example, Petzschner and colleagues (2019), using the HB Attention task, detected attention-related HEP differences over right central electrodes within a time window of 524-620 ms. Our study reports similar HEP modulation patterns, identifying two significant clusters—one early and one late—over fronto-central electrodes, but without right lateralization.

These attentional modulations of neural markers of interoceptive processing are thought to reflect precision adjustments through changes in post-synaptic gain of superficial pyramidal cells, which encode PE (Ainely et al., 2016; Feldman & Friston, 2010). This neurophysiological activity can be detected by EEG, producing the observed HEP effects. However, our analysis of interindividual variability in ΔHEPatt (the difference between HEPs across the two attention conditions) showed no significant association with the variability in ΔHEPpe, where ΔHEPatt can be considered an index of intentional interoceptive attention regulation. This result suggests relatively independent precision adjustment processes between overt, intentional attention and covert attention driven by uncertainty estimates.

Individual differences in interoceptive processing and empathic learning

These distinct quantifications of interoceptive attention and precision offer insights into how interoceptive predictive processes relate to individual differences in adaptive learning and decision-making. Specifically, our results showed that ΔHEPatt was negatively associated with learning rate but positively correlated with the decision-making parameter ζ. This suggests that while the ability to intentionally modulate of interoceptive salience may interfere with implicit learning and social inference it may help to align individuals' implicit probabilistic beliefs with behavioural choices. Different aspects of social interaction might be distinctly influenced by ability to modulate interoceptive attention.

Regarding the individual differences in feedback-related HEP modulation, we found that higher EC was linked to better learning of empathic preferences. Autistic traits, while not predicting learning parameters, were associated with reduced incorporation of learned social probabilities during decision-making. Moreover, limited modulation of interoceptive markers of surprise (ΔHEPpe) during empathic learning was associated with higher autistic traits. Individuals with autistic traits may exhibit reduced fluctuations in interoceptive precision, potentially due to rigid social and bodily priors, overestimation of volatility, limited adaptation to new contexts and impaired interoceptive integration (Lawson et al, 2017; Palmer et al., 2017; Proff et al., 2024; Quattrocki & Friston, 2014), factors that are significantly intertwined. The dissociation between interoceptive and exteroceptive domains is supported by findings of atypical multisensory integration (including interoceptive-exteroceptive integration) in autism (Meilleur et al., 2020; Noel et al., 2018; Proof et al., 2022), as well as neurophysiological evidence showing altered connectivity or activity in the insula (Di Martino et al., 2009; Ebisch et al., 2011; Green et al., 2016). These impairments which can lead to inflexible interoceptive precision and inability to downregulate PEs have been suggested to underly the autistic psychopathology, which can be more pronounced in volatile social environments, requiring flexible adjustments for interpersonal emotion regulation (Sandhu et al., 2023).

Finally, the positive association between personal distress and increased belief in social volatility (ω3) aligns with previous findings showing a link between trait anxiety and higher estimates of environmental instability in a non-social task (de Becker et al, 2016). Trait anxiety has been associated with persistently high interoceptive PEs and maladaptive interoceptive precision (Paulus and Stein, 2010). As a result, dysregulated attention to bodily signals may interfere with optimal exteroceptive processing (Marshall et al., 2020; Rent et al., 2024). Therefore, minimizing interoceptive PEs through accurate predictions or flexible precision adjustments may be critical for the optimal processing of social information.

Limitations and future steps

One main limitation of this study is the absence of a non-social control task that would allow comparing how interoceptive processing relates to learning under uncertainty in conditions with and without emotional and social implications. However, previous research using such a control task has shown a dissociation between behaviour in social and non-social conditions and a relationship between trait cognitive empathy and learning in this social task (Arbel et al, 2021). Another limitation concerns the absence of analysis controlling for the potential electrical artifacts that cardiac activity may induce on HEPs (Park, Correia, Ducorps, & Tallon-Baudry, 2014). However, not only the procedures adopted here (i.e. artifact rejection) are now standard in the field (ref) as our main results were observed on later time windows of the HEP (i.e. >450ms after the R-peak), thought to be less susceptible to cardiac field artifacts. Additionally, the study could benefit from more trials per condition to improve the statistical reliability of HEP measurements. While having 35-40 trials per condition is considered adequate for low signal-to-noise ratio ERP studies, increasing the number of trials would better the quality of HEP data, as HEP is a subtle signal, usually accompanied by high noise. Similarly, the power in the regression analyses examining interindividual differences could be improved, as the variability in individual traits may require a sample size larger than 40 participants. Yet, previous studies with similar designs have used sample sizes comparable to ours (de Becker et al., 2016; Sevgi et al., 2020). Lastly, even though social interactions including empathic approaches and emotion regulation often have high levels of ambiguity and uncertainty the repetitive presentation of different distress scenarios and changing emotion regulation preferences over the duration of the study is something that deviates from real-life interactions. Although we attempted to address this issue with a cover story, it was difficult to fully mimic a naturalistic social environment. Despite managing to render most part of the learning procedure implicit and thus more likely to engage intuitive and affective subconscious mechanisms, the absence of a non-social task makes it difficult to confidently draw conclusions regarding the specific nature and the mechanisms underlying the learning process.

Conclusion

In summary, our findings offer insights into how interoceptive processing relate to learning and decision-making processes during dynamic empathic interactions. By demonstrating that HEP fluctuations track precision-weighted prediction errors in response to social feedback, we shed light into the underlying mechanisms of interoceptive predictive processing in volatile social environments. Such interoceptive modulations are not just sensitive to violations of expected outcomes but also track contextual changes related to volatility fluctuations. Our findings also suggest that these computational mechanisms of the brain for fine-tuning bodily reactions based on higher order social predictions are likely separate from overt attentional regulation ones. Even though HEP modulations during the decision-making process were not linked to learning rates, they predicted response accuracy, further highlighting how interoceptive precision modulations facilitate decision making and learning. The associations we observed between empathic and autistic traits and the indices of interoceptive integration indicate (potentially distinct) adaptive mechanisms regarding one's sensitivity to internal body signals when facing social uncertainty. While the absence of a non-social control and certain design limitations prevent us from drawing more comprehensive conclusions, our study provides a valuable starting point for future work. Investigating interoceptive precision adjustments across different social contexts could inform the development of tailored approaches to address empathy-related difficulties and presents a promising direction for further exploration in the emerging field of computational psychiatry. Overall, our findings contribute to the understanding of brain's adaptive interoceptive mechanisms within volatile social settings, enriching a view of empathy as an embodied, dynamic and context-sensitive process.

Chapter 5. Transcutaneous auricular vagus nerve stimulation modulates adaptive empathy

5.1. Introduction

Several bodily functions, like respiration and heartbeat are controlled by the vagus nerve, i.e., the tenth cranial nerve, which is a major part of the parasympathetic nervous system. The vagus nerve is comprised by 25% efferent and 75% afferent fibres being key for the bidirectional communication between the brain and the body (Berthoud and Neuhuber, 2000). The vagal tone, i.e. the operational balance of activity between the parasympathetic and sympathetic nervous system through the vagus nerve, has been suggested to control, apart from bodily functions, a large range of adaptive social behaviours (Damasio et al., 1991; Proges, 2007). For example, it influences mammal's physiological states to facilitate either 'fight or flight' responses (with decreased vagal input) or social engagement (with increased vagal input). Being the largest somatic nerve, it conveys information from major internal organs to the brainstem, a central interoceptive hub (Critchley & Harrison, 2013), and from there to several other areas including the insula, hippocampus and locus coeruleus, while also induces changes to several neurotransmitters, including serotonin and noradrenaline (Leusden et al., 2015). These brain systems have been associated with social attention, learning, emotion processing, empathy and interoception (Critchley & Harrison, 2013; Critchley & Garfinkel, 2017). While initially direct current stimulation was applied subcutaneously directly into the vagus nerve, in 2001 a non-invasive stimulation of the auricular branch of the vagus nerve (taVNS) was developed and widely used up to date (Yap et al., 2020)). taVNS is applied on the left afferent branch of the vagus, at low intensity and thus it does not interfere with the efferent parasympathetic signals to the heart. Due to its neuromodulatory effects, vagus nerve stimulation (VNS) has been used as a therapeutic tool for conditions such as epilepsy and depression, with potential applications for neurodevelopmental and emotion regulation disorders (Beekwilder & Beems, 2010; Engineer et al., 2017).

Vagus nerve stimulation has been recently suggested as a promising tool to experimentally manipulate the communication between the body and the brain, thereby directly affecting interoceptive processing (Pacioroek & Skora, 2020). Indeed, a recent study showed that taVNS can affect the awareness of cardiac signals as it improved performance in the heartbeat discrimination task, a measure of interoceptive accuracy, that is, the ability to objectively perceive one's own internal bodily signals (Villani et al., 2019). Several more recent studies also come in support of this hypothesis (Richter et al., 2021; Poppa et al., 2022; Ventura-Bort & Weymar, 2024). For instance, Poppa and colleagues' (2022) study provided direct evidence on the link between taVNS and cardiac interoception by showing that taVNS modulated the Heart-Evoked-Potential (HEP), an effect localized to the insula, operculum, somatosensory cortex, and orbital and ventromedial prefrontal regions (Poppa et al., 2022). Importantly, the HEP is an EEG signal time-locked to the heartbeat, thought to reflect cortical processing of heartbeats, and associated with interoceptive accuracy (Mai et al, 2018), empathic sensitivity (Fukushima et al., 2011), affective predictions (Gentsch et al., 2018) and belief updating of cardiac dynamics (Ainley et al., 2016; Pfeiffer & DeLucia, 2017). Together these studies suggest that taVNS can be a useful tool for the manipulation of interoceptive processing and its impact on affective and social behaviour.

Indeed, the vagus nerve, and especially the more recently developed ventral branch in mammals, has been postulated by the polyvagal theory to play a central role in social engagement (Porges, 2007). Several studies accord with such hypothesis, as vagal activity, estimated through respiratory sinus arrhythmia (RSA) and heart rate variability (HRV), has been associated with pro-social traits (Kogan et al., 2014), cooperation (Beffara et al., 2016), and altruistic behaviour (Bornemann et al., 2016). More recent evidence also supports a causal role for the vagus nerve in social cognition as it has been shown that taVNS can improve emotion recognition of other's faces and bodies (Colzato et al., 2017; Sellaro et al., 2018), and modulate attention allocation to salient social cues (Maraver et al., 2020). Our bodily physiological states and the perception of them (that is, interoception) have been shown to influence the processing of social information, affecting emotion recognition (Hubner et al., 2021), empathy (Grynberg & Pollatos, 2015) and social behaviour (Ambrosecchia et al., 2017). Importantly, interoceptive signal have been suggested to be critical in learning and decision-making, especially under conditions of uncertainty (Dunn et al., 2010, Katkin et al, 2001). As most social situations come with high levels of complexity and uncertainty, listening accurately and adaptively to your bodily signals is crucial for navigating social interactions. Given the vagus nerve's proposed role in social behaviour (Porges, 2007), and the importance of interoceptive predictions in socio-affective processing, manipulating afferent interoceptive activity may impact key processes of social learning and decision-making. Yet, to the best of our knowledge, no study thus far tested whether taVNS can influence social adaptive learning.

Regarding the impact of VNS on learning, while there is significant evidence for its influence on memory and learning in non-human animals (Clark et al., 1998, 1999; Driskill et al., 2022; Pena et al., 2014), there are limited and mixed findings regarding the VNS effects on reinforcement learning in humans. Specifically, despite research in animals demonstrating a clear positive effect on reward learning (Bowles et al., 2022; Han et al., 2018), two recent studies in humans provide opposing results (Kühnel et al., 2020; Weber et al., 2021). In line with previous findings in non-human animals, Weber and colleagues (2021) demonstrated that direct VNS in epilepsy patients improved performance in a reward-based forced-choice learning task. Conversely, Kühnel and colleagues (2020), using a computational approach to model the behaviour in a reinforcement learning go/no-go task, found that accuracy and learning rate decreased in the active taVNS compared to the sham condition. Despite both studies showing that VNS influences reinforcement learning, it is still unclear how associative reinforcement learning is affected by VNS. Moreover, given that VNS has been found to have neuromodulatory effects on the dopaminergic (Han et al., 2018), noradrenergic (Roosevelt et al., 2006) and cholinergic (Bowles et al., 2022) systems, where each has distinct effects on associative learning under uncertainty (Marshall et al., 2018), we do not yet know how VNS can impact social reinforcement learning under different conditions of uncertainty.

In a previous study, we examined how brain-body dynamics can be linked to adaptive empathic behaviour by investigating the modulation of HEPs during a reinforcement learning task. Findings showed prediction errors of empathic preferences to be linked to HEP modulations during the feedback phase, suggesting cardiac dynamics to be involved in social interoceptive inference. However, it has been proven notoriously difficult to interfere with these brain-body dynamics and thereby examine the related causal influence of autonomic afferents in social perception and behaviour. Recent evidence on the cognitive and social effects of taVNS make this technique a promising and safe tool to manipulate interoceptive afferents and autonomic activity. Thus, capitalizing on the role of vagus nerve and interoception on associative learning and social cognition, we examined whether and how taVNS can improve empathic learning and decision-making under conditions of uncertainty. Specifically, we used a probabilistic empathic reinforcement learning task while applying active vs sham taVNS. In this adaptive empathy task, participants had to learn, over the course of several trials, the preferred emotion regulation strategy (between reappraisal and distraction) of a virtual character. Importantly, this preference was not stable and changed over time under different volatility phases, i.e. changing every few trials (volatile block) or after longer periods of stable preference (stable block). The behavioural data was modelled using a hierarchical Bayesian learning model, i.e. the Hierarchical Gaussian Filter, where social beliefs and the related uncertainty are represented at 3 different levels (Mathys et al., 2011; 2014). This modelling approach provides individual learning parameters and adaptive learning rates for each representation level, indices of prediction error (PE), as well as response parameters that additionally capture differences in decision making. Our main hypothesis was that learning rates, especially of the second and third level, and overall performance would be improved with the application of taVNS.

5.2. Methods

5.2.1. Participants

Our participants were psychology students from the University of Kent. In total, 68 participants were included in our analyses [aged 18-35, (M=20.5, SD=2.5), 47 females]. With this sample size, we can detect medium effect sizes (Cohen's d=0.5) with 80% power and significance level of a=0.05. From the initial 70 participants recruited, two were excluded due to technical issues or due to failure to pass attention checks. Their participation was rewarded with credits in the research participation scheme of the University of Kent. All participants provided written informed consent before the beginning of the experiment. The study was approved by the School of Psychology University of Kent Ethics Committee and all safety requirements were ensured. TaVNS is a non-invasive technique that posits no major risk to the participants but to minimize any potential implications. Inclusion criteria required: 1) no history of neurological disorders, 2) no history of brain surgery, tumour, or intracranial metal implantation, 3) no known cardiovascular abnormalities, 4) no pregnancy, 5) no susceptibility to seizures or migraine, and 6) no pacemaker or other implanted devices; 7) no history of syncope; 8) no particularly irritable/sensitive skin. Before the beginning of the experimental procedure, participants were informed regarding possible adverse effects (i.e. itch, minor skin irritation and mild pain at the point of stimulation that typically disappears a few minutes after stimulation offset).

5.2.2. Questionnaires

Participants completed 5 online questionnaires prior to the empathy task in the lab session. The questionnaires were the same used in EEG study reported in Chapter 4, including self-related responses to the distress scenarios, the AQ (Baron-Cohen et al., 2001), the IRI (Davis, 1980), the DERS (Dan-Glauser & Scherer, 2012). We also added the MAIA to collect participants subjective beliefs regarding the processing of and attention to their own bodily sensations. Detailed descriptions of the first 4 questionnaires can be found in Chapter 4 and for MAIA in Chapter 2.

5.2.3. Adaptive empathy task

After completing the questionnaires at home, participants had to perform the AET as described in Chapter 4, but instead of the use of EEG, this time we applied sham or active taVNS. TaVNS was applied in a between-subjects single-blind design, thus participants were randomly assigned to the sham or active stimulation group. The order of the ER preferences in the task was counter-balanced between participants (i.e. half of the participants starting with reappraisal being the preferred regulation response for the virtual character, and for the other half of them, distraction was the correct response at ~70% of times). Stimulus presentation and design was the same as described in Chapter xx. Overall, the empathy task lasted around 35 minutes with two breaks. After the end of the procedure, participants were asked to answer a few debriefing questions.

5.2.4. TAVNS set up and device

We used the Transcutaneous Electrical Nerve Stimulation device (V-TENS Plus; https://bodyclock.co.uk/) for stimulation with custom-built electrodes (Villani et al, 2019). For the active stimulation, the electrodes were placed on the anterior wall of the external ear canal where the tragus is. Sham stimulation was performed by placing the electrode on the left earlobe, as this area is free of vagal innervation (Peuker & Filler, 2002). TaNVS was always applied on the left ear for both groups due to the lateral organization of the vagus nerve. As the right brunch of the vagus nerve comprises both afferent and efferent fibres, taVNS on this side could induce cardiac side effects. Conversely, the left branch comprises only afferent fibres and thus is not directly involved in the regulation of cardiac activity. Clinical trials and meta-analysis confirm the absence of any arrhythmic effects of taVNS when applied on the left ear (Kim et al., 2022; Kreuser et al., 2012).

Here, we applied continuous current stimulation on the left ear with the following parameters for all participants: pulse width=250µs and frequency=25Hz. The intensity of the stimulation was adjusted across participants so that it remained just below their individual perceptual threshold. To this aim, the experimenter slowly increased the intensity level up to the level that the participant felt some sensation (e.g., tingling, itching), which was barely detected without causing any pain or discomfort. We then decreased the intensity and then increased again repeating the process 3 times to ensure the intensity stimulation level for each participant, i.e. just below perceptual awareness. The intensity was applied continuously at that specified level continuously throughout the task. The average intensity was 0.20mA (SD=0.08) for the active stimulation group and 0.19mA (SD=0.09) for the sham stimulation group.

5.2.5. Computational modelling

To model the behavioural responses from the AET we applied a hierarchical Bayesian inference approach and specifically we used a 3-level HGF model. The structure and parameters of the model are described in detailed in Chapter xx. With this approach after fitting the data of each participant we obtained the individual learning and decision-making parameters of interest. Specifically, to examine the effect of taVNS mainly used the parameters: learning rate in the second (LR2) and third (LR3) level; the difference between the learning rates for the volatile and stable period, for both the second (ΔLR2) and third level (ΔLR3); precision-weighted predictive errors (pwPEs) for both second and third level; as well as the ζ parameter. The learning parameter LR2 of the second level captures individual differences regarding how fast ER preferences of the target are updated from trial to trial. The learning rate at the third level captures the magnitude of the trialwise updates on the volatility learning rate. ΔLR2 and ΔLR3 reflect the adaptation of each learning rate in response to volatility changes (Lawson et al., 2017). Following the results of the previous chapter on pwPEs and social predictions and individual differences in interoceptive processing as well as social and autistic traits, we examined how pwPEs are modulated by taVNS and individual traits. Lastly, the parameter ζ of the response model captures individual differences in the use of formed beliefs during decision-making, where higher ζ values denote a more deterministic response behaviour, or in other words more consistent use of the updates in ER beliefs for their responses.

5.2.6. Statistical analysis

Using the model parameters and estimates extracted, we conducted a series of statistical analysis using the R software. We ran a 2x2 mixed ANOVA with the mean learning rate for the second and third level LR (LR2, LR3) as the within-subjects factor and taVNS group (active, sham) as the between-subjects factor. Also, to measure possible taVNS influence on learning adaptability to changes in volatility, we estimated the difference between mean learning rate in the volatile and stable period, for both the second and third level, and ran an 2x2 mixed ANOVA with ΔLR (ΔLR2, ΔLR3) as the within-subjects factor and taVNS as between-subjects factor. In addition, we run an independent samples t-test to compare the response parameter ζ between the sham and active stimulation group.

Moreover, to test how individual differences affect task performance and potentially moderate taVNS effects, we ran different linear regressions with the following DVs: the mean learning rate of the second (LR2) and third (LR3) HGF level, the learning rate difference in the second (ΔLR2) and third level (ΔLR3) between the stable and volatile phase, the mean pwPE in the stable and volatile phase of the second level of the HGF model. The predictors were the questionnaires scores and their interaction with stimulation group. Specifically, following the analysis in the previous chapters, as predictors we used the affective dimensions of the IRI (i.e. empathic concern and personal distress), and the total AQ and MAIA score and the emotion regulation preferences (ERP) as a covariate. Finally, taVNS group was entered as an interaction term. Thus, the general formula of these regressions is the following:

DV ~ (IRI_EC + IRI_PD + AQ + MAIA) * taVNS + ERP

5.3. Results

TaVNS effects on accuracy, and learning and response model variables

First, the 2x2 mixed ANOVA on accuracy revealed an expected main effect of volatility (F(1,67)=49.250, p<0.001, η^2p =0.424), with higher performance in the stable phase (M=57.3, SD=8.12) compared to the volatile phase (M=49.5, SD=7.30). Conversely, no main effect of taVNS (F(1,67)=0.405, p=0.527, η^2p =0.006), or interaction effect between taVNS and volatility was revealed (F(1,67)=0.541, p=0.465, η^2p =0.008). This shows no general taVNS effect on learning performance in terms of accuracy.

To further analyse how participants' learning was affected by taVNS, we entered the mean learning rate of the second and third levels (LR2, LR3) in a 2x2 mixed ANOVA. The results showed a significant main effect of LR type (F(1,67)=29.625, p<0.001, η^2p =0.313) but no significant main effect of taVNS (F(1,67)=1.661, p=0.202, η^2p =0.026) nor interaction effect between LR type and taVNS (F(1,67)=0.058, p=0.942, η^2p =0.001). Therefore, no effect of taVNS on mean learning rates was revealed.

We also analysed how participants modulated their learning rate according to environmental volatility by entering the learning rate difference (ΔLR2, ΔLR3) between the stable and volatile phase parameters in a 2x2 mixed ANOVA. The ANOVA showed a significant main effect of DL type (F(1,67)=18.957, p<0.001, η^2p =0.221) and a significant main effect of taVNS (F(1,67)=4.686, p=0.034, η^2p =0.065). Importantly, we also found an interaction effect between ΔLR type and taVNS (F(1,67)=4.817, p=0.032, η^2p =0.067). Pairwise comparisons (with Bonferroni corrections) showed that ΔLR2 was significantly higher for the active stimulation group (M=0.209, SD=0.357) compared to the sham stimulation group (M=0.073, SD=0.076; t(1,66.695)=2.387, p=0.037, Cohen’s d=0.544; fig. 2). Conversely, ΔLR3 did not differ between the active and sham stimulation groups (t(1,66.804)=-0.2073, p=0.836, Cohen's d=0.049). These results suggest that the taVNS effect on learning pertains mainly to the learning of the second-level contingencies (with a moderate to large effect size) by helping participants adapt to the virtual character's changes in ER preferences.

Lastly, we examined the effect of taVNS on pwPEs of the second level in the stable and volatile phases in a 2x2 mixed ANOVA. The results showed a significant main effect of phase (F(1,67)=9.145, p=0.004, η^2p =0.120) as pwPEs were higher in the volatile phase (M=0.026, SD=0.021) compared to the stable phase (M=0.016, SD=0.028). However, we did not find a significant main effect of taVNS (F(1,67)=0.006, p=0.937, η^2p =0.001) nor a pwPE phase x taVNS interaction (F(1,67)=0.342, p=0.560, η^2p =0.005).

Regarding the response model results, we found that there was no difference between the active and sham stimulation for the model parameter ζ (t(1,66.339)=0.455, p=0.649, Cohen's d=0.057). Therefore, taVNS did not have any observable effect on decision-making.

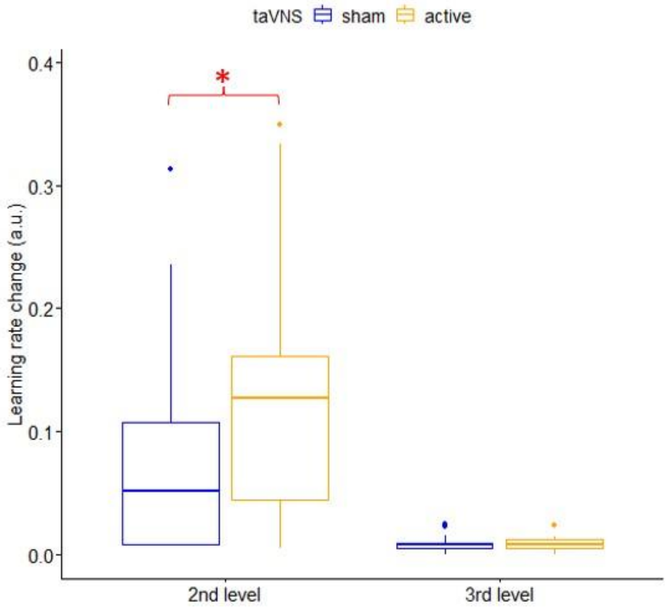


Figure 5.1. Learning rate change across representation levels. Participants in the active stimulation group had increased mean modulation of the learning of ER preferences (ΔLR2) due to the volatility changes compared to the sham group (p<0.037). Conversely, no difference was found between the two stimulation groups for the modulation of the volatility learning (ΔLR3).

Individual differences

To examine how the learning parameters of the HGF model were related to individual traits, we ran linear regressions on the learning parameters presented above, with the taVNS group and the individual traits measured in the questionnaires as predictors. Prior to this, we conducted a series of t-tests to examine initial differences between the participants in the two stimulation groups for the traits measured. This analysis showed no significant differences in any of these traits (ps > 0.493).

First, regarding learning rates, we found that the mean learning rate in the second level was the only one predicted by empathic concern (t = 2.390, p = 0.029), while all other predictors did not reach significance (ps > 0.05). Specifically, participants with higher empathic concern learned the ER contingencies faster. Conversely, individual traits did not influence the general update of changes in volatility, as the mean learning rate on the third level was not predicted by any questionnaires.

The regression on the learning rate difference ΔLR2 showed that it was predicted by taVNS activation (t = 2.423, p = 0.021), confirming our ANOVA results. All other predictors did not reach significance (ps > 0.05), as only empathic concern approached significance (t = 1.984, p = 0.053). Interestingly, and in contrast to the results observed in the ANOVA, the volatility learning rate difference ΔLR3 was significantly predicted by taVNS activation (t = 3.494, p = 0.001), showing a higher change in volatility learning rate in response to volatility changes in the active group compared to the sham group. This suggests an effect of taVNS in updating beliefs about volatility only when accounting for individual differences.

This analysis further revealed AQ as significant predictor (t = 3.620, p = 0.001), with people scoring higher on the AQ scale also having a higher update of volatility learning rate between the volatile and stable phases. The interaction between AQ score and taVNS was also significant (t = -3.636, p = 0.001), as in the sham group we observed that participants with a high AQ score had a more adaptive volatility learning rate, while in the active stimulation group this correlation was absent. Moreover, the same pattern was observed with the MAIA score, as the interaction with taVNS appeared as a significant predictor (t = -2.060, p = 0.047), with MAIA scores being positively correlated to ΔLR3 in the sham group but showing no correlation in the active group. This result similarly indicates that individuals with higher interoceptive awareness had a greater update of the volatility learning rate, an effect that was minimized with the application of taVNS.

Regarding the prediction error regressions (pwPEs) at the second level of HGF, for the stable phase of the task, we found the MAIA score to be a significant predictor (t = -2.513, p = 0.017), as people with higher interoceptive awareness had lower pwPEs. The MAIA x taVNS interaction was also a significant predictor (t = 2.155, p = 0.038), as this negative association was mainly observed in the sham stimulation group, while in the active group the correlation was absent. Lastly, in the volatile phase, we found empathic concern to be a significant predictor (t = 2.683, p = 0.011), where individuals with higher empathic concern also had a higher mean pwPE. In addition, the Empathic concern x taVNS interaction (t = -2.477, p = 0.018) was also significant, as this positive correlation observed in the sham group was diminished in the active stimulation group.

Regression Model	Predictor	t-value	p-value	Interpretation
LR2	Empathic Concern	2.390	0.029	Higher EC → Faster learning of ER preferences
ΔLR3	MAIA x taVNS	-2.060	0.047	IS positively predicted ΔLR3 mainly in the sham group
	Autism	3.620	0.001	Higher AQ → Higher volatility learning rate change (ΔLR3)
	Autism x taVNS	-3.636	0.001	AQ correlated with ΔLR3 more strongly in the sham group
pwPW (stable phase)	MAIA	-2.513	0.017	Higher IS → Lower pwPEs (Stable Phase)
	MAIA x taVNS	2.155	0.038	IS effect mainly in sham group
pwPW (volatile phase)	Empathic Concern	2.683	0.011	Higher EC → Higher pwPEs (volatile phase)
	Empathic Concern x taVNS	-2.477	0.018	EC effect mainly in sham group

Table 5.1. Regression Results on Individual Differences. This table presents the significant regression results examining the relationship between individual traits (e.g., interoceptive sensibility, empathic concern, autistic traits) and learning trajectories. Predictors include questionnaire scores and their interactions with taVNS stimulation (active vs. sham). LR2 = Second-level learning rate; ΔLR2, ΔLR3 = Learning rate differences between stable and volatile phases at the second and third hierarchical levels, respectively; pwPEs = Precision-weighted prediction errors. Significant interaction effects indicate that the relationship between individual traits and learning parameters differs between the active and sham taVNS groups.

5.4. Discussion

Afferent interoceptive signals have been suggested to facilitate social cognition (Grynberg & Pollatos, 2015; Ondobaka et al., 2017; Shah et al., 2017) and decision making under uncertainty (Dunn et al., 2010; Katkin et al, 2001; Kandasamy et al, 2016). Transcutaneous auricular vagus nerve stimulation (taVNS) has emerged as a promising tool to intervene in these brain-body pathways (Paciorek & Skora, 2020; Weng et al., 2021), demonstrating beneficial impact in a variety of clinical and non-clinical applications (Johnson & Wilson, 2018; Yap et al., 2020). Our study demonstrates that taVNS improved adaptive social behaviour in an empathy task where participants had to (implicitly) learn the emotion regulation preferences of a virtual character which were either stable or volatile. To examine learning under different uncertainty conditions we modelled our behavioural data using a Bayesian model for hierarchically coupled representation levels (Mathys et al., 2014). While taVNS did not seem to influence overall accuracy and learning rates in this task, results showed that active taVNS was associated with greater updating of the learning rate regarding the ER preferences of the target. This effect is primarily associated with calculations on the second (estimation uncertainty) representation level. Moreover, we showed that the taVNS modulates the relationship between individual differences in empathic concern and interoceptive awareness and social learning. These results extend previous findings suggesting a positive impact of vagus stimulation on associative learning and social perception (Colzato et al., 2017; Weber et al., 2021) by showing, for the first time, how taVNS can affect social learning under volatility conditions. These effects are likely explained by the neuromodulatory role that taVNS has on the noradrenergic and dopaminergic systems and on its effect on interoceptive processing.

While several studies on animals have demonstrated a positive impact of vagal stimulation on learning and memory (Clark et al., 1998, 1999; Driskill et al., 2022; Frausto Pena et al., 2014), research involving humans is scarce while yielding mixed results. For instance, Weber and colleagues (2021) found that taVNS improved reinforcement learning by increasing reward sensitivity and inducing a shift of accuracy-speed trade-offs towards maximizing rewards. Conversely, a study with healthy participants showed that taVNS impaired decision-making and reduced the learning rate in an approach-avoidance task (Kühnel et al, 2020). Another study in humans examining the effects of taVNS in fear extinction revealed no stimulation related effect (Burger et al., 2018), failing to replicate previous findings in rats (Alvarez-Dieppa et al., 2016). These mixed results highlight the need to understand the precise mechanisms underlying taVNS effects on learning. Which processes are affected? Under which conditions? In our study of associative learning in a social context, total accuracy scores and mean learning rates did not differ between sham and active simulation groups. Critically, though, we observed a greater modulation of the learning rate of the target's ER preferences between the high and low volatility blocks. This effect primarily pertains to the second level of the model representing the belief about the stability of the target's ER preferences. A similar taVNS effect was revealed on volatility learning rate changes (ΔLR3) when controlling for individual differences in our regression analysis, an effect not found in the related ANOVA though. Nonetheless, these two level are tightly linked, with the second-level learning rate being controlled by the belief about the stability of this volatility, reflected on the third level parameters. In other words, how fast the social environment patterns are expected to change influences the way we assess a prediction error in a social interaction. In more volatile environments, an unexpected error has more changes to signal a significant change and thus being evaluating according, and informing our expectations, thereby increasing the learning rate of social beliefs updating. In contrast, in environments with lower volatility, such errors could be attributed to noise, i.e. not relevant changes, thus leading to fewer updates of social beliefs.

Previous research suggests that vagus stimulation mainly impacts the activity of the noradrenergic (Roosevelt et al., 2006) and dopaminergic (Han et al., 2018) systems, which have distinct effects on learning under volatility. Noradrenaline is involved in signalling contextual changes and influences learning of uncertain events due to volatility changes while dopamine affects adaptive responses, i.e. the ability to adapt one's responses in face of unexpected events (Marshall et al., 2016; Lawson et al., 2021). Specifically, two studies using HGF modelling found that a noradrenaline antagonist modulated the learning rate at the second (Lawson et al., 2021) and third level (Marshall et al., 2016). Conversely, dopamine has been found to influence behavioural adaptability to environmental volatility (Marshall et al, 2016). In our study, the observed effect of taVNS on the modulation of the learning rate of ER contingencies due to volatility changes suggests a noradrenaline effect, making participants more adaptive to contextual changes (Angela & Dayan, 2005; Sales et al., 2019). Because we did not observe any general effect on learning performance or on the decision-making parameter of the response model, we speculate that taVNS did not significantly affect dopamine levels.

Even though, no other study has thus far investigated the role of taVNS on adaptive behaviour in a social context, our results are consistent with the polyvagal theory which postulates that vagal activity to be critical for social engagement and adaptive behaviour (Porges, 2007). Therefore, another possible explanation of our findings is that taVNS increased vagal tone made participants more relaxed and engaged, thereby more sensitive to changes in the social environment. These effects might have been also combined with enhanced brain-body communication and reduction of interoceptive PE as evidence in tVNS studies suggest (Villani et al, 2019; Poppa et al, 2022; Richter et al., 2020). Given that interoception is linked to empathic and emotion regulation abilities, these improvements in interoceptive processing might have increased the saliency of social cues. Supporting this, previous studies have shown that increased vagus activity is associated with enhanced processing of emotional and social cues (Colzato et al., 2017; Maraver et al., 2020), as well as with higher compassion, cooperation and pro-social traits (Kogan et al., 2014; Oehrn et al., 2022; Stellar et al., 2015). Additionally, it has been shown that taVNS can also improve cognitive emotion regulation (De Smet et al., 2021), with vagal tone being an index of emotional regulation (Pinna & Edwards, 2020). Therefore, we can postulate that improved self-regulation and interoceptive processing could have increased attunement to the emotion regulation needs of others. Indeed, a recent meta-analysis showed that better adaptive self-regulation is positively correlated to compassion as well as cognitive and affective empathy (Kampf et al., 2023).

Our regression analyses exploring how individual differences relate to adaptive empathic learning under volatility showed that taNVS increased the modulation of the learning rates between the stable and volatility phases (ΔLR2 and ΔLR3). We further found that empathic concern was positively correlated with mean second level learning rate (LR2) and the second learning rate difference (ΔLR2). Higher empathic concern was also associated with increased surprise, as indexed by pwPEs, in ER contingencies in the volatile phase, likely driven by higher learning rates in conditions of high volatility. Previous research has shown that empathic traits predict performance in this task, but not on a similar associative learning task devoid of social meaning, highlighting the usefulness of this paradigm to the study of adaptive social behaviour (Arber et al, 2021). We extend these findings by showing that empathic (concern) traits are associated with increased social learning in stable conditions and increased surprise signalling under volatile conditions. This suggests that people more attuned and concerned with other's feelings are better at learning others' emotion regulation preferences and more sensitive to violations of such expected preferences in conditions of high uncertainty, two abilities that seem key to flexible and adaptive empathic behaviour. On the other hand, the absence of significant associations with volatility learning suggests that this aspect of affective empathy primarily taps onto better processing of task specific socio-affective contingencies.

We also found that people with higher AQ scores exhibited higher update of the volatility learning rate (ΔLR3). This is consistent with a previous study revealing, using a non-social task, that individuals with autism tend to overlearn about volatility in the face of environmental change compared to neurotypicals (Lawson et al., 2017). These findings were attributed to a general overestimation of volatility in ASD, while computational pupillometric analysis revealed heightened trial-wise surprise and pwPEs, suggesting aberrant phasic noradrenaline activity. Interestingly, in our study we found that the correlation between autistic traits and ΔLR3 was mainly observed in the sham stimulation group and diminished in the active stimulation one. Thus, in support of proposals that vagus nerve stimulation could be a promising intervention tool to alleviate symptoms in neurodevelopmental disorders (Jin & Kong, 2016; Yap et al., 2020) by regulating noradrenaline activity. Autistic individuals is possible to show lower vagal tone (Ming et al., 2005), atypical locus coeruleus-noradrenergic activity (Chatham et al., 2022), difficulties navigating high-uncertainty situations (Jenkinson et al., 2020), as well as aberrant belief updating (Lawson et al., 2014; Van de Cruys et al., 2014)Taken together, taVNS could potentially help these autistic individuals to interact under volatile social scenarios by optimizing estimations of uncertainty.. However, it is not always clear what an optimal estimation of uncertainty might be, as this could be heavily contextually dependant. For example, psychopathology in autism has been considered not as a suboptimal estimation of uncertainty and belief updating per se, but rather as an interpersonal mismatch thereof (Bolis et al., 2017). Here, taVNS could be used as a tool of interpersonalised psychiatry (Bolis et al., 2023), with the aim of regulating such interpersonal mismatches in social interactions.

A similar pattern of taVNS interaction on ΔLR3 was observed with the MAIA score as a positive corelation observed in the sham stimulation group was absent in the active group.

This result also suggests that taNVS might regulate the autonomous nervous system in a way that minimizes how individual differences in somatic processing are reflected to social understanding. The higher ΔLR3 for individuals with higher interoceptive awareness is a result not clearly expected from previous literature. However, based on research linking interoceptive accuracy to intuitive learning and decision making (Dunn et al, 2010; Katkin et al, 2001, Kandasamy et al, 2016), we can hypothesize that not easily detectable contextual changes are more likely to be unconsciously detected by individuals with heightened interoceptive abilities. Although, we did not observe any relationship between MAIA score and learning rate parameters of the second level to support this speculation, we found that pwPEs at the second level were lower in the stable phase for participants with higher reported interoceptive awareness, probably indicating more optimal calculations of expected uncertainty under relatively stable conditions. This observed correlation disappeared with the application of taVNS, as in previous interactions. Interoceptive processes have been shown to be critical for intuitive decision making and learning under certainty (Dunn et al., 2010; Katkin et al, 2001; Kandasamy et al, 2016), while has also been suggested that the coherence between subjective feelings of arousal and actual physiological states underpins the updates of uncertainty representations in dynamic environments (Biddell et al., 2024). Thus, by facilitating interoceptive processing (Villani et al, 2019; Poppa et al, 2022), taVNS may also facilitate social interactions under uncertainty in various ways.

However, the interpretations of how taVNS may modulate the relationship between individual differences, as measured by questionnaires, and task performance, needs to be approached cautiously due to our limited sample size, which restricts our ability to draw confident conclusions. Moreover, although we found a taVNS effect on adaptive empathic learning, it was only observed in the rate of change? of our model parameters and not in other indices of overall performance. This could be explained by a specific stimulation-dependent effect related to the update of volatility representations. However, we cannot fully rule out the influence of other factors also affecting empathic learning. For instance, the learning of ER preferences might have been influenced by order in which the preferred ER strategies were presented, i.e. which was the first ER strategy preferred by the target. This issue can be addressed with a larger sample size to increase the statistical power to determine if and how learning is affected by the order of ER strategy preference and how this can be influenced by taVNS. Another limitation of our study is the absence of physiological measurement indexing vagal activity or neuromodulations changes due to taVNS. Thus, even though our findings, based on past research, suggest a stimulation-dependent effect on social learning, probably mediated via the noradrenergic system, we cannot draw such mechanistic conclusions with certainty. Future studies incorporating pupillometry and HRV could help identifying how these taVNS effects are related to neuro-physiological modulations. Lastly, we cannot be certain that the observed taVNS effects were social-specific or associated with improved learning in general. Vagus nerve stimulation has been shown to affect domain-general general associative learning and additionally improve social engagement and processing. (REF) Running a similar protocol with a non-social learning task could help clarify the specificity of taVNS effects.

Conclusion

In summary, this is the first study showing a relationship between vagus nerve stimulation and learning under volatility, particularly on a social context, with taVNS promoting adaptive empathetic learning. taVNS was also found to mediate how individual differences in autistic and interoceptive traits influence different aspects of empathetic learning. Our modelling approach highlights specific facets of uncertainty representations influenced by taVNS and suggests potential underlying mechanisms. However, while these results have promising implications for therapeutic applications in both clinical and subclinical populations, more research combining computational, behavioural and neuro-physiological methods, is needed to further elucidate on the underlying neurocognitive mechanisms.

Chapter 6. General Discussion

6.1. Overview

Understanding and navigating the social world is a dynamic process that requires continuous integration of external social cues and internal bodily states. Influenced by predictive processing (PP) frameworks, this thesis approaches social understanding as an active, embodied probabilistic inference process shaped by interoceptive processing, learned contingencies, and social uncertainty. Theoretical and computational models have highlighted how uncertainty plays a pivotal role in these integration processes, influencing how we predict, perceive and respond to social interactions to minimize surprise. However, several gaps remain in our understanding of how interoceptive processing within social contexts with different sources of uncertainty, ranging from perceptual ambiguity to environmental volatility, affects social cognition and adaptive decision-making. This thesis sought to address these gaps across four empirical chapters. Specifically, this research explored:

1. How interoception interacts with uncertainty intolerance (IU) and alexithymia, to influence anxiety and social perception.
2. How interpersonal emotion contingencies (IEC) influence social perception under varying conditions of social uncertainty.
3. How interoceptive signals underpin and affect social learning and adaptive empathy, particularly in volatile social environments.
4. How do individual differences in interoceptive sensibility, autistic, empathic traits and alexithymia modulate these processes.

The first two empirical chapters primarily examined how interoceptive inference affects emotion perception and affective empathy. Using path analysis and diffusion modelling, the studies of Chapter 2 demonstrated that IU and alexithymia mediate the relationship between interoceptive sensibility and anxiety, while also shaping emotion recognition and affective empathy. Chapter 3 explored further the role of contextual uncertainty on emotion recognition. Specifically, these studies demonstrated how fluctuations in IEC expectancy affect emotion recognition biases. Using HGF modelling they also showed how implicit learning of IEC is modulated by physiological arousal and individual differences in interoceptive sensibility and alexithymia. The final two empirical chapters examined how interoceptive processing underpins adaptive empathy and social learning under different uncertainty conditions. Chapter 4 demonstrated that heartbeat-evoked potentials (HEPs) track precision-weighted prediction errors (pwPEs) during volatile social feedback, with affective empathy and autistic traits modulating interoceptive integration and adaptive learning. Finally, Chapter 5 introduced transcutaneous vagus nerve stimulation (taVNS) as a neuromodulatory tool, revealing its capacity to enhance adaptive empathic learning.

This General Discussion aims to synthesize the key findings and situate them within broader theoretical frameworks. It will start by providing an integrative summary of how interoception and uncertainty jointly influence social cognition across the four studies. Then, it will theoretically anchor these results in PP and social cognition models, highlighting the implications for emotion perception, empathy and (mal)adaptive learning. Finally, it will reflect on methodological strengths and limitations, offering suggestions for future investigations.

6.2. Synthesis of key findings

6.2.1. Anxiety, interoception, and uncertainty in social interactions

One overarching theme across the empirical studies is the complex interplay between (negative) emotional experiences - whether expressed as trait anxiety, empathic distress, or autonomic arousal during social scenarios- interoceptive and social processes, and perceptions of uncertainty. Chapter 2, using path analysis, demonstrated that IU and alexithymia mediate the relationship between facets of IS and anxiety and affective empathy. Specifically, individuals who distrust their bodily sensations and worry or fixate on ambiguous interoceptive cues tend to exhibit heightened trait anxiety and empathic distress. In line with findings linking IU to social anxiety (Carleton et al., 2010), our results further suggest that IU can act as a reinforcer for maladaptive interpretations of bodily signals and as barrier to emotion regulation in uncertain social contexts (Cosmoiu and Nedelcea, 2022; Freeston & Komes, 2023).

Facets of these findings aligned with those revealed in the experimental dynamic socio-affective conditions of Chapter 3. Specifically, it was found that lower autonomic arousal during implicit learning of IEC predicted higher learning rates and volatility estimates, suggesting that heightened arousal, commonly observed in anxiety, may disrupt belief updating by amplifying noise, or self-centred predictions, in interoceptive-exteroceptive integration. Chapter 4 expanded on this by linking aversive emotional reactivity to social learning in an interactive and volatile social context where expected and unexpected uncertainty were manipulated. Specifically, our findings revealed that individuals with elevated PD exhibited reduced learning of social volatility, mirroring evidence indicating that heightened anxiety correlates with underestimation of volatility changes in non-threating environments (Pulcu & Browning, 2019), likely reflecting inflexible learning and avoidance patterns.

Chapter 5 offers a potential route to mitigate maladaptive learning and decision-making. In this study, we found that active taVNS enhanced learning rate adjustments between stable and volatile social conditions, likely by improving vagal tone, noradrenaline signalling, and interoceptive accuracy, allowing individuals to recalibrate social and bodily predictions to volatility shifts and ease social anxiety. Notably, this speaks to Chapter 2 findings suggesting that positive interoceptive appraisal buffers anxiety.

6.2.2. Interoception and uncertainty in emotion perception

The findings of Chapter 2 and 3 underscore how interoception and uncertainty in tandem shape how we perceive and interpret emotional cues, especially under conditions of ambiguity. In Chapter 2, we observed that IU mediates the relationship between facets of IS (Not Worrying and Trust) and emotion recognition of ambiguous facial expressions. Surprisingly, higher IU predicted better identification and more efficient processing of these expressions, as revealed by DDM, possibly reflecting attentional mechanisms driving individuals to process uncertain stimuli in greater detail (Fergus & Carleton, 2016).

Building on this, Chapter 3 introduces a dynamic learning paradigm for emotion perception, manipulating different uncertainty sources, wherein short-term socio-affective expectations shape how participants categorize ambiguous facial expressions. Specifically, implicitly learned IEC were shown to affect emotion recognition and enhance emotion egocentricity bias (EEB), particularly when emotion expressions were more difficult to categorize. In addition, individuals who more readily notice their bodily signals and integrate them into their emotional awareness demonstrated sharper tracking of shifting interpersonal affective states. Interestingly, individuals who trust and listen to their body for insight exhibited a tighter coupling between physiological arousal and learning, suggesting that believing in the reliability of one's bodily responses can amplify the integration of internal feedback with external social information. It was also revealed that those with greater awareness of the significance of their bodily sensations for their felt emotional experience could more efficiently incorporate information about learned IEC during emotion categorization, though it was not replicated in the lab study.

The findings of Chapter 2 and 3 frame emotion perception as an embodied inference process, where the brain dynamically integrates interoceptive and exteroceptive signals to resolve uncertainty, based on uncertainty estimates, interpersonal context (IEC), autonomic arousal and IS traits.

6.2.3. Interoception in adaptive empathic learning

Chapter 4 and 5 extend the thesis scope by examining more complex and dynamic social cognition processes to elucidate how interoceptive processing relates to precision adjustments during adaptive empathic learning under different conditions of volatility. Chapter 4 reveals that fluctuations in HEP amplitudes track precision-weighted prediction errors (pwPEs) during empathic feedback, aligning with prior evidence (Gentsch et al., 2018; Fournagan et al., 2024) suggesting HEP's sensitivity to affective probabilistic expectations in social contexts. We also found that unconscious interoceptive processing, as indexed again by HEP fluctuations during the decision-making, predicted response accuracy, highlighting the importance of interoceptive precision regulation for adaptive decision-making. Notably, enhanced modulation of interoceptive attention predicted participants' optimal use of inferred social contingencies during empathic responses. This finding aligns with Chapter 2, where heightened body-related emotional awareness was similarly associated with optimal decision-making. These results suggest that controlling interoceptive attention and bodily emotional awareness can facilitate contextualizing interoceptive cues, thereby utilizing them to inform intuitive decision-making. Additionally, we found that autistic traits modulated the integration of social predictions with bodily cues, suggesting reduced interoceptive integration and/or use of contextual information related to social volatility to fine-tune interoceptive precision.

Building on this, Chapter 5 experimentally manipulated interoceptive processing by applying taVNS to modulate adaptive interoceptive precision. Arguably, by activating noradrenergic pathways and accentuating the salience of interoceptive signals, taVNS improved participants ability to track changes in social uncertainty and adapt empathically to others' emotion regulation needs, suggesting a causal role for this neural pathway in adaptive empathy. Notably, this intervention weakened the link between autistic and interoceptive sensibility traits on volatility learning, suggesting that vagal enhancement can potentially be use as (interoceptive) neuromodulation in populations prone to misestimating unexpected uncertainty. These results also parallel Chapter 2's finding that lower autonomic arousal facilitates efficient learning, as taVNS may be a way to decouple disruptive arousal from social inference.

Importantly, Bayesian model comparison across all studies involving social learning – either simple associative learning or reinforcement learning (RL) – revealed that HGF models with volatility representations outperformed simpler model-free learning (Rescorla-Wanger) or non-volatility HGF models. This suggests that adaptive learning rates tuned by volatility estimates better account for the dynamic interpersonal learning of these studies. In sum, these results emphasize this thesis's central tenet: social cognition is an embodied negotiation of uncertainty, mediated by interoceptive precision and appraisal and amenable to intervention.

6.3. Theoretical implications

The empirical findings across the four chapters coalesce into a unified theoretical framework that positions social cognition as an embodied predictive process, where the brain resolves uncertainty by dynamically weighting interoceptive and exteroceptive signals. Rooted in predictive processing (PP) principles, this framework integrates hierarchical inference, precision-weighting, and neuromodulation to explain how individuals navigate ambiguity in social interactions. Next, I will discuss how our key findings fit into this framework and their implications related to theories and models of emotion perception, adaptive social learning, maladaptive interoceptive inference.

6.3.1. Emotion perception under uncertainty

The results from Chapters 2 and 3 portray emotion perception as an embodied, context-sensitive process, in line with PP accounts that emphasize continuous reweighting of internal bodily signals and external social cues under uncertainty, rather than fixed, context-insensitive simulations. Simulation accounts hold that perceiving another's affect involves reenacting their state in one's own body, and "feeling" what they feel, and then projecting these feelings onto them (Galleze, et al., 2004). Yet, our observation that IEC can bias emotion recognition challenges simulation as the main route to social understanding. If simulation was primary, it should be largely immune to contextual and learning effects. Instead, our results showed that participants used probabilistic associations between observed emotions and their own interoceptive/affective cues to categorize ambiguous emotional expressions. This supports a more flexible, learning-based perspective, such as the Learned Matching Hypothesis (LMH; Heyes, 2018) and PP, in which domain-general learning mechanisms forge bidirectional links between observed exteroceptive signals and interoceptive states. These associations are shaped by interpersonal regularities throughout development but are also relatively malleable even after short-term exposure to different conditions (Fotopoulou & Tsakiris, 2017; Heyes, 2018; Riva et al., 2016). We suggest that such interpersonal cross-modal bidirectional links not only underpin flexible mirroring but also shape interpersonal expectations that guide swift emotion judgments. Barsalou's (2013) pattern completion account further complements this perspective, proposing that perceiving someone's emotional cues partially reactivates multimodal representations (situated conceptualizations) that integrate perceptual, motor, and interoceptive information from past experiences. Thus, mirroring and projective simulation emerges not as innate, inflexible processes but as the result of learned cross-modal associations that produce context-relevant pattern activations.

A similar stance emerges from enactivist and direct perception theories (De Jaegher, 2009; De Jaegher et al., 2010, Zahavi, 2010b), which argue that emotional understanding is typically an immediate, socially embedded process shaped by dynamic embodied interactions, cultural norms, and situational goals. However, direct perception can be constrained by perceptual ambiguity or limited interactive cues and affordances, prompting greater reliance on cognitive or reflective processes (Gallagher & Varga, 2014). Our findings, which show increased reliance on IEC for the perception of more ambiguous expressions, may suggest a shift toward simulation mechanisms or inferential processes that incorporate contextual changes. Participants may have relied on top-down affective priors, comprising contextual influences and implicit inferences, to inform perception. Nonetheless, our experimental design does not allow us to confidently exclude non-perceptual processes, such as explicit inference or priming effects. Similarly, we cannot draw confidently argue that it provides evidence for the affective realism hypothesis, which posits that affective states colour the perception of ambiguous cues (Anderson et al., 2012; Barrett & Bar, 2009).

Mechanistically, PP's hierarchical Bayesian inference offers a unifying explanation for the exteroceptive-interoceptive integration during perceptual inference (Gendron & Barrett, 2018; Otten et al., 2017; Seth & Friston, 2016). Specifically, precision-weighting mechanisms regulate the extent to which the brain relies on top-down contextual priors versus bottom-up sensory or bodily input, depending on their respective reliability or uncertainty, to minimize surprise across hierarchical levels. This accounts not only for the observed top-down effects of learned IEC but also for their stronger influence in the expected congruency block, where short-term priors aligned with higher-order expectations about interpersonal congruency formed over a lifetime (as reflected in the EEB responses in the Neutral Expectancy Block; Heyes, 2018; Riva et al., 2016). These top-down predictions also played a greater role in resolving uncertainty when participants encountered highly ambiguous facial expressions, likely overriding noisy bottom-up signals. While theories emphasizing cross-modal associations, such as Barsalou's (2013) pattern completion and the LMH (Heyes, 2018), highlight learning mechanisms that flexibly shape social perception and behaviour, interoceptive PP provides a more cohesive mechanistic framework for social inference under uncertainty by incorporating the effects of context, past experiences, and interoceptive processes across multiple levels of personal and interpersonal abstraction.

This interoceptive inference framework has also been applied to emotion perception of the self (Seth & Friston, 2016; Quigley et al., 2021), reconciling traditionally opposing perspectives in emotion research, particularly regarding the tension between bodily primacy versus cognitive evaluation. In this context, individual traits, such as IU, function as high level priors, either driving negative interpretations of ambiguous bodily cues or improving perceptual efficiency of ambiguous social signal. Consequently, one's "precision modulator" profile, shaped by traits like IU, IS or alexithymia, can facilitate or hinder emotion perception of self and others, suggesting that social functioning is a dynamic interplay of bodily cues, learned associations, situational demands and developmental social trajectories.

Rather than indicating rigid simulations, our findings suggest that the body is involved in emotion perception through the dynamic integration of interoceptive and exteroceptive signals, shaped by interpersonal statistical regularities, past experience, uncertainty conditions, and individual traits within a hierarchical Bayesian inference. Accounting for these factors, interoceptive PP provides a unifying account of social perception as an embodied process of precision-weighting to minimize interpersonal uncertainty (De Bruin & Strijbos, 2015; De Bruin & Michael, 2021; Quadt, 2017).

6.3.2. Interoception in social learning and adaptive empathy under uncertainty

Beyond emotion perception, our findings from Chapters 2–4 underscore the different ways in which social learning and empathic responsiveness are intertwined with interoceptive processing, with bodily signals guiding belief updating in uncertain interpersonal contexts that involve multiple nested levels of psychophysiological inference.

Our findings demonstrating that HGF models outperformed simpler reinforcement-learning ones suggest that learning from volatile social information involves structured (hierarchical) inference, where individuals flexibly adjust predictions in response to environmental uncertainty rather than relying on past reinforcement alone. Framing learning in terms of PP, we can interpret this structured inference as multilevel prediction error minimisation, with higher-level priors shaping lower-level updates (Friston et al., 2016; Soltani & Izquierdo, 2019). In contrast, standard RL typically seeks to maximize external rewards without accounting for individual differences, top-down effect or dynamic uncertainty in a hierarchical probabilistic fashion.

Beyond high-level volatility estimates, our results indicate a role for interoceptive awareness and interoceptive signals as key modulators of learning, functioning either as high-level priors (e.g., IS effects on learning in Chapters 3 and 5), shaping belief updating about social contingencies, or low-level physiological signals that interact with uncertainty estimates to modulate learning rates (e.g., arousal and learning rate correlation in Chapter 3). This resonates with accounts postulating that social learning involves domain-general computational principles of prediction error minimisation along with social-cognitive processes such as mentalising and affective learning (Friston et al., 2016; Olsson et al., 2020; Puscian et al., 2022). Thus, our findings support the thesis that most complex social learning processes are better captured by model-based learning approaches, reinforcing hierarchical Bayesian accounts of learning. This suggests that multilevel inference - spanning sensory, affective, social, and cognitive domains- optimises adaptive responses in uncertain environments.

This framework highlights that the brain tracks not only the occurrence of outcomes but also the likelihood that these outcomes vary unpredictably over time, necessitating flexible recalibration of precision estimates across timescales while distinguishing between expected and unexpected uncertainty (Soltani & Izquierdo, 2019). In this context, and particularly within interoceptive PP, we can interpret our findings, such as the modulation of HEPs at feedback during empathic learning, as potential evidence for interoceptive pwPE associated with multimodal social predictions. Recent models of interoceptive social inference postulate that higher order social priors generate top-down predictions to interpret and explain away multimodal afferent signals (Ondobaka et al., 2017; Quattrocki & Friston, 2014). Precision-weighting lies at the core of the brain's computational process for minimizing surprise, dynamically balancing confidence between predictions and afferent signals, and allocating attention between self and other.

In the context of the present research (particularly regarding our findings on HEP modulations according to socio-affective predictions), each feedback moment represents an opportunity for the brain to update its model of the other person's needs. When socio-affective predictions are strong (narrow priors) during low volatility phases, the brain assigns high precision to interoceptive predictions, thereby reducing the salience of incoming social and interoceptive feedback and the extent of belief updating. Conversely, when unexpected social uncertainty increases, higher-level multimodal predictions recalibrate interoceptive and exteroceptive expectations and related precisions, and specifically increasing the relative precision of afferent signals, thus increasing the gain of interpersonal learning.

This suggests that HEPs (and interoceptive signals more broadly) do not reflect passive readouts of bodily state or outcome valence, but instead index ongoing recalibrations at the interface of internal cues and external (social) information. Importantly, while older theories such as the Somatic Marker Hypothesis (SMH) treat bodily signals as heuristic markers that automatically bias decision-making (Damasio, 1996), our results suggest a more flexible interpretation, while emphasizing the role of post-feedback bodily cues for learning, in line with newer evaluations of the SMH (Dunn et al., 2006). Interoceptive signals may be contextually weighted based on their estimated reliability and interpreted according to personality traits, rather than taken at face value, echoing proposed functions of interoceptive inference in decision-making (Dunn et al., 2010; Gu & FitzGerald, 2014).

This framework underlines the different ways in which the interplay between exteroceptive and interoceptive predictive processing underpin adaptive, hierarchical social inference. As low-level inputs, afferent interoceptive signals provide real-time PE signals that inform immediate learning adjustments. Our findings on HEPs and evoked heart responses, corroborate prior research and support the Arousal Coherence hypothesis which posits that autonomic fluctuations track shifts in uncertainty and learning rates (Biddell et al., 2024), with the extent to which interoceptive signals are integrated into higher-order affective and cognitive processes impacting adaptive behaviour. Thus, flexible social interactions depend on higher-level social priors, which generate precision-weighted high-level predictions to contextualise and interpret interoceptive inputs, distinguishing expected variability from truly unexpected changes. In this way, interoception functions as an embodied anchor for contextualising social cues, where differential processing of interoceptive signals scaffolds social inference.

Interoceptive precision-weighting is critical to interpersonal attunement and allostatic interpersonal regulation. Allostatic regulation is an anticipatory, active inference process, dynamically adjusting autonomic states in response to expected social contingencies (Atzil & Barrett, 2017; Burleso & Quigley, 2021). Bodily signals reflecting meaningful changes in another's behaviour are used to modulate empathic responses, whereas downregulating precision on irrelevant interoceptive signals facilitates adaptive action, as evidenced by our finding that lower HEP amplitudes predicted response accuracy. This aligns with suggestions that suppressing afferent PEs during action is crucial to prevent interference with motor planning (Seth, 2013; Pezzulo et al., 2015). From an allostatic perspective, the body remains poised to modulate arousal while attuning to another's affective fluctuations, illustrating how self-regulation and social inference operate in tandem. Maintaining stable affective functioning in uncertain social contexts requires continuous calibration of internal bodily signals with perceived external demands, enabling anticipation of another's physiological state and adaptive responses. This interpersonal uncertainty reduction, spanning multiple timescales, supports attunement and social learning, aligning with models suggesting that interpersonal neurophysiological coupling facilitates social interactions (Bolis et al., 2023; Pan et al., 2021; Quadt, 2017).

Within interoceptive PP, individual traits can be view as meta-priors, operating at higher levels of embodied social inference, modulating the precision and appraisal of social and interoceptive cues. In the context of our social learning modelling, HGF can capture these individual differences in its parameter space, that reflect higher order priors that develop and change across longer timescales (Mathys et al., 2011). Our findings revealed that differences in IS, affective empathy, and autistic traits modulate social learning in different ways. For instance, empathic concern appears to enhance precision/attention in others' emotion regulation needs (EC and learning rate in Chapters 4-5), while greater reliance on bodily sensations for insight predicts increased coherence between arousal and learning rate fluctuations (Chapter 3). However, further research is needed to examine how increased interoceptive integration or coherence between autonomic fluctuations and felt arousal affects adaptive learning under uncertainty, as postulated by the Arousal Coherence hypothesis (Biddell et al., 2024).

Our findings also highlight promising means of enhancing interoceptive integration to improve adaptive social cognition, with taVNS suggested as a tool for boosting interoceptive processing and motivated behaviour (Neuser et al., 2020; Paciorek & Scora, 2020; Villani et al., 2019). The observed improvements in social adaptability to volatility could have stemmed from multiple, overlapping mechanisms. From a Polyvagal Theory perspective (Porges, 2009), boosting vagal regulation controls sympathetic hyperarousal, creating a physiological milieu conducive to calm-yet-alert engagement with uncertain social signals. In parallel, evidence suggests that taVNS can increase interoceptive accuracy (Villani et al., 2019), HEP amplitudes (Poppa et al., 2022), and HRV (Ventura-Bort & Weymar, 2024), all markers of superior interoceptive precision and homeostatic regulation. This improved calibration of bodily signals may underpin the observed modulation in adaptive social learning, supporting the argument that the observed fluctuations in our HEPs reflect changes in interoceptive precision. Additionally, taVNS is known to activate the locus coeruleus (LC) and its noradrenergic system, integral for updating beliefs when contingencies shift (Angela & Dayan, 2005; Lawson et al., 2021). According to active inference models, state-action PEs arising from LC firing drive learning-rate adjustments, enabling rapid reconfiguration of internal models, ascending NA input to cortical areas, including the ACC, optimizing the balance between plasticity and stability (Sales et al., 2019). Thus, tVNS appears to promote optimal precision-weighting of interoceptive and exteroceptive signals, fostering a functional brain–body coupling that supports flexible yet stable social inference under shifting uncertainty.

6.3.3. Maladaptive interoceptive inference and clinical implications

IU, anxiety and bodily and social uncertainty

While we have focused mainly on how interoception can facilitate social inference and adaptive decision-making by reducing interpersonal uncertainty, our findings also underscore how internal and external uncertainty can be related to maladaptive interoceptive processing and social behaviour, across several conditions such as IU, anxiety and autism. Our findings from Chapter 2 suggest that individuals who exhibit low trust in their bodily sensations and display heightened worry and fixation on ambiguous interoceptive cues report significantly higher levels of trait anxiety and empathic distress (PD). This corroborates Freeston and Komes' (2023) proposal that IU is fundamentally tied to interoceptive discrepancies, where even minor bodily uncertainty is experienced as a significant somatic error, leading to heightened distress and difficulty internalizing safety cues. This proposal builds upon previous interoceptive PP models of anxiety and psychopathology (Paulus et al., 2019) as well as on the Generalized Unsafety Theory of Stress (Brosschot et al., 2016), both of which posit that chronic uncertainty misattribution underlies maladaptive anxiety responses. Our results showing that trust in bodily signals mitigates PD, but only when IU is low, underscore the dual role of interoception as both a resource (when bodily signals are correctly interpreted) and a liability (when uncertainty is met with maladaptive interoceptive appraisals). Thus, interventions aimed at reducing IU and improving interoceptive trust may have downstream benefits for social functioning, particularly in individuals prone to social anxiety and distress in interpersonal contexts.

From a computational psychiatry perspective, IU and anxiety can be understood as a failure to appropriately calibrate interoceptive PE, thereby amplifying bodily uncertainty, and reinforcing maladaptive anxiety responses and avoidance of ambiguity, which in turn could lead to emotion dysregulation in social interactions (Freeston & Komes, 2023; Paulus & Stein, 2010). Both conditions might be characterised by rigid priors that inflate PE and overweight uncertainty as threat, leading to overgeneralized fear responses, difficulties in updating beliefs about safety, driving a state of chronic alarm. Our findings further suggest how this maladaptive inference could unfold in dynamic social interactions as individuals with high PD underestimated social volatility (Chapter 4), corroborating previous evidence showing that trait anxiety is linked to reduced learning rate adaptation to volatility in non-threating conditions (Pulcu & Browning, 2019). Interestingly, in threating environments anxious individual tend to overestimate volatility, treating most unexpected events as signal threat and abruptly increasing unexpected uncertainty. Furthermore, while not directly linked to anxiety, we also found that higher arousal, possibly reflecting unresolved PEs, impaired the learning of IEC (Chapter 3). The inability to minimize or downregulate PEs draws attention and cognitive resources away from task-relevant information, thus interfering with exteroceptive-interoceptive integration and social learning. Rather than supporting flexible social cognition, heightened arousal and worry can generate a self-reinforcing cycle of hypervigilance, misestimation of social volatility and perception social behaviours as unpredictable. Taken together, our findings extend maladaptive interoceptive inference accounts by suggesting that anxiety-related learning responses in social contexts are driven by misestimations of uncertainty and social safety, as well as a sense of body distrust (Paulus & Stein, 2010; Pulcu & Browning, 2019; Sandhu et al., 2023).

Autism, interoceptive inference and social uncertainty

Anxiety exhibits high comorbidity with autism (Hollocks et al., 2019), with IU found to mediate this relationship (Jenkinson et al., 2020), leading others to attribute this comorbidity to IU (Stark et al., 2021). Given that atypicalities in interoceptive and uncertainty processing have also been observed in autism, interoceptive PP has similarly been proposed as an explanatory framework, with most accounts suggesting that individuals with ASD exhibit impaired multimodal integration, overprecise priors, and limited contextualization of interoceptive cues (Lawson et al., 2014; Palmer et al., 2017; Proof et al., 2023). These three factors interact, yielding a highly diverse phenotype, as ASD profiles manifest in varied ways depending on contextual factors.

Our findings on autist traits and social learning in a largely neurotypical sample can be explained through this lens. The reduced interoceptive integration observed for individuals with higher autistic traits aligns with Proff and colleagues (2022) recent theoretical proposal building on PP and coherence theories of autism. According to their model, autistic individuals exhibit atypical interoceptive-exteroceptive integration, leading to difficulties contextualising interoceptive cues within broader social and environmental contexts. This results in ineffective predictions and failure to attenuate interoceptive PEs leading to abnormal interoceptive gain control, reflected in the anterior insular and cingulate cortex activity, both critical brain networks for interoceptive precision-weighting and social inference (Eilam-Stock et al., 2014; Harrison et al., 2021; Singer et al., 2009). This perspective resonates with our HEP findings, which show that autistic traits are linked to reduced differentiation of interoceptive signals during social learning, suggesting maladaptive interoceptive precision-weighting according to socio-affective predictions and difficulty incorporating bodily cues during dynamic social interactions. Furthermore, our finding also revealed reduced integration of social predictions during decision-making (Chapter 4), aligning with perspectives that consider autism not as deficit in processing social information per se, but as a difficulty in using relevant contextual information for action selection (Palmer et al., 2015; Sevgi et al., 2020).

Autistic individuals may also exhibit atypical volatility estimation, preferring structured, predictable environments, and rule-based learning over dynamically shifting social contingencies (Sandhu et al., 2023). Indeed, our finding that autistic traits are linked to overlearning of volatility shift (sham condition; Chapter 5) align with previous findings and theoretical suggestions (Lawson et al., 2017; Sandhu et al., 2023). Notably, this overaction to volatility resembles learning patterns reported in anxious individuals. Given the high comorbidity of anxiety, autism and IU, it has been postulated that this learning behaviour in ASD might indicate heightened anxiety under conditions of increased uncertainty rather than core autistic features (Sandhu et al., 2023), a hypothesis that is important to be investigated. Importantly, taVNS mitigated this learning pattern, possibly by recalibrating interoceptive precision/accuracy and noradrenaline signal, both of which are implicated as problematic in Bayesian accounts of ASD (Lawson et al., 2014; Quattrocki & Friston, 2014).

Despite this evidence, our findings across the two studies did not demonstrate any significant deficit of adaptive empathy and response accuracy for individuals with high autistic traits, which could be interpreted in several ways. For instance, individuals with autistic traits but without actual clinical symptomatology might effectively use alternative/compensatory mechanisms, such as pattern identification, to respond adaptively under social uncertainty. Another suggestion is that psychopathology in autism should not be related to suboptimal estimation of social uncertainty and belief updating per se but rather arising from difficulties in interpersonal attunement in real-life dynamic interaction, as second-person social neuroscience accounts (Redcay & Schilbach, 2019; Bolis et al., 2023) and Interaction Theories of social cognition suggest (Gallagher & Varga. 2015). This view also addresses the high heterogeneity of autism, as observed in the variability of our results as well, where different predictive strategies, interoceptive precision-weighting, and volatility estimation give rise to diverse social interaction profiles. This aligns with contemporary efforts in computational psychiatry highlighting the need for a multi-dimensional, data-driven approach to mental health that captures interpersonal dynamics rather than relying on static diagnostic categories (Petzschner et al., 2017; Stephan & Mathys, 2014). In this line, the "autism space" framework proposes that autism is best understood as a multi-scale, interaction condition rather than a fixed diagnostic category, emphasizing mismatches in interpersonal inference, societal structures rather than intrinsic deficits (Bolis et al., 2023). Thus, proposed novel approaches like generative embedding move beyond traditional individualized psychiatry (Brodersen et al., 2011; Stephan & Mathys, 2014), facilitating the identification of mechanistic subtypes through the combination of Bayesian modelling, machine learning, and real-time psychophysiological measurements across individual, social and biological levels.

This approach has been suggested as crucial for understanding other conditions beyond autism that can be conceptualized as disorders of social interaction, like social anxiety, depression and schizophrenia (Bolis et al., 2018; Redcay & Schilbach, 2019). The dynamic nature of our tasks could have captured such disruptions of interpersonal attunement, though the development of ecologically valid, real-time interactive paradigms is necessary. This perspective also aligns with Fuchs' (2013) view that psychopathology critical involves disruptions in subjective and intersubjective temporality. Autism and anxiety could, therefore, be also viewed as disturbances of interpersonal temporality, where individuals experience misalignment with the rhythms of social interactions, leading to difficulties in social adaptation and attunement. This interpersonal misattunement is further affected by and affects intrapersonal bodily attunement (Fuchs 2017). Given that predictive social engagement relies on synchronizing with others at multiple temporal scales, atypical volatility estimation and interoceptive-exteroceptive integration in autism and anxiety may also reflect disruptions in embodied temporal coordination, rather than isolated cognitive deficits. This suggests that interventions targeting interpersonal synchrony, such as real-time biofeedback or interoceptive recalibration, could enhance social adaptation by restoring dynamic temporal alignment between self and others.

6.4. Methodological Evaluation

6.4.1. Strengths

A significant strength of this work is its multi-method integrative approach, combining behavioural, physiological, neural stimulation and computational techniques along with various self-report assessments to provide a multifaceted investigation of interoception in social cognition under uncertainty. Using various computational models, the research captures important aspects of the dynamic and hierarchical nature of predictive social inference. This integration of diverse methodologies supports a mechanistic understanding of interoceptive contributions to emotion perception, empathy, and adaptive learning, directly linking our findings to popular computational neuroscience perspectives, particularly within the interoceptive PP framework.

A key strength of this work is the development of novel experimental paradigms to investigate EEB and adaptive empathic learning under conditions of uncertainty. The newly designed EEB task introduces a contextually modulated affective induction paradigm for probabilistic inference of IEC that enhances ecological validity while maintaining a subtle distinction between self and other. This design is crucial for probing affective projection mechanisms, ensuring that IEC are learned implicitly. Similarly, the adaptive empathy task advances prior work (Kozakevich Arbel et al., 2021) by incorporating different levels of social volatility, allowing the study of how individuals dynamically adjust their learning rates in response to shifting environmental conditions requiring optimal tracking of expected and unexpected uncertainty. These paradigms provide controlled means of testing social inference mechanisms while enhancing affective engagement and ecological validity.

This thesis also applies advanced computational modelling, using DDM to examine perceptual and decision-making processes of emotion perception under uncertainty as well as HGF to model individual differences in adaptive social learning within a meta-Bayesian computational neuroscience framework (Daunizeau et al., 2010). In contrast to standard RL models, this Bayesian approach allows for hierarchical belief updating and precision-weighting of uncertainty, capturing confidence estimates alongside variable learning rates. The use of Bayesian model selection across studies ensures that model validity is formally tested and optimally fitted to the data, reinforcing the thesis's methodological precision.

A further methodological strength is the application of various interoceptive measures, ranging from self-report (MAIA) to neurophysiological markers (HEPs), enabling a multi-level analysis of predictive processing, tapping onto different aspects of interceptive processing. This thesis is also the first to apply an interoceptive intervention (taVNS) in a reinforcement-based social learning task, offering novel insights into the neuromodulatory effects of vagal stimulation on interoceptive processing and adaptive empathy.

Finally, cross-study consistency and replicability underpin the reliability of the findings. Key results, such as the role of IU in interoceptive-affective interactions, the effect of IEC on emotion recognition, the role of interoception in adaptive social learning and the superiority of Bayesian learning models with volatility estimates over simpler RL approaches were consistently observed across multiple experiments. This coherence strengthens the thesis's theoretical claims, reinforcing interoceptive inference as a fundamental mechanism underlying social cognition in uncertain environments.

6.4.2. Limitations

Despite its strengths, this thesis also presents several methodological limitations that should be addressed in future research. Firstly, sample size and generalisability are important points for consideration. The predominantly undergraduate sample limits the extent to which these findings generalise to other populations, especially with respect to age, gender, and clinical relevance. Moreover, some individual differences analyses, particularly those examining IS, IU, and autistic traits, may have been underpowered, necessitating larger samples to better capture their effects on social learning.

Additionally, task constraints may limit ecological validity. While the emotion recognition task provides valuable insights into EEB, the use of static facial expressions does not fully capture the dynamic, context-dependent nature of real-world emotion perception. Similarly, the adaptive empathy task, although advancing prior paradigms, remains a highly structured predetermined environment, which does not reflect the complexity of interpersonal interactions in natural settings, particularly due to the absence of a real partner, which fundamentally alters how individuals behave in real-life interactions and adapt to ensure interpersonal emotion regulation. Another limitation concerns the absence of a non-social control condition. Without a comparison task, it remains unclear how interoceptive processing influences and is influenced by social learning, and how it differs in non-social conditions with different emotional implications. In addition, the EEB paradigm should be further developed to clarify the impact of perceptual noise and the potential contribution of explicitly learned interpersonal associations. This task could further be improved by separately modulating IEC and emotion recognition in distinct trial sets within the same dual-task paradigm.

Furthermore, while this research incorporates HEPs and ECG recordings, additional physiological measures (e.g., HRV, pupillometry), particularly in the context of taVNS-induced effects on social inference, would strengthen claims regarding the impact of vagal and neuromodulation on interoceptive processing and social inference mechanisms. Lastly, the effects of IS, affective empathy, and autistic traits on social learning parameters modelled by HGF were not fully replicated across studies. While several findings of this research suggested a modulatory role of these traits in belief updating and adaptive learning, some of their effects were either small or were not replicated, indicating that future research with larger samples is needed to clarify their exact contributions.

6.5. Future directions

The findings of this thesis along with related theoretical proposals, point to several potential pathways for future research to further advance our understanding of interoceptive processing in social cognition under uncertainty. Second-person neuroscience provides a promising methodological and theoretical framework to advance our understanding of the psychophysiological mechanisms underpinning effective social interactions (Bolis et al., 2018; Redcay & Schilbach, 2019). Thus, a critical next step is the integration of real-time dyadic interaction paradigms in social cognition research. Future studies could investigate whether interpersonal physiological coupling, such as synchronisation of autonomic responses, influences social belief updating. Additionally, neural synchronisation during social learning could be examined to test whether interoceptive predictions and precision modulate interpersonal neurophysiological and behavioural alignment.

Methodological improvements could also involve the use of taVNS combined with other physiological and neural markers, including HEPs, pupillometry, and fMRI, to determine its effects on interoceptive processing, noradrenergic modulation, and adaptive social inference. Combined with computational approaches, this would facilitate a clearer characterisation of how vagal stimulation influences interoceptive precision-weighting and adaptive social behaviour. Another direction of high importance is the direct manipulation of interoceptive predictions, for example, using respiratory resistance paradigms that would allow for experimental modulations of afferent and efferent interoceptive signals to examine their influence on perceptual and social learning processes (Harrison et al., 2021). This could provide a critical causal test of interoceptive inference mechanisms in social cognition.

Beyond HGF, we can employ alternative Bayesian models, particularly active inference approaches, to explore the interaction between action planning and interoceptive predictions, as well as to investigate hierarchical inference beyond volatility learning. This extends interactive social inference beyond passive belief updating to examine how individuals actively shape their bodily and sensory input through goal-directed behaviour, providing key insights into interoceptive predictive processing effects in social cognition. Additionally, machine learning techniques, such as generative embedding, could be applied for unsupervised classification of individual differences in psychophysiological states, learning, and decision-making (Brodersen et al., 2011). This approach combines mechanistic modelling with discriminative classification, enabling latent space representations of various subtypes of interoceptive processing and social inference. Along with larger and more diverse samples, this could help identify, among others, distinct interoceptive phenotypes linked to adaptive or maladaptive social cognition.

From a translational perspective, developing interventions aimed at improving interoceptive integration and appraisal, such as biofeedback, mindfulness-based training, or vagal stimulation protocols, could offer therapeutic pathways for populations with atypical interoceptive processing, including anxiety and autism.

Finally, developmental studies should investigate how bodily self-organisation, self-regulation, and uncertainty processing evolve across the lifespan, considering sociocultural influences on interoceptive awareness and social learning trajectories. Examining early social interactions and attachment profiles in relation to interoceptive processes and interpersonal regulation could provide critical insights into how embodied social cognition develops over time.

6.6. Conclusion

This thesis enriches emerging perspectives of social cognition as an embodied anticipatory process, where interoception and uncertainty dynamically shape how we perceive, learn, and adapt to the social world. By integrating behavioural, physiological, and computational approaches, this research highlights several mechanisms through which interoceptive signals interact with emotion perception, social learning, and interpersonal emotion regulation, particularly in uncertain social environments.

Across four empirical chapters, the findings suggest that intolerance of uncertainty, interoceptive sensibility, and alexithymia can modulate social and affective processes, while hierarchical Bayesian modelling captures how uncertainty estimates, physiological process and individual traits influence belief updating in volatile environments. Additionally, short-term interpersonal emotion contingencies were found to modulate the emotion egocentricity bias, while model comparisons indicated that social learning was informed by higher-order beliefs about volatility. Crucially, the neural markers of interoceptive processing, namely HEPs, were found to track pwPE during empathic feedback related to social predictions, and predict adaptive decision-making, suggesting that flexible processing of bodily signals guides social inference under volatility. Furthermore, this work indicates taVNS as a neuromodulatory tool that could recalibrate interoceptive precision, thereby supporting adaptive social learning.

Taken together, these observations indicate that social cognition is not merely a passive reading of external signals but potentially a dynamic negotiation of embodied predictions, influenced by individual traits and different sources of uncertainty. By tentatively bridging predictive processing models with emerging clinical insights, this thesis highlights potential avenues for interventions, such as biofeedback or vagal neuromodulation, aimed at alleviating social anxiety and enhancing resilience. Future research should refine and extend these insights into real-world dyadic interactions, developmental trajectories, and clinical populations, leveraging advances in computational psychiatry and second-person neuroscience.

References

Adams, K. L., Edwards, A., Peart, C., Ellett, L., Mendes, I., Bird, G., & Murphy, J. (2022). The association between anxiety and cardiac interoceptive accuracy: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *140*, 104754.

Adams, R. B., & Kveraga, K. (2015). Social vision: Functional forecasting and the integration of compound social cues. *Review of Philosophy and Psychology*, *6*, 591-610.

Adolfi, F., Couto, B., Richter, F., Decety, J., Lopez, J., Sigman, M., ... & Ibáñez, A. (2017). Convergence of interoception, emotion, and social cognition: a twofold fMRI meta-analysis and lesion approach. *Cortex*, *88*, 124-142.

Ainley, V., Apps, M. A., Fotopoulou, A., & Tsakiris, M. (2016). 'Bodily precision': a predictive coding account of individual differences in interoceptive accuracy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160003.

Ainley, V., Maister, L., & Tsakiris, M. (2015). Heartfelt empathy? No association between interoceptive awareness, questionnaire measures of empathy, reading the mind in the eyes task or the director task. *Frontiers in Psychology*, *6*, 554.

Allen, M. (2020). Unravelling the neurobiology of interoceptive inference. *Trends in Cognitive Sciences*, *24*(4), 265-266.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482.

Allen, M., Legrand, N., Correa, C. M. C., & Fardo, F. (2020). Thinking through prior bodies: autonomic uncertainty and interoceptive self-inference.

Allen, M., Varga, S., & Heck, D. H. (2023). Respiratory rhythms of the predictive mind. *Psychological Review*, *130*(4), 1066.

Alvarez-Dieppa, A. C., Griffin, K., Cavalier, S., & McIntyre, C. K. (2016). Vagus nerve stimulation enhances extinction of conditioned fear in rats and modulates arc protein, CaMKII, and GluN2B-containing NMDA receptors in the basolateral amygdala. *Neural Plasticity*, *2016*(1), 4273280.

Ambrosecchia, M., Ardizzi, M., Russo, E., Ditaranto, F., Speciale, M., Vinai, P., ... & Gallese, V. (2017). Interoception and autonomic correlates during social interactions. Implications for anorexia. *Frontiers in human neuroscience*, *11*, 219.

Anderson, E. C., Carleton, R. N., Diefenbach, M., & Han, P. K. (2019). The relationship between uncertainty and affect. *Frontiers in psychology*, *10*, 2504.

Anderson, E., Siegel, E., White, D., &, L. F. (2012). Out of sight but not out of mind: unseen affective faces influence evaluations and social impressions. *Emotion*, *12*(6), 1210.

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681-692.

Arslanova, I., Galvez-Pol, A., Kilner, J., Finotti, G., & Tsakiris, M. (2022). Seeing through each other's hearts: Inferring others' heart rate as a function of own heart rate perception and perceived social intelligence. *Affective Science*, *3*(4), 862-877.

Atzil, S., & Barrett, L. F. (2017). Social regulation of allostasis: Commentary on "Mentalizing homeostasis: The social origins of interoceptive inference" by Fotopoulou and Tsakiris. *Neuropsychoanalysis*, *19*(1), 29-33.

Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nature human behaviour*, *2*(9), 624-636.

Azevedo, R. T., Garfinkel, S. N., Critchley, H. D., & Tsakiris, M. (2017). Cardiac afferent activity modulates the expression of racial stereotypes. *Nature Communications*, *8*(1), 13854.

Azevedo, R. T., von Mohr, M., & Tsakiris, M. (2023). From the viscera to first impressions: phase-dependent cardio-visual signals bias the perceived trustworthiness of faces. *Psychological Science*, *34*(1), 120-131.

Azzalini, D., Rebollo, I., & Tallon-Baudry, C. (2019). Visceral signals shape brain dynamics and cognition. *Trends in cognitive sciences*, *23*(6), 488-509.

Babo-Rebelo, M., Buot, A., & Tallon-Baudry, C. (2019). Neural responses to heartbeats distinguish self from other during imagination. *NeuroImage*, *191*, 10-20.

Bagby, R. M., Taylor, G. J., & Parker, J. D. (1994). The twenty-item Toronto Alexithymia Scale—II. Convergent, discriminant, and concurrent validity. Journal of psychosomatic research, 38(1), 33-40.

Baiano, C., Job, X., Santangelo, G., Auvray, M., & Kirsch, L. P. (2021). Interactions between interoception and perspective-taking: Current state of research and future directions. *Neuroscience & Biobehavioral Reviews*, *130*, 252-262.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modelling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).

Baltazar, M., Hazem, N., Vilarem, E., Beaucousin, V., Picq, J. L., & Conty, L. (2014). Eye contact elicits bodily self-awareness in human adults. *Cognition*, *133*(1), 120-127.

Banellis, L., & Cruse, D. (2020). Skipping a beat: heartbeat-evoked potentials reflect predictions during interoceptive-exteroceptive integration. *Cerebral Cortex Communications*, *1*(1), tgaa060.

Barel, E., & Cohen, A. (2018). Effects of acute psychosocial stress on facial emotion recognition. *Psychology*, *9*(3), 403-412.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. Journal of autism and developmental disorders, 34(2), 163-175.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*, 5-17.

Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1325-1334.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature reviews neuroscience*, *16*(7), 419-429.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current directions in psychological science*, *20*(5), 286-290.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160011.

Barsalou, L. W. (2013). Mirroring as pattern completion inferences within situated conceptualizations. Cortex, 49(10), 2951-2953.

Beffara, B., Bret, A. G., Vermeulen, N., & Mermillod, M. (2016). Resting high frequency heart rate variability selectively predicts cooperative behavior. *Physiology & Behavior*, *164*, 417-428.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245-249.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214-1221.

Berardis, D. D., Campanella, D., Nicola, S., Gianna, S., Alessandro, C., Chiara, C., ... & Ferro, F. M. (2008). The impact of alexithymia on anxiety disorders: a review of the literature. *Current Psychiatry Reviews*, *4*(2), 80-86.

Berntson, G. G., & Khalsa, S. S. (2021). Neural circuits of interoception. *Trends in neurosciences*, *44*(1), 17-28.

Biddell, H., Solms, M., Slagter, H., & Laukkonen, R. (2024). Arousal coherence, uncertainty, and well-being: an active inference account. *Neuroscience of consciousness*, *2024*(1), niae011.

Bijsterbosch, J. M., Hasenack, B., van Rooijen, B., Sternheim, L. C., Boelen, P. A., Dijkerman, H. C., & Keizer, A. (2023). Intolerable feelings of uncertainty within the body: Associations between interoceptive awareness, intolerance of uncertainty, and body dissatisfaction. *Journal of Adolescence*, *95*(8), 1678-1688.

Boelen, P. A., & Reijntjes, A. (2009). Intolerance of uncertainty and social anxiety. *Journal of anxiety disorders*, *23*(1), 130-135.

Bolis, D., & Schilbach, L. (2020). 'I interact therefore I am': the self as a historical product of dialectical attunement. *Topoi*, *39*, 521-534.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2018). Beyond autism: Introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology*, *50*(6), 355-372.

Bolis, D., Dumas, G., & Schilbach, L. (2023). Interpersonal attunement in social interactions: from collective psychophysiology to inter-personalized psychiatry and beyond. *Philosophical Transactions of the Royal Society B*, *378*(1870), 20210365.

Bornemann, B., Kok, B. E., Boeckler, A., & Singer, T. (2016). Helping from the heart: Voluntary upregulation of heart rate variability predicts altruistic behavior. *Biological psychology*, *119*, 54-63.

Bowles, S., Hickman, J., Peng, X., Williamson, W. R., Huang, R., Washington, K., ... & Welle, C. G. (2022). Vagus nerve stimulation drives selective circuit modulation through cholinergic reinforcement. *Neuron*, *110*(17), 2867-2885.

Bredemeier, K., & Berenbaum, H. (2008). Intolerance of uncertainty and perceived threat. *Behaviour Research and Therapy*, *46*(1), 28-38.

Brennan, G. M., & Baskin-Sommers, A. R. (2020). Aggressive realism: More efficient processing of anger in physically aggressive individuals. *Psychological science*, *31*(5), 568-581.

Brewer, R., Cook, R., & Bird, G. (2016). Alexithymia: a general deficit of interoception. *Royal Society open science*, *3*(10), 150664.

Brosschot, J. F., Verkuil, B., & Thayer, J. F. (2016). The default response to uncertainty and the importance of perceived safety in anxiety and stress: An evolution-theoretical perspective. *Journal of anxiety disorders*, *41*, 22-34.

Browning, M., Behrens, T. E., Jocham, G., O'reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature neuroscience*, *18*(4), 590-596.

Burger, A. M., Van Diest, I., van der Does, W., Hysaj, M., Thayer, J. F., Brosschot, J. F., & Verkuil, B. (2018). Transcutaneous vagus nerve stimulation and extinction of prepared fear: a conceptual non-replication. *Scientific reports*, *8*(1), 11471.

Burleson, M. H., & Quigley, K. S. (2021). Social interoception and social allostasis through touch: legacy of the somatovisceral afference model of emotion. *Social neuroscience*, *16*(1), 92-102.

Butera, C. D., Harrison, L., Kilroy, E., Jayashankar, A., Shipkova, M., Pruyser, A., & Aziz-Zadeh, L. (2023). Relationships between alexithymia, interoception, and emotional empathy in autism spectrum disorder. *Autism*, *27*(3), 690-703.

Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. Current biology, 16(19), 1905-1910.

Carleton, R. N. (2016). Into the unknown: A review and synthesis of contemporary models involving uncertainty. *Journal of anxiety disorders*, *39*, 30-43.

Carleton, R. N., Collimore, K. C., & Asmundson, G. J. (2010). “It's not just the judgements—It's that I don’t know”: Intolerance of uncertainty as a predictor of social anxiety. *Journal of Anxiety Disorders*, *24*(2), 189-195.

Carleton, R. N., Duranceau, S., Shulman, E. P., Zerff, M., Gonzales, J., & Mishra, S. (2016). Self-reported intolerance of uncertainty and behavioural decisions. *Journal of behavior therapy and experimental psychiatry*, *51*, 58-65.

Carleton, R. Nicholas. "The intolerance of uncertainty construct in the context of anxiety disorders: Theoretical and practical perspectives." *Expert review of neurotherapeutics* 12.8 (2012): 937-947.

Catmur, C., Mars, R. B., Rushworth, M. F., & Heyes, C. (2011). Making mirrors: premotor cortex stimulation enhances mirror and counter-mirror motor facilitation. Journal of Cognitive Neuroscience, 23(9), 2352-2362.

Ceunen, E., Vlaeyen, J. W., & Van Diest, I. (2016). On the origin of interoception. *Frontiers in psychology*, *7*, 743.

Chatham, C. H., Huppert, T. J., & Müller, R. A. (2022). Measures of tonic and phasic activity of the locus coeruleus—norepinephrine system in children with autism spectrum disorder: An event-related potential and pupillometry study. Autism Research, 15(10), 1792-1807.

Childs, J. E., Alvarez-Dieppa, A. C., McIntyre, C. K., & Kroener, S. (2015). Vagus nerve stimulation as a tool to induce plasticity in pathways relevant for extinction learning. *Journal of visualized experiments: JoVE*, (102), 53032.

Cisler, J. M., Olatunji, B. O., Feldner, M. T., & Forsyth, J. P. (2010). Emotion regulation and the anxiety disorders: An integrative review. *Journal of psychopathology and behavioral assessment*, *32*, 68-82.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, K. B., Naritoku, D. K., Smith, D. C., Browning, R. A., & Jensen, R. A. (1999). Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nature neuroscience*, *2*(1), 94-98.

Clark, K. B., Smith, D. C., Hassert, D. L., Browning, R. A., Naritoku, D. K., & Jensen, R. A. (1998). Posttraining electrical stimulation of vagal afferents with concomitant vagal efferent inactivation enhances memory storage processes in the rat. *Neurobiology of learning and memory*, *70*(3), 364-373.

Coll, M. P., Hobson, H., Bird, G., & Murphy, J. (2021). Systematic review and meta-analysis of the relationship between the heartbeat-evoked potential and interoception. *Neuroscience & Biobehavioral Reviews*, *122*, 190-200.

Colzato, L. S., Sellaro, R., & Beste, C. (2017). Darwin revisited: The vagus nerve is a causal element in controlling recognition of other's emotions. *Cortex*, *92*, 95-102.

Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: from origin to function. *Behavioral and brain sciences*, *37*(2), 177-192.

Cook, R., Brewer, R., Shah, P., & Bird, G. (2013). Alexithymia, not autism, predicts poor recognition of emotional facial expressions. *Psychological science*, *24*(5), 723-732.

Cooper, R. M., Rowe, A. C., & Penton-Voak, I. S. (2008). The role of trait anxiety in the recognition of emotional facial expressions. *Journal of Anxiety Disorders*, *22*(7), 1120-1127.

Corcoran, A. W., & Hohwy, J. (2017). Allostasis, interoception, and the free energy principle: Feeling our way forward.

Couto, B., Adolfi, F., Velasquez, M., Mesow, M., Feinstein, J., Canales-Johnson, A., ... & Ibanez, A. (2015). Heart evoked potential triggers brain responses to natural affective scenes: a preliminary study. *Autonomic Neuroscience*, *193*, 132-137.

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, *13*(4), 500-505.

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current opinion in psychology*, *17*, 7-14.

Critchley, H. D., & Garfinkel, S. N. (2018). The influence of physiological signals on cognition. *Current Opinion in Behavioral Sciences*, *19*, 13-18.

Critchley, H.D. (2005) Neural mechanisms of autonomic, affective, and cognitive integration. J. Comp. Neurol. 493, 154-166.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion review*, *8*(2), 144-153.

Damasio, A. R. (1994). Descartes' error and the future of human life. *Scientific American*, *271*(4), 144-144.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351(1346), 1413-1420.

Dan-Glauser, E. S., & Scherer, K. R. (2012). The difficulties in emotion regulation scale (DERS). *Swiss Journal of Psychology*.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS one*, *5*(12), e15554.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS one*, *5*(12), e15554.

Davis, M. H. (1980). Interpersonal reactivity index.

De Berardis, D., Campanella, D., Gambi, F., La Rovere, R., Sepede, G., Core, L., ... & Ferro, F. M. (2007). Alexithymia, fear of bodily sensations, and somatosensory amplification in young outpatients with panic disorder. *Psychosomatics*, *48*(3), 239-246.

De Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature communications*, *7*(1), 10996.

de Bézenac, C. E., Swindells, R. A., & Corcoran, R. (2018). The necessity of ambiguity in self–other processing: A psychosocial perspective with implications for mental health. *Frontiers in Psychology*, *9*, 2114.

De Bruin, L., & Michael, J. (2021). Prediction error minimization as a framework for social cognition research. *Erkenntnis*, *86*, 1-20.

De Bruin, L., & Strijbos, D. (2015). Direct social perception, mindreading and Bayesian predictive coding. *Consciousness and Cognition*, *36*, 565-570.

De Jaegher, H. (2009). Social understanding through direct perception? Yes, by interacting. *Consciousness and cognition*, *18*(2), 535-542.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition?. *Trends in cognitive sciences*, *14*(10), 441-447.

De Smet, S., Baeken, C., Seminck, N., Tilleman, J., Carrette, E., Vonck, K., & Vanderhasselt, M. A. (2021). Non-invasive vagal nerve stimulation enhances cognitive emotion regulation. *Behaviour research and therapy*, *145*, 103933.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, *3*(2), 71-100.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in psychology*, *9*, 580.

Desmedt, O., Corneille, O., & Luminet, O. (2024). The conceptualization and measurement of interoception. In *Interoception: A Comprehensive Guide* (pp. 35-74). Cham: Springer International Publishing.

Deuter, C. E., Nowacki, J., Wingenfeld, K., Kuehl, L. K., Finke, J. B., Dziobek, I., & Otte, C. (2018). The role of physiological arousal for self-reported emotional empathy. *Autonomic Neuroscience*, *214*, 9-14.

Di Martino, A., Shehzad, Z., Kelly, C., Roy, A. K., Gee, D. G., Uddin, L. Q., ... & Milham, M. P. (2009). Relationship between cingulo-insular functional connectivity and autistic traits in neurotypical adults. *American Journal of Psychiatry*, *166*(8), 891-899.

Di Tella, M., Adenzato, M., Catmur, C., Miti, F., Castelli, L., & Ardito, R. B. (2020). The role of alexithymia in social cognition: Evidence from a non-clinical population. *Journal of Affective Disorders*, *273*, 482-492.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., ... & Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS computational biology*, *10*(9), e1003810.

Dietel, F. A., Möllmann, A., Bürkner, P. C., Wilhelm, S., & Buhlmann, U. (2021). Interpretation bias across body dysmorphic, social anxiety and generalized anxiety disorder—A multilevel, diffusion model account. *Cognitive Therapy and Research*, 1-15.

Dillon, D. G., Lazarov, A., Dolan, S., Bar-Haim, Y., Pizzagalli, D. A., & Schneier, F. R. (2022). Fast evidence accumulation in social anxiety disorder enhances decision making in a probabilistic reward task. *Emotion*, *22*(1), 1.

Domschke, K., Stevens, S., Pfleiderer, B., & Gerlach, A. L. (2010). Interoceptive sensitivity in anxiety and anxiety disorders: an overview and integration of neurobiological findings. *Clinical psychology review*, *30*(1), 1-11.

Driskill, C. M., Childs, J. E., Itmer, B., Rajput, J. S., & Kroener, S. (2022). Acute vagus nerve stimulation facilitates short term memory and cognitive flexibility in rats. *Brain sciences*, *12*(9), 1137.

Du, S., & Martinez, A. M. (2011). The resolution of facial expressions of emotion. *Journal of Vision*, *11*(13), 24-24.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., ... & Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. Psychological science, 21(12), 1835-1844.

Dunn, B. D., Evans, D., Makarova, D., White, J., & Clark, L. (2012). Gut feelings and the reaction to perceived inequity: The interplay between bodily responses, regulation, and perception shapes the rejection of unfair offers on the ultimatum game*. Cognitive, Affective, & Behavioral Neuroscience, 12*, 419–429.

Ebisch, S. J., Gallese, V., Willems, R. M., Mantini, D., Groen, W. B., Romani, G. L., ... & Bekkering, H. (2011). Altered intrinsic functional connectivity of anterior and posterior insula regions in high-functioning participants with autism spectrum disorder. *Human brain mapping*, *32*(7), 1013-1028.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior research methods, 42, 351-362.

Eilam-Stock, T., Xu, P., Cao, M., Gu, X., Van Dam, N. T., Anagnostou, E., ... & Fan, J. (2014). Abnormal autonomic and associated brain activities during rest in autism spectrum disorder. *Brain*, *137*(1), 153-171.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and emotion*, *14*(2), 131-149.

Eisenberg, N., Fabes, R. A., Miller, P. A., Fultz, J., Shell, R., Mathy, R. M., & Reno, R. R. (1989). Relaton of sympathy and personal distress to prosocial behavior: A multmethod study. Journal of Personality and Social Psychology, 57(1), 55-66.

Engelen, T., Buot, A., Grèzes, J., & Tallon-Baudry, C. (2023). Whose emotion is it? Perspective matters to understand brain-body interactions in emotions. *NeuroImage*, *268*, 119867.

Englis, B. G., Vaughan, K. B., & Lanzetta, J. T. (1982). Conditioning of counter-empathetic emotional responses. *Journal of Experimental Social Psychology*, *18*(4), 375-391.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, *4*, 215.

Feldman, M. J., Jolink, T. A., Alvarez, G. M., Fendinger, N. J., Gaudier-Diaz, M. M., Lindquist, K. A., & Muscatell, K. A. (2023). The roles of inflammation, affect, and interoception in predicting social perception. *Brain, Behavior, and Immunity*, *112*, 246-253.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature human behaviour*, *3*(5), 426-435.

FeldmanHall, O., Glimcher, P., Baker, A. L., & Phelps, E. A. (2016). Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk. *Journal of Experimental Psychology: General*, *145*(10), 1255.

Fergus, T. A., & Carleton, R. N. (2016). Intolerance of uncertainty and attentional networks: Unique associations with alerting. *Journal of Anxiety Disorders*, *41*, 59-64.

Folz, J., Fiacchino, D., Nikolić, M., van Steenbergen, H., & Kret, M. E. (2022). Reading Your Emotions in My Physiology? Reliable Emotion Interpretations in Absence of a Robust Physiological Resonance. Affective science, 3(2), 480–497.

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*, *19*(1), 3-28.

Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., ... & Stephan, K. E. (2021). TAPAS: an open-source software package for translational neuromodeling and computational psychiatry. *Frontiers in psychiatry*, *12*, 680811.

Freeston, M., & Komes, J. (2023). Revisiting uncertainty as a felt sense of unsafety: The somatic error theory of intolerance of uncertainty. *Journal of Behavior Therapy and Experimental Psychiatry*, *79*, 101827.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148-158.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. Philosophical transactions of the Royal Society B: Biological sciences, 364(1521), 1211-1221.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862-879.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, *29*(1), 1-49.

Fuchs, T. (2013). Temporality and psychopathology. *Phenomenology and the cognitive sciences*, *12*(1), 75-104.

Fuchs, T. (2017). Intercorporeality and interaffectivity. *Intercorporeality: Emerging socialities in interaction*, *7853*, 3-23.

Fuchs, T., & Koch, S. C. (2014). Embodied affectivity: on moving and being moved. *Frontiers in psychology*, *5*, 508.

Fukushima, H., Terasawa, Y., & Umeda, S. (2011). Association between interoception and empathy: evidence from heartbeat-evoked brain potential. *International Journal of Psychophysiology*, *79*(2), 259-265.

Füstös, J., Gramann, K., Herbert, B. M., & Pollatos, O. (2013). On the embodiment of emotion regulation: interoceptive awareness facilitates reappraisal. *Social cognitive and affective neuroscience*, *8*(8), 911-917.

Gaebler, M., Daniels, J. K., Lamke, J. P., Fydrich, T., & Walter, H. (2013). Heart rate variability and its neural correlates during emotional face processing in social anxiety disorder. *Biological psychology*, *94*(2), 319-330.

Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction?. *Journal of consciousness studies*, *8*(5-6), 83-108.

Gallagher, S. (2007). Simulation trouble. *Social neuroscience*, *2*(3-4), 353-365.

Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, *195*(6), 2627-2648.

Gallagher, S., & Varga, S. (2014). Social constraints on the direct perception of emotions and intentions. Topoi, 33(1), 185-199.

Gallagher, S., & Varga, S. (2015). Social cognition and psychopathology: a critical overview. *World Psychiatry*, *14*(1), 5-14.

Gallese, V. (2009). Mirror neurons, embodied simulation, and the neural basis of social identification. *Psychoanalytic dialogues*, *19*(5), 519-536.

Gallese, V. (2014). Bodily selves in relation: embodied simulation as second-person perspective on intersubjectivity. *Philosophical transactions of the royal society B: biological sciences*, *369*(1644), 20130177.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493-501.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in cognitive sciences*, *8*(9), 396-403.

Gangopadhyay, N. (2014). Introduction: embodiment and empathy, current debates in social cognition. *Topoi*, *33*(1), 117-127.

Garcia, J., Kovner, R., & Green, K. F. (1970). Cue properties vs palatability of flavors in avoidance learning. *Psychonomic Science*, *20*(5), 313-314.

Garfinkel, S. N., & Critchley, H. D. (2016). Threat and the body: how the heart supports fear processing. *Trends in cognitive sciences*, *20*(1), 34-46.

Garfinkel, S. N., Barrett, A. B., Minati, L., Dolan, R. J., Seth, A. K., & Critchley, H. D. (2013). What the heart forgets: Cardiac timing influences memory for words and is modulated by metacognition and interoceptive sensitivity. *Psychophysiology*, *50*(6), 505-512.

Garfinkel, S. N., Minati, L., Gray, M. A., Seth, A. K., Dolan, R. J., & Critchley, H. D. (2014). Fear from the heart: sensitivity to fear stimuli depends on individual heartbeats. *Journal of Neuroscience*, *34*(19), 6573-6582.

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological psychology*, *104*, 65-74.

Gendron, M., & Barrett, L. F. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, *1*(4), 316-339.

Gendron, M., & Barrett, L. F. (2018). Emotion perception as conceptual synchrony. *Emotion Review*, *10*(2), 101-110.

Gentsch, A., Sel, A., Marshall, A. C., & Schütz-Bosbach, S. (2019). Affective interoceptive inference: Evidence from heart-beat evoked brain potentials. *Human Brain Mapping*, *40*(1), 20-33.

Godinić, D., & Obrenovic, B. (2020). Effects of economic uncertainty on mental health in the COVID-19 pandemic context: social identity disturbance, job uncertainty and psychological well-being model.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gordon, R. M. (1995). Simulation without introspection or inference from Me to You. In M. Davies & T. Stone (Eds.), *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

Gray, M. A., Beacher, F. D., Minati, L., Nagai, Y., Kemp, A. H., Harrison, N. A., & Critchley, H. D. (2012). Emotional appraisal is influenced by cardiac afferent information. *Emotion*, *12*(1), 180.

Green, S. A., Hernandez, L., Bookheimer, S. Y., & Dapretto, M. (2016). Salience network connectivity in autism is related to brain and behavioral markers of sensory overresponsivity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(7), 618-626.

Gross, J. J., John, O. P., & Richards, J. M. (2000). The dissociation of emotion expression from emotion experience: A personality perspective. *Personality and Social Psychology Bulletin*, *26*(6), 712-726.

Grynberg, D., & Pollatos, O. (2015). Perceiving one's body shapes empathy. *Physiology & behavior*, *140*, 54-60.

Gu, X., & FitzGerald, T. H. (2014). Interoceptive inference: homeostasis and decision-making. *Trends Cogn Sci*, *18*(6), 269-70.

Gu, X., FitzGerald, T. H., & Friston, K. J. (2019). Modeling subjective belief states in computational psychiatry: interoceptive inference as a candidate framework. *Psychopharmacology*, *236*, 2405-2412.

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, *521*(15), 3371-3388.

Gu, Y., Gu, S., Lei, Y., & Li, H. (2020). From uncertainty to anxiety: How uncertainty fuels anxiety in a process mediated by intolerance of uncertainty. *Neural Plasticity*, *2020*(1), 8866386.

Guo, K., Soornack, Y., & Settle, R. (2019). Expression-dependent susceptibility to face distortions in processing of facial expressions of emotion. *Vision research*, *157*, 112-122.

Gupta, A., Bansal, R., Alashwal, H., Kacar, A. S., Balci, F., & Moustafa, A. A. (2022). Neural substrates of the drift-diffusion model in brain disorders. *Frontiers in computational neuroscience*, *15*, 678232.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), 367-388.

Han, W., Tellez, L. A., Perkins, M. H., Perez, I. O., Qu, T., Ferreira, J., ... & de Araujo, I. E. (2018). A neural circuit for gut-induced reward. *Cell*, *175*(3), 665-678.

Harricharan, S., Nicholson, A. A., Thome, J., Densmore, M., McKinnon, M. C., Théberge, J., ... & Lanius, R. A. (2020). PTSD and its dissociative subtype through the lens of the insula: Anterior and posterior insula resting-state functional connectivity and its predictive validity using machine learning. *Psychophysiology*, *57*(1), e13472.

Harrison, O. K., Garfinkel, S. N., Marlow, L., Finnegan, S. L., Marino, S., Köchli, L., ... & Fleming, S. M. (2021). The Filter Detection Task for measurement of breathing-related interoception and metacognition. *Biological Psychology*, *165*, 108185.

Harrison, O. K., Köchli, L., Marino, S., Luechinger, R., Hennel, F., Brand, K., ... & Stephan, K. E. (2021). Interoception of breathing and its relationship with anxiety. *Neuron*, *109*(24), 4080-4093.

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, *5*(1), 60-65.

Hazem, N., Beaurenaut, M., George, N., & Conty, L. (2018). Social contact enhances bodily self-awareness. *Scientific reports*, *8*(1), 4195.

Herbert, B. M., Herbert, C., & Pollatos, O. (2011). On the relationship between interoceptive awareness and alexithymia: is interoceptive awareness related to emotional awareness?. *Journal of personality*, *79*(5), 1149-1175.

Herman, A. M., & Tsakiris, M. (2021). The impact of cardiac afferent signaling and interoceptive abilities on passive information sampling. *International Journal of Psychophysiology*, *162*, 104-111.

Herman, A. M., Esposito, G., & Tsakiris, M. (2021). Body in the face of uncertainty: The role of autonomic arousal and interoception in decision-making under risk and ambiguity. *Psychophysiology*, *58*(8), e13840.

Heyes, C. (2010). Where do mirror neurons come from?. *Neuroscience & Biobehavioral Reviews*, *34*(4), 575-583.

Heyes, C. (2011). Automatic imitation. *Psychological bulletin*, *137*(3), 463.

Heyes, C. (2018). Empathy is not in our genes. Neuroscience & Biobehavioral Reviews, 95, 499-507.

Heyes, C., & Catmur, C. (2022). What happened to mirror neurons?. *Perspectives on Psychological Science*, *17*(1), 153-168.

Hickman, L., Seyedsalehi, A., Cook, J. L., Bird, G., & Murphy, J. (2020). The relationship between heartbeat counting and heartbeat discrimination: A meta-analysis. *Biological Psychology*, *156*, 107949.

Hirsch, C., Meynen, T., & Clark, D. (2004). Negative self-imagery in social anxiety contaminates social interactions. *Memory*, *12*(4), 496-506.

Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: a framework for understanding uncertainty-related anxiety. *Psychological review*, *119*(2), 304.

Hodgson, A. R., Freeston, M. H., Honey, E., & Rodgers, J. (2017). Facing the unknown: Intolerance of uncertainty in children with autism spectrum disorder. *Journal of applied research in intellectual disabilities*, *30*(2), 336-344.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, *47*, 75-85.

Hollocks, M. J., Lerh, J. W., Magiati, I., Meiser-Stedman, R., & Brugha, T. S. (2019). Anxiety and depression in adults with autism spectrum disorder: a systematic review and meta-analysis. *Psychological medicine*, *49*(4), 559-572.

Holzer, P. (2017). Interoception and gut feelings: Unconscious body signals' impact on brain function, behavior and belief processes. *Processes of believing: The acquisition, maintenance, and change in creditions*, 435-442.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Hübner, A. M., Trempler, I., & Schubotz, R. I. (2022). Interindividual differences in interoception modulate behavior and brain responses in emotional inference. *NeuroImage*, *261*, 119524.

Hübner, A. M., Trempler, I., Gietmann, C., & Schubotz, R. I. (2021). Interoceptive sensibility predicts the ability to infer others' emotional states. *Plos one*, *16*(10), e0258089.

Hunt, P. A., Denieffe, S., & Gooney, M. (2017). Burnout and its relationship to empathy in nursing: a review of the literature. *Journal of Research in Nursing*, *22*(1-2), 7-22.

Hwang, Y. I., Arnold, S., Srasuebkul, P., & Trollor, J. (2020). Understanding anxiety in adults on the autism spectrum: An investigation of its relationship with intolerance of uncertainty, sensory sensitivities and repetitive behaviours. *Autism*, *24*(2), 411-422.

Isomura, T., Parr, T., & Friston, K. (2019). Bayesian filtering with multiple internal models: toward a theory of social intelligence. *Neural computation*, *31*(12), 2390-2431.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, *24*(2), 187-192.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241-7244.

Jacoby, R. J., Abramowitz, J. S., Buck, B. E., & Fabricant, L. E. (2014). How is the Beads Task related to intolerance of uncertainty in anxiety disorders?. *Journal of Anxiety Disorders*, *28*(6), 495-503.

Jazaieri, H., Morrison, A. S., Goldin, P. R., & Gross, J. J. (2015). The role of emotion and emotion regulation in social anxiety disorder. *Current psychiatry reports*, *17*, 1-9.

Jenkinson, Richard, Elizabeth Milne, and Andrew Thompson. "The relationship between intolerance of uncertainty and anxiety in autism: A systematic literature review and meta-analysis." *Autism* 24.8 (2020): 1933-1944.

Jin, Y., & Kong, J. (2017). Transcutaneous vagus nerve stimulation: a promising method for treatment of autism spectrum disorders. *Frontiers in Neuroscience*, *10*, 609.

Johnson, R. L., & Wilson, C. G. (2018). A review of vagus nerve stimulation as a therapeutic intervention. *Journal of inflammation research*, 203-213.

Johnston, P. J., McCabe, K., & Schall, U. (2003). Differential susceptibility to performance degradation across categories of facial emotion—a model confirmation. *Biological Psychology*, *63*(1), 45-58.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. (2017). Social learning through prediction error in the brain. *NPJ science of learning*, *2*(1), 8.

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive ability predicts survival on a London trading floor. *Scientific reports*, *6*(1), 32986.

Karukivi, M., Hautala, L., Kaleva, O., Haapasalo-Pesu, K. M., Liuksila, P. R., Joukamaa, M., & Saarijärvi, S. (2010). Alexithymia is associated with anxiety among adolescents. *Journal of affective disorders*, *125*(1-3), 383-387.

Katkin, E. S., Wiens, S., & Öhman, A. (2001). Nonconscious fear conditioning, visceral perception, and the development of gut feelings. *Psychological Science*, *12*(5), 366-370.

Keysers, C., & Gazzola, V. (2014). Hebbian learning and predictive mirror neurons for actions, sensations and emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1644), 20130175.

Keysers, C., Silani, G., & Gazzola, V. (2024). Predictive coding for the actions and emotions of others and its deficits in autism spectrum disorders. *Neuroscience & Biobehavioral Reviews*, 105877.

Kim, A. Y., Marduy, A., de Melo, P. S., Gianlorenco, A. C., Kim, C. K., Choi, H., ... & Fregni, F. (2022). Safety of transcutaneous auricular vagus nerve stimulation (taVNS): A systematic review and meta-analysis. *Scientific reports*, *12*(1), 22055.

Kinnaird, E., Stewart, C., & Tchanturia, K. (2019). Investigating alexithymia in autism: A systematic review and meta-analysis. *European Psychiatry*, *55*, 80-89.

Knight, L. K., Stoica, T., Fogleman, N. D., & Depue, B. E. (2019). Convergent neural correlates of empathy and anxiety during socioemotional processing. *Frontiers in human neuroscience*, *13*, 94.

Kogan, A., Oveis, C., Carr, E. W., Gruber, J., Mauss, I. B., Shallcross, A., ... & Keltner, D. (2014). Vagal activity is quadratically related to prosocial traits, prosocial emotions, and observer perceptions of prosociality. *Journal of personality and social psychology*, *107*(6), 1051.

Kolassa, I. T., Kolassa, S., Bergmann, S., Lauche, R., Dilger, S., Miltner, W. H., & Musial, F. (2009). Interpretive bias in social phobia: An ERP study with morphed emotional schematic faces. *Cognition and emotion*, *23*(1), 69-95.

Kozakevich Arbel, E., Shamay-Tsoory, S. G., & Hertz, U. (2021). Adaptive empathy: Empathic response selection as a dynamic, feedback-based learning process. Frontiers in psychiatry, 1248.

Kreuzer, P. M., Landgrebe, M., Husser, O., Resch, M., Schecklmann, M., Geisreiter, F., ... & Langguth, B. (2012). Transcutaneous vagus nerve stimulation: retrospective assessment of cardiac safety in a pilot study. *Frontiers in psychiatry*, *3*, 70.

Kühnel, A., Teckentrup, V., Neuser, M. P., Huys, Q. J., Burrasch, C., Walter, M., & Kroemer, N. B. (2020). Stimulation of the vagus nerve reduces learning in a go/no-go reinforcement learning task. *European Neuropsychopharmacology*, *35*, 17-29.

Ladouceur, R., Talbot, F., & Dugas, M. J. (1997). Behavioral expressions of intolerance of uncertainty in worry: Experimental findings. *Behavior modification*, *21*(3), 355-371.

Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). Anxiety impedes adaptive social learning under uncertainty. *Psychological science*, *31*(5), 592-603.

Lamm, C., Porges, E. C., Cacioppo, J. T., & Decety, J. (2008). Perspective taking is associated with specific facial responses during empathy for pain. Brain Research, 1227, 153–161.

Lapidus, R. C., Puhl, M., Kuplicki, R., Stewart, J. L., Paulus, M. P., Rhudy, J. L., ... & Tulsa 1000 Investigators. (2020). Heightened affective response to perturbation of respiratory but not pain signals in eating, mood, and anxiety disorders. *PLoS One*, *15*(7), e0235346.

Lawson, R. P., Bisby, J., Nord, C. L., Burgess, N., & Rees, G. (2021). The computational, pharmacological, and physiological determinants of sensory learning under uncertainty. *Current Biology*, *31*(1), 163-172.

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature neuroscience*, *20*(9), 1293-1299.

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. Frontiers in human neuroscience, 8, 302.

Lee, S. J., Lee, M., Kim, H. B., & Huh, H. J. (2024). The Relationship Between Interoceptive Awareness, Emotion Regulation and Clinical Symptoms Severity of Depression, Anxiety and Somatization. *Psychiatry Investigation*, *21*(3), 255.

Lehmann, K., Bolis, D., Friston, K. J., Schilbach, L., Ramstead, M. J., & Kanske, P. (2024). An active-inference approach to second-person neuroscience. *Perspectives on Psychological Science*, *19*(6), 931-951.

Lenggenhager, B., Azevedo, R. T., Mancini, A., & Aglioti, S. M. (2013). Listening to your heart and feeling yourself: Effects of exposure to interoceptive signals during the ultimatum game. *Experimental Brain Research, 230(2),* 233–241.

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior research methods*, *49*, 513-537.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*, 775-784.

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in human neuroscience*, *7*, 547.

Luft, C. D. B., & Bhattacharya, J. (2015). Aroused with heart: Modulation of heartbeat evoked potential by arousal induction and its oscillatory correlates. *Scientific reports*, *5*(1), 15717.

Maisel, M.E., Stephenson, K.G., South, M., Rodgers, J., Freeston, M.H., & Gaigg, S.B. (2016). Modeling the cognitive mechanisms linking autism symptoms and anxiety in adults. Journal of Abnormal Psychology, 125, 692–703.

Maister, L., Hodossy, L., & Tsakiris, M. (2017). You fill my heart: Looking at one’s partner increases interoceptive accuracy. *Psychology of Consciousness: Theory, Research, and Practice*, *4*(2), 248.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Maraver, M. J., Steenbergen, L., Hossein, R., Actis-Grosso, R., Ricciardelli, P., Hommel, B., & Colzato, L. S. (2020). Transcutaneous vagus nerve stimulation modulates attentional resource deployment towards social cues. *Neuropsychologia*, *143*, 107465.

Marchi, F., & Newen, A. (2015). Cognitive penetrability and emotion recognition in human facial expressions. *Frontiers in psychology*, *6*, 828.

Marshall, A. C., Gentsch, A., Schröder, L., & Schütz-Bosbach, S. (2018). Cardiac interoceptive learning is modulated by emotional valence perceived from facial expressions. *Social cognitive and affective neuroscience*, *13*(7), 677-686.

Marshall, L., Mathys, C., Ruge, D., De Berker, A. O., Dayan, P., Stephan, K. E., & Bestmann, S. (2016). Pharmacological fingerprints of contextual uncertainty. *PLoS Biology*, *14*(11), e1002575.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. Frontiers in human neuroscience, 8, 825

Mazefsky, C. A., Herrington, J., Siegel, M., Scarpa, A., Maddox, B. B., Scahill, L., & White, S. W. (2013). The role of emotion regulation in autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(7), 679-688.

McGovern, H. T., & Otten, M. (2024). Priors and prejudice: hierarchical predictive processing in intergroup perception. *Frontiers in Psychology*, *15*, 1386370.

Medford, N., Quadt, L., & Critchley, H. (2024). Interoception and psychopathology. In *Phenomenological Neuropsychiatry: How Patient Experience Bridges the Clinic with Clinical Neuroscience* (pp. 155-174). Cham: Springer International Publishing.

Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The multidimensional assessment of interoceptive awareness, version 2 (MAIA-2). PloS one, 13(12), e0208034.

Melloni, M., Lopez, V., & Ibanez, A. (2014). Empathy and contextual social cognition. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 407-425.

Ming, X., Julu, P. O., Brimacombe, M., Connor, S., & Daniels, M. L. (2005). Reduced cardiac parasympathetic activity in children with autism. *Brain and Development*, *27*(7), 509-516.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72-80.

Morgado, P., Sousa, N., & Cerqueira, J. J. (2015). The impact of stress in decision making in the context of uncertainty. *Journal of Neuroscience Research*, *93*(6), 839-847.

Moutoussis, M., Bentall, R. P., El-Deredy, W., & Dayan, P. (2011). Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. *Cognitive neuropsychiatry*, *16*(5), 422-447.

Moutoussis, M., Dolan, R. J., & Friston, K. J. (2014). *The computational anatomy of psychosis*. *Frontiers in Psychology*, 5, 130.

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2018). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry (Cambridge, Mass.)*, *2*, 50.

Mul, C. L., Stagg, S. D., Herbelin, B., & Aspell, J. E. (2018). The feeling of me feeling for you: Interoception, alexithymia and empathy in autism. *Journal of autism and developmental disorders*, *48*, 2953-2967.

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, *13*, 1039172.

Nair, T. K., Waslin, S. M., Rodrigues, G. A., Datta, S., Moore, M. T., & Brumariu, L. E. (2024). A meta-analytic review of the relations between anxiety and empathy. *Journal of anxiety disorders*, *101*, 102795.

Neuser, M. P., Teckentrup, V., Kühnel, A., Hallschmid, M., Walter, M., & Kroemer, N. B. (2020). Vagus nerve stimulation boosts the drive to work for rewards. *Nature communications*, *11*(1), 3555.

Nicholson, T., Williams, D. M., Grainger, C., Lind, S. E., & Carruthers, P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder. *Consciousness and Cognition*, *70*, 11-24.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, *9*(3), 184-211.

Nitschke, J. P., & Bartz, J. A. (2023). The association between acute stress & empathy: A systematic literature review. *Neuroscience & Biobehavioral Reviews*, *144*, 105003.

Noel, J. P., Lytle, M., Cascio, C., & Wallace, M. T. (2018). Disrupted integration of exteroceptive and interoceptive signaling in autism spectrum disorder. *Autism Research*, *11*(1), 194-205.

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of mathematical psychology*, *76*, 117-130.

Oberman, L. M., Winkielman, P., & Ramachandran, V. S. (2007). Face to face: Blocking facial mimicry can selectively impair recognition of emotional expressions. *Social neuroscience*, *2*(3-4), 167-178.

Oehrn, C. R., Molitor, L., Krause, K., Niehaus, H., Schmidt, L., Hakel, L., ... & Weber, I. (2022). Non-invasive vagus nerve stimulation in epilepsy patients enhances cooperative behavior in the prisoner's dilemma task. *Scientific reports*, *12*(1), 10255.

Olatunji, B. O., Deacon, B. J., Abramowitz, J. S., & Valentiner, D. P. (2007). Body vigilance in nonclinical and anxiety disorder samples: structure, correlates, and prediction of health concerns. *Behavior Therapy*, *38*(4), 392-401.

Oliveira-Silva, P., & Gonçalves, O. F. (2011). Responding empathically: A question of heart, not a question of skin. *Applied psychophysiology and biofeedback*, *36*, 201-207.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197-212.

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

Otten, M., Seth, A. K., & Pinto, Y. (2017). A social Bayesian brain: How social knowledge can shape visual perception. *Brain and cognition*, *112*, 69-77.

Ozsivadjian, A., Hollocks, M. J., Magiati, I., Happé, F., Baird, G., & Absoud, M. (2021). Is cognitive inflexibility a missing link? The role of cognitive inflexibility, alexithymia and intolerance of uncertainty in externalising and internalising behaviours in young people with autism spectrum disorder. *Journal of child psychology and psychiatry*, *62*(6), 715-724.

Paciorek, A., & Skora, L. (2020). Vagus nerve stimulation as a gateway to interoception. *Frontiers in Psychology*, *11*, 1659.

Palmer, C. E., & Tsakiris, M. (2018). Going at the heart of social cognition: is there a role for interoception in self-other distinction?. *Current opinion in psychology*, *24*, 21-26.

Palmer, C. J., Lawson, R. P., & Hohwy, J. (2017). Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychological bulletin*, *143*(5), 521.

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition*, *36*, 376-389.

Palser, E. R., Fotopoulou, A., Pellicano, E., & Kilner, J. M. (2018). The link between interoceptive processing and anxiety in children diagnosed with autism spectrum disorder: Extending adult findings into a developmental sample. *Biological Psychology*, *136*, 13-21.

Palser, E. R., Galvez-Pol, A., Palmer, C. E., Hannah, R., Fotopoulou, A., Pellicano, E., & Kilner, J. M. (2021). Reduced differentiation of emotion-associated bodily sensations in autism. *Autism*, *25*(5), 1321-1334.

Palser, E. R., Palmer, C. E., Galvez-Pol, A., Hannah, R., Fotopoulou, A., & Kilner, J. M. (2018). Alexithymia mediates the relationship between interoceptive sensibility and anxiety. *PloS one*, *13*(9), e0203212.

Pandey, R., Saxena, P., & Dubey, A. (2011). Emotion regulation difficulties in alexithymia and mental health. Europe's journal of psychology, 7(4), 604-623.

Park, H. D., & Blanke, O. (2019). Heartbeat-evoked cortical responses: Underlying mechanisms, functional roles, and methodological considerations. *Neuroimage*, *197*, 502-511.

Park, H. D., & Tallon-Baudry, C. (2014). The neural subjective frame: from bodily signals to perceptual consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130208.

Park, H. D., Correia, S., Ducorps, A., & Tallon-Baudry, C. (2014). Spontaneous fluctuations in neural responses to heartbeats predict visual detection. *Nature neuroscience*, *17*(4), 612-618.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Paulus, M. P., & Stein, M. B. (2010). Interoception in anxiety and depression. *Brain structure and Function*, *214*, 451-463.

Paulus, M. P., Feinstein, J. S., & Khalsa, S. S. (2019). An active inference approach to interoceptive psychopathology. *Annual review of clinical psychology*, *15*(1), 97-122.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* Oxford University Press.

Pavy, F., Zaman, J., Van den Noortgate, W., Scarpa, A., von Leupoldt, A., & Torta, D. M. (2024). The effect of unpredictability on the perception of pain: a systematic review and meta-analysis. *Pain*, *165*(8), 1702–1718.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, *24*, 1234-1251.

Peña, D. F., Childs, J. E., Willett, S., Vital, A., McIntyre, C. K., & Kroener, S. (2014). Vagus nerve stimulation enhances extinction of conditioned fear and modulates plasticity in the pathway from the ventromedial prefrontal cortex to the amygdala. *Frontiers in behavioral neuroscience*, *8*, 327.

Petzschner, F. H., Weber, L. A., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: toward a joint framework for differential diagnosis. *Biological psychiatry*, *82*(6), 421-430.

Peuker, E. T., & Filler, T. J. (2002). The nerve supply of the human auricle. *Clinical Anatomy, 15(1*), 35–37.

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

Pfeifer, G., Garfinkel, S. N., van Praag, C. D. G., Sahota, K., Betka, S., & Critchley, H. D. (2017). Feedback from the heart: Emotional learning and memory is controlled by cardiac cycle, interoceptive accuracy and personality. *Biological Psychology*, *126*, 19-29.

Pfeiffer, C., & De Lucia, M. (2017). Cardio-audio synchronization drives neural surprise response. *Scientific reports*, *7*(1), 14842.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational behavior and human decision processes*, *68*(3), 208-224.

Pinna, T., & Edwards, D. J. (2020). A systematic review of associations between interoception, vagal tone, and emotional regulation: Potential applications for mental health, wellbeing, psychological flexibility, and chronic conditions. *Frontiers in psychology*, *11*, 1792.

Pollatos, O., Traut-Mattausch, E., & Schandry, R. (2009). Differential effects of anxiety and depression on interoceptive accuracy. *Depression and anxiety*, *26*(2), 167-173.

Pollatos, O., Traut-Mattausch, E., Schroeder, H., & Schandry, R. (2007). Interoceptive awareness mediates the relationship between anxiety and the intensity of unpleasant feelings. *Journal of anxiety disorders*, *21*(7), 931-943.

Poppa, T., Benschop, L., Horczak, P., Vanderhasselt, M. A., Carrette, E., Bechara, A., ... & Vonck, K. (2022). Auricular transcutaneous vagus nerve stimulation modulates the heart-evoked potential. *Brain Stimulation*, *15*(1), 260-269.

Porges, S. W. (1993). Body perception questionnaire. Laboratory of Developmental Assessment: University of Maryland

Porges, S. W. (2007). The polyvagal perspective. *Biological psychology*, *74*(2), 116-143.

Porges, S. W. (2009). The polyvagal theory: new insights into adaptive reactions of the autonomic nervous system. *Cleveland Clinic journal of medicine*, *76*(Suppl 2), S86.

Preece, D. A., Mehta, A., Petrova, K., Sikka, P., Bjureberg, J., Becerra, R., & Gross, J. J. (2023). Alexithymia and emotion regulation. *Journal of affective disorders*, *324*, 232-238.

Prochazkova, E., & Kret, M. E. (2017). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. Neuroscience & Biobehavioral Reviews, 80, 99-114.

Proff, I., Williams, G. L., Quadt, L., & Garfinkel, S. N. (2022). Sensory processing in autism across exteroceptive and interoceptive domains. *Psychology & Neuroscience*, *15*(2), 105.

Pulcu, E., & Browning, M. (2019). The misestimation of uncertainty in affective disorders. *Trends in cognitive sciences*, *23*(10), 865-875.

Puścian, A., Bryksa, A., Kondrakiewicz, L., Kostecki, M., Winiarski, M., & Knapska, E. (2022). Ability to share emotions of others as a foundation of social learning. *Neuroscience & Biobehavioral Reviews*, *132*, 23-36.

Quadt, L. (2017). *Action-oriented predictive processing and social cognition*. Johannes Gutenberg-Universität Mainz.

Quadt, L., Critchley, H. D., Garfinkel, S. N., Tsakiris, M., & De Preester, H. (2018). Interoception and emotion: Shared mechanisms and clinical implications. *The interoceptive mind: From homeostasis to awareness*, *123*.

Quattrocki, E., & Friston, K. (2014). Autism, oxytocin and interoception. Neuroscience & Biobehavioral Reviews, 47, 410-430

Quigley, K. S., Kanoski, S., Grill, W. M., Barrett, L. F., & Tsakiris, M. (2021). Functions of interoception: From energy regulation to experience of the self. *Trends in neurosciences*, *44*(1), 29-38.

Quintana, D. S., Guastella, A. J., Outhred, T., Hickie, I. B., & Kemp, A. H. (2012). Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition. *International journal of psychophysiology*, *86*(2), 168-172.

Radoman, M., Phan, K. L., & Gorka, S. M. (2019). Neural correlates of predictable and unpredictable threat in internalizing psychopathology. *Neuroscience letters*, *701*, 193-201.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

Razran, G. (1961). The observable and the inferable conscious in current Soviet psychophysiology: Interoceptive conditioning, semantic conditioning, and the orienting reflex*. Psychological Review, 68(2),* 81–147.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495-505.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Current research and theory, 64-99.

Richter, F., García, A. M., Rodriguez Arriagada, N., Yoris, A., Birba, A., Huepe, D., ... & Sedeño, L. (2021). Behavioral and neurophysiological signatures of interoceptive enhancements following vagus nerve stimulation. *Human brain mapping*, *42*(5), 1227-1242.

Riva, F., Triscoli, C., Lamm, C., Carnaghi, A., & Silani, G. (2016). Emotional egocentricity bias across the life-span*. Frontiers in aging neuroscience, 8, 74*.

Roosevelt, R. W., Smith, D. C., Clough, R. W., Jensen, R. A., & Browning, R. A. (2006). Increased extracellular concentrations of norepinephrine in cortex and hippocampus following vagus nerve stimulation in the rat. *Brain research*, *1119*(1), 124-132.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36.

Sahib, A., Chen, J., Cárdenas, D., & Calear, A. L. (2023). Intolerance of uncertainty and emotion regulation: A meta-analytic and systematic review. *Clinical Psychology Review*, *101*, 102270.

Saini, F., Ponzo, S., Silvestrin, F., Fotopoulou, A., & David, A. S. (2022). Depersonalization disorder as a systematic downregulation of interoceptive signals. *Scientific Reports*, *12*(1), 22123.

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model. *PLoS computational biology*, *15*(1), e1006267.

Sandhu, T. R., Xiao, B., & Lawson, R. P. (2023). Transdiagnostic computations of uncertainty: towards a new lens on intolerance of uncertainty. *Neuroscience & Biobehavioral Reviews*, *148*, 105123.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755-1758.

Sato, W., & Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, *104*(1), 1-18.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current opinion in Psychology*, *17*, 15-21.

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology*, *18*(4), 483-488.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological bulletin*, *147*(3), 293–327.

Seeley, W. W. (2019). The salience network: a neural system for perceiving and responding to homeostatic demands. *Journal of Neuroscience*, *39*(50), 9878-9882.

Sel, A., Azevedo, R. T., & Tsakiris, M. (2017). Heartfelt self: cardio-visual integration affects self-face recognition and interoceptive cortical processing. *Cerebral Cortex*, *27*(11), 5144-5155.

Sellaro, R., de Gelder, B., Finisguerra, A., & Colzato, L. S. (2018). Transcutaneous vagus nerve stimulation (tVNS) enhances recognition of emotions in faces but not bodies. *Cortex*, *99*, 213-223.

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

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1708), 20160007.

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in cognitive sciences*, *22*(11), 969-981.

Shah, P., Catmur, C., & Bird, G. (2017). From heart to mind: Linking interoception, emotion, and theory of mind. *Cortex, 93*, 220-223.

Shamay-Tsoory, S. G., & Hertz, U. (2022). Adaptive empathy: a model for learning empathic responses in response to feedback. *Perspectives on Psychological Science*, *17*(4), 1008-1023.

Shanton, K., & Goldman, A. (2010). Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 527-538.

Sheppes, G., Scheibe, S., Suri, G., & Gross, J. J. (2011). Emotion-regulation choice. *Psychological science*, *22*(11), 1391-1396.

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *Journal of neuroscience*, *33*(39), 15466-15476.

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current biology*, *24*(18), R875-R878.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. Annals of the New York Academy of Sciences, 1156, 81–96.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in cognitive sciences*, *13*(8), 334-340.

Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. Science, 303(5661), 1157 1162.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science, 303(5661), 1157-1162.

Slotta, T., Witthöft, M., Gerlach, A. L., & Pohl, A. (2021). The interplay of interoceptive accuracy, facets of interoceptive sensibility, and trait anxiety: A network analysis. *Personality and Individual Differences*, *183*, 111133.

Smeets, T., Dziobek, I., & Wolf, O. T. (2009). Social cognition under stress: differential effects of stress-induced cortisol elevations in healthy young men and women. *Hormones and behavior*, *55*(4), 507-513.

Smith, B. M., Twohy, A. J., & Smith, G. S. (2020). Psychological inflexibility and intolerance of uncertainty moderate the relationship between social isolation and mental health outcomes during COVID-19. *Journal of contextual behavioral science*, *18*, 162-174.

Smith, R., Badcock, P., & Friston, K. J. (2021). Recent advances in the application of predictive coding and active inference models within clinical neuroscience. *Psychiatry and Clinical Neurosciences*, *75*(1), 3-13.

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., ... & Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS computational biology*, *16*(12), e1008484.

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R., & Phelps, E. A. (2015). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, *29*(4), 695-701.

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*(10), 635-644.

Stark, E., Stacey, J., Mandy, W., Kringelbach, M. L., & Happé, F. (2021). 'Uncertainty attunement'has explanatory value in understanding autistic anxiety. *Trends in Cognitive Sciences*, *25*(12), 1011-1012.

Steinbeis, N., & Singer, T. (2014). Projecting my envy onto you: Neurocognitive mechanisms of an offline emotional egocentricity bias. NeuroImage, 102, 370-380.

Stellar, J. E., Cohen, A., Oveis, C., & Keltner, D. (2015). Affective and physiological responses to the suffering of others: compassion and vagal activity. *Journal of personality and social psychology*, *108*(4), 572.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current opinion in neurobiology*, *25*, 85-92.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004-1017.

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology & behavior*, *106*(1), 5-15.

Stevenson, R. J., Francis, H. M., Oaten, M. J., & Schilt, R. (2018). Hippocampal dependent neuropsychological tests and their relationship to measures of cardiac and self-report interoception. *Brain and Cognition*, *123*, 23-29.

Stoica, T., & Depue, B. (2020). Shared characteristics of intrinsic connectivity networks underlying interoceptive awareness and empathy. *Frontiers in Human Neuroscience*, *14*, 571070.

Sunahara, C. S., Rosenfield, D., Alvi, T., Wallmark, Z., Lee, J., Fulford, D., & Tabak, B. A. (2022). Revisiting the association between self-reported empathy and behavioral assessments of social cognition. *Journal of Experimental Psychology: General*, *151*(12), 3304.

Surcinelli, P., Codispoti, M., Montebarocci, O., Rossi, N., & Baldaro, B. (2006). Facial emotion recognition in trait anxiety. *Journal of anxiety disorders*, *20*(1), 110-117.

Tajadura-Jiménez, A., & Tsakiris, M. (2014). Balancing the “inner” and the “outer” self: Interoceptive sensitivity modulates self–other boundaries. *Journal of Experimental Psychology: General*, *143*(2), 736.

Taylor, G. J. (2000). Recent developments in alexithymia theory and research. *The Canadian Journal of Psychiatry*, *45*(2), 134-142.

Terasawa, Y., Moriguchi, Y., Tochizawa, S., & Umeda, S. (2014). Interoceptive sensitivity predicts sensitivity to the emotions of others. *Cognition and Emotion*, *28*(8), 1435-1448.

Thibodeau, M. A., Carleton, R. N., Gómez-Pérez, L., & Asmundson, G. J. (2013). “What if I make a mistake?”: Intolerance of uncertainty is associated with poor behavioral performance. *The Journal of nervous and mental disease*, *201*(9), 760-766.

Thompson, N. M., Uusberg, A., Gross, J. J., & Chakrabarti, B. (2019). Empathy and emotion regulation: An integrative account. *Progress in brain research*, *247*, 273-304.

Tibi-Elhanany, Y., and Shamay-Tsoory, S. G. (2011). Social cognition in social anxiety: first evidence for increased empathic abilities. *Isr. J. Psychiatry Relat. Sci.* 48, 98–106.

Tomova, L., von Dawans, B., Heinrichs, M., Silani, G., & Lamm, C. (2014). Is stress affecting our ability to tune into others? Evidence for gender differences in the effects of stress on self-other distinction. *Psychoneuroendocrinology*, *43*, 95-104.

Trevisan, D. A., Mehling, W. E., & McPartland, J. C. (2021). Adaptive and maladaptive bodily awareness: Distinguishing interoceptive sensibility and interoceptive attention from anxiety-induced somatization in autism and alexithymia. *Autism research*, *14*(2), 240-247.

Trilla, I., Weigand, A., & Dziobek, I. (2021). Affective states influence emotion perception: evidence for emotional egocentricity. Psychological research, 85(3), 1005-1015.

Tsakiris, M. (2017). The multisensory basis of the self: From body to identity to others. *Quarterly journal of experimental psychology*, *70*(4), 597-609.

Tsakiris, M., & Critchley, H. (2016). Interoception beyond homeostasis: affect, cognition and mental health. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160002.

Tsakiris, M., Jiménez, A. T., & Costantini, M. (2011). Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body-representations. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1717), 2470-2476.

Vabba, A., Porciello, G., Monti, A., Panasiti, M. S., & Aglioti, S. M. (2023). A longitudinal study of interoception changes in the times of COVID-19: Effects on psychophysiological health and well-being. Heliyon, 9(4).

Van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological science*, *15*(1), 71-74.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., De-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological review*, *121*(4), 649.

Van Diest, I. (2019). Interoception, conditioning, and fear: the panic threesome. *Psychophysiology*, *56*(8), e13421.

Vazard, J. (2024). Feeling the unknown: emotions of uncertainty and their valence. *Erkenntnis*, *89*(4), 1275-1294.

Ventura-Bort, C., & Weymar, M. (2024). Transcutaneous auricular vagus nerve stimulation modulates the processing of interoceptive prediction error signals and their role in allostatic regulation. *Human Brain Mapping*, *45*(3), e26613.

Villani, V., Tsakiris, M., & Azevedo, R. T. (2019). Transcutaneous vagus nerve stimulation improves interoceptive accuracy. *Neuropsychologia*, *134*, 107201.

Von Mohr, M., Finotti, G., Ambroziak, K. B., & Tsakiris, M. (2020). Do you hear what I see? An audio-visual paradigm to assess emotional egocentricity bias. Cognition and Emotion, 34(4), 756-770.

Von Mohr, M., Finotti, G., Villani, V., & Tsakiris, M. (2021). Taking the pulse of social cognition: cardiac afferent activity and interoceptive accuracy modulate emotional egocentricity bias. Cortex, 145, 327-340.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: an R Package Providing Distribution Functions for the Wiener Diffusion Model. *R Journal*, *6*(1).

Weber, I., Niehaus, H., Krause, K., Molitor, L., Peper, M., Schmidt, L., ... & Oehrn, C. R. (2021). Trust your gut: vagal nerve stimulation in humans improves reinforcement learning. *Brain communications*, *3*(2), fcab039.

Weilenmann, S., Schnyder, U., Parkinson, B., Corda, C., Von Kaenel, R., & Pfaltz, M. C. (2018). Emotion transfer, emotion regulation, and empathy-related processes in physician-patient interactions and their association with physician well-being: a theoretical model. *Frontiers in psychiatry*, *9*, 389.

Weng, H. Y., Feldman, J. L., Leggio, L., Napadow, V., Park, J., & Price, C. J. (2021). Interventions and manipulations of interoception. *Trends in neurosciences*, *44*(1), 52-62.

Werner, N. S., & Schandry, R. (2024). The Impact of Interoception on Learning, Memory, and Decision-Making. Interoception: A Comprehensive Guide, 151-184.

Werner, N. S., Duschek, S., Mattern, M., & Schandry, R. (2009). Interoceptive sensitivity modulates anxiety during public speaking. *Journal of Psychophysiology*, *23*(2), 85-94.

Werner, N. S., Duschek, S., Mattern, M., & Schandry, R. (2009). The relationship between pain perception and interoception. *Journal of Psychophysiology*, *23*(1), 35-42.

Werner, N. S., Peres, I., Duschek, S., & Schandry, R. (2010). Implicit memory for emotional words is modulated by cardiac perception. *Biological psychology*, *85*(3), 370-376.

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of mathematical psychology*, *54*(1), 39-52.

White, C. N., Skokin, K., Carlos, B., & Weaver, A. (2016). Using decision models to decompose anxiety-related bias in threat classification. *Emotion*, *16*(2), 196.

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-regulation*, *2*, 371-392.

Wichary, S., Mata, R., & Rieskamp, J. (2016). Probabilistic inferences under emotional stress: how arousal affects decision processes. *Journal of Behavioral Decision Making*, *29*(5), 525-538.

Wigboldus, D. H., Sherman, J. W., Franzese, H. L., & Knippenberg, A. V. (2004). Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition*, *22*(3), 292-309.

Wilkerson, W. S. (2001). Simulation, theory, and the frame problem. Philosophical Psychology, 14(2), 141–153.

Woelk, S. P., & Garfinkel, S. N. (2024). Dissociative Symptoms and Interoceptive Integration.

Wood, A., Rychlowska, M., Korb, S., & Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences*, *20*(3), 227-240.

Yap, J. Y., Keatch, C., Lambert, E., Woods, W., Stoddart, P. R., & Kameneva, T. (2020). Critical review of transcutaneous vagus nerve stimulation: challenges for translation to clinical practice. *Frontiers in neuroscience*, *14*, 284.

Zahavi, D. (2008). Simulation, projection and empathy. *Consciousness and cognition*, *17*(2), 514-522.

Zahavi, D. (2011a). Empathy and mirroring: Husserl and Gallese. In *Life, subjectivity & art: Essays in honor of Rudolf Bernet* (pp. 217-254). Dordrecht: Springer Netherlands.

Zahavi, D. (2011b). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology*, *2*(3), 541-558.

Zaki, J. (2020). Integrating empathy and interpersonal emotion regulation. *Annual review of psychology*, *71*(1), 517-540.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nature neuroscience*, *15*(5), 675-680.

Zaman, J., De Peuter, S., Van Diest, I., Van den Bergh, O., & Vlaeyen, J. W. (2016). Interoceptive cues predicting exteroceptive events. *International Journal of Psychophysiology*, *109*, 100-106.

Zamariola, G., Frost, N., Van Oost, A., Corneille, O., & Luminet, O. (2019). Relationship between interoception and emotion regulation: New evidence from mixed methods. *Journal of Affective Disorders*, *246*, 480-485.

Appendix

Appendix A: Chapter 2

From Interoception to Anxiety, Empathic Distress and Emotion Perception: The Role of Intolerance of Uncertainty and Alexithymia

A.1. Anxiety model (study 2)

Model 1b

The model 1b (Study 2) predicting anxiety based on the data of study 2 included the 5 MAIA subscales with significant zero-order correlations with anxiety, alexithymia and IU scores. This model was the same with the previous ones, with the differences in subscales included, namely Self-Regulation and Body Listening instead of Not Distracting and Noticing. The model had good fit to the data (χ^2 (13)=13.823, p=0.386, CFI=0.998, and RMSEA=0.020) explained 59.5% of the variance in anxiety (R^2=0.595), which corresponds to a very large effect size (f^2=1.47).

The direct regressions showed that anxiety was negatively predicted by the MAIA subscales Self-Regulation (β=-3.679, SE=0.825, p<0.001), and Trust (β=-1.765, SE=0.695, p=0.011) and positively predicted by MAIA subscale Emotional Awareness (β=2.727, SE=0.852, p=0.001), alexithymia (β=0.188, SE=0.065, p=0.004) and IU (β=0.208, SE=0.035, p<0.001). IU was negatively predicted by Not Worrying (β=-5.855, SE=1.609, p<0.001), Trust (β=-6.382, SE=1.509, p<0.001), and positively by alexithymia (β=0.776, SE=0.135, p<0.001). Lastly, alexithymia was negatively associated with Not Worrying (β=-1.977, SE=0.948, p=0.037) and Trust (β=-2.691, SE=0.875, p=0.002).

Regarding the indirect paths, alexithymia mediated the relationship between Trust and anxiety (β=-0.505, SE=0.240, p=0.035), and IU mediated the effect of Not Worrying on anxiety (β=-1.217, SE=0.393, p=0.002), as well as the effect of Trust on anxiety (β=-1.327, SE=0.386, p=0.001). Finally, similarly to the results from Study 1, the indirect pathway from Trust to alexithymia through alexithymia and IU was also significant (β=-0.434, SE=0.176, p=0.014).

Model 1b – Anxiety

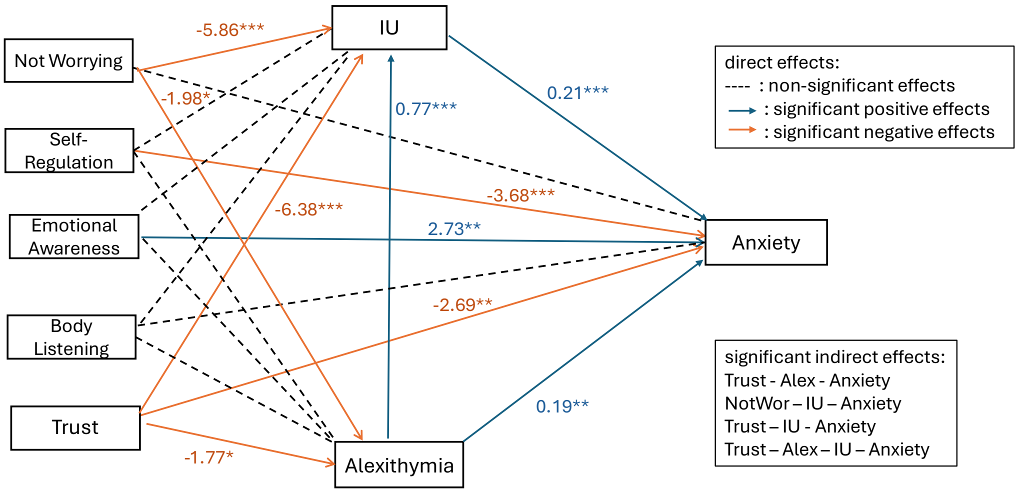


Figure A.1. Path Analysis Results for Anxiety Model (replication). This figure presents the path analysis examining the direct and indirect effects on trait anxiety (STAI). Model predictors are 5 dimensions of MAIA (Not Worrying, Self-Regulation, Emotional Awareness, Body Listening, Trust), alexithymia (TAS-20) and Intolerance of Uncertainty (IU scale). The negative and positive direct and indirect effects are accompanied by regression coefficients (β), with asterisks denoting the significance level (*p<0.05, **p<0.01, *p<0.001).

A.2. Correlation Matrixes

	Notic	NotDis	NotWorry	AttReg	EmAwar	SelfReg	BodyLis	Trust	MAIA	IU	Alex	Mean	SD
MAIA Noticing												3.50	0.66
MAIA Not Distracting	-0.16											1.64	0.76
MAIA Not Worrying	-0.18	0.03										2.13	0.70
MAIA Attention-Regulation	0.27	-0.09	0.16									3.02	0.69
MAIA Emotional Awareness	0.47	-0.19	-0.14	0.45								3.45	0.80
MAIA Self-Regulation	0.21	-0.03	0.10	0.54	0.36							2.89	0.77
MAIA Body Listening	0.27	-0.05	-0.01	0.53	0.48	0.56						2.88	0.89
MAIA Trust	0.18	0.11	0.09	0.47	0.28	0.44	0.38					3.21	0.98
MAIA total	0.42	0.21	0.26	0.79	0.61	0.70	0.69	0.66				2.78	0.41
IU	**0.17**	**-0.24**	**-0.40**	-0.03	**0.19**	-0.03	0.05	**-0.17**	-0.13			78.02	19.30
Alexithymia	-0.07	**-0.23**	**-0.16**	-0.09	0.05	-0.06	-0.04	**-0.25**	**-0.20**	**0.41**		54.34	10.76
STAI	0.10	**-0.20**	**-0.37**	**-0.31**	0.01	**-0.28**	**-0.19**	**-0.56**	**-0.42**	**0.54**	**0.49**	51.64	10.36

Table A.1. Correlation Matrix Study 1. Zero order correlations between all MAIA dimensions, Intolerance of Uncertainty (IU), Alexithymia (TAS-20) and trait anxiety (STAI), with bold values denoting significant correlations (p<0.05) across all participants of Study 1. Mean scores and standard deviations (SD) for each questionnaire also presented.

	Notic	NotDis	NotWorry	AttReg	EmAwar	SelfReg	BodyLis	Trust	MAIA	IU	Alex	STAI	IRI_EC	IRI_PD	Mean
MAIA Noticing															4.01
MAIA Not Distracting	0.05	1.00													3.24
MAIA Not Worrying	0.09	0.06	1.00												3.16
MAIA Attention-Regulation	0.48	-0.15	0.15	1.00											5.30
MAIA Emotional Awareness	0.64	-0.13	0.15	0.45	1.00										4.03
MAIA Self-Regulation	0.53	-0.03	0.23	0.55	0.66	1.00									3.77
MAIA Body Listening	0.52	0.01	0.19	0.34	0.57	0.65	1.00								3.45
MAIA Trust	0.34	0.00	0.25	0.48	0.47	0.59	0.54	1.00							3.71
MAIA total	0.70	0.11	0.39	0.81	0.72	0.81	0.67	0.70	1.00						3.85
IU	-0.06	-0.10	**-0.39**	-0.15	-0.11	**-0.25**	**-0.20**	**-0.46**	**-0.31**	1.00					78.57
Alexithymia	-0.08	-0.07	**-0.24**	-0.10	**-0.16**	**-0.20**	**-0.19**	**-0.33**	**-0.24**	0.55	1.00				61.06
STAI	-0.13	-0.05	**-0.37**	**-0.26**	**-0.17**	**-0.45**	**-0.31**	**-0.54**	**-0.43**	0.67	0.50	1.00			48.69
IRI Empathic Concern	**0.22**	**0.20**	-0.09	-0.08	**0.20**	0.00	0.11	0.01	0.06	-0.02	-0.04	0.04	1.00		19.10
IRI Personal Distress	-0.10	-0.08	**-0.41**	**-0.24**	**-0.21**	**-0.31**	**-0.22**	**-0.44**	**-0.39**	0.60	0.50	0.58	0.25	1.00	14.23
ER accuracy total	-0.03	0.07	0.01	-0.17	0.07	0.00	0.03	-0.01	-0.05	0.11	-0.11	0.17	-0.03	0.07	0.84
ER accuracy negative	0.06	0.06	0.14	0.00	0.18	0.08	0.15	0.09	0.12	-0.02	-0.04	-0.08	0.04	-0.04	0.85
ER accuracy positive	-0.08	0.03	-0.10	**-0.20**	-0.07	-0.07	-0.07	-0.09	-0.16	0.16	-0.08	-0.06	-0.06	0.13	0.83
ER drift rate negative	0.09	0.10	0.06	0.06	0.14	-0.04	0.09	0.00	0.10	0.02	0.01	-0.01	0.13	0.01	1.08
ER drift rate positive	0.10	-0.04	0.08	0.22	0.10	0.07	0.07	0.10	0.18	-0.16	0.06	0.10	0.15	-0.09	-0.77
ER boundary separation	0.12	-0.02	0.18	0.04	0.13	0.14	0.08	0.17	0.15	-0.10	-0.11	0.03	0.10	-0.06	2.05
ER bias	-0.04	-0.06	-0.06	-0.07	0.04	0.01	0.00	0.03	-0.05	0.03	0.04	0.09	-0.24	-0.01	0.44

Table A.2. Correlation Matrix Study 2. Zero order correlations between all MAIA dimensions, Intolerance of Uncertainty (IU), Alexithymia (TAS-20) and trait anxiety (STAI), emotion recognition accuracy for positive and negative stimuli and DDM parameters, with bold values denoting significant correlations (p<0.05) across all participants of Study 2. Mean scores and standard deviations (SD) for each questionnaire also presented.

Appendix B: Chapter 3

Emotion recognition and Interpersonal Emotion Contingencies

B.1. Debriefing questionnaire

Participant reported moderate engagement the roulette game, on average with, when asked to report it in a Likert-scale ranging 1(engaged) to 9 (not engaged). The mean engagement across participants' responses in study 1 was M=5.2 (SD=2.1) and for study 2 M=4.9 (SD=2.3).

In the question: 'How much sad/angry/frustrated did you feel?', ranging from 1(little) to (very), the average intensity of feelings was M=4.9 (SD=1.8) in study 1 and M=4.7 (SD=1.8) in study 2.

The assessment of the intensity of positive feelings after a winning outcome in the game showed that there was moderate to high satisfaction with mean intensity of M=6.2 (SD=1.9) in study 1 and M=5.9 in study 2.

From the open question regarding the structure and underlying rationale of the task, none reported, across the 2 studies, awareness of the manipulations of the probability of interpersonal emotion congruency.

B.2. Emotion recognition task - Reaction times

Study 1

The 3x2x2 (Block x Trial x Noise) ANOVA on RT revealed a significant main effect of Block (F(2,84)=3.351, p=0.040, η^2_p=0.074) with higher RT in the neutral block, and lower in the expected incongruency block. Bonferroni-corrected pairwise comparison showed that only the difference between the neutral and incongruent blocks reached significance (t(1,42)=5.10, p<0.001, Cohen's d=0.857). The main effects of Trial (F(1,42)=0.029, p=0.864, η^2_p=0.001) and Noise (F(1,42)=2.043, p=0.160, η^2_p=0.046) were not significant. Of all interactions, only the one between Trial x Noise was close to significance (F(1,42)=3.814, p=0.054, η^2_p=0.083) as in the congruent trials, RTs were lower in conditions of high noise in comparison to low noise (t(131)=2.049, p=0.042, Cohen's d=0.121), whereas in the incongruent trials there was no difference (t(131)=0.142, p=0.853, Cohen's d=0.009). All other interactions were not significant (p>0.139).

The 2x2x2 (Block (Congruent block, Incongruent block) x Trial Expectancy (Expected, Unexpected) x Noise (High, Low)) ANOVA didn't show a significant main effect of Expectancy (F(1,43)=0.236, p=0.630) or any significant interaction.

Study 2

The 3x2x2 (Block x Trial x Noise) ANOVA on RT revealed a significant main effect of Block (F(2,88)=13.002, p<0.001, η^2_p=0.228) with lower RT in the incongruent block (M=0.47s, SD=0.15) compared to the congruent (M=0.51s, SD=0.18) and neutral block (M=0.54s, SD=0.15). The main effect of Trial was also significant (F(1,44)=6.408, p=0.015, η^2_p=0.127) with congruent trials having lower RTs (M=0.49s, SD=0.15) than incongruent ones (M=0.52s, SD=0.17). The main effect of Noise and its interactions did not reach significance (ps>0.095).

The 2x2x2 (Exp_Cong x Expectancy x Noise) ANOVA revealed a significant interaction between Block and Expectancy (F(1,44)=7.825, p=0.008, η^2_p=0.012) where higher RT in congruent trials was higher than incongruent ones in the expected congruency block, with the reverse pattern observed in the expected incongruency block (Fig.3). All other main effects and interactions did not reach significance (p>0.150).

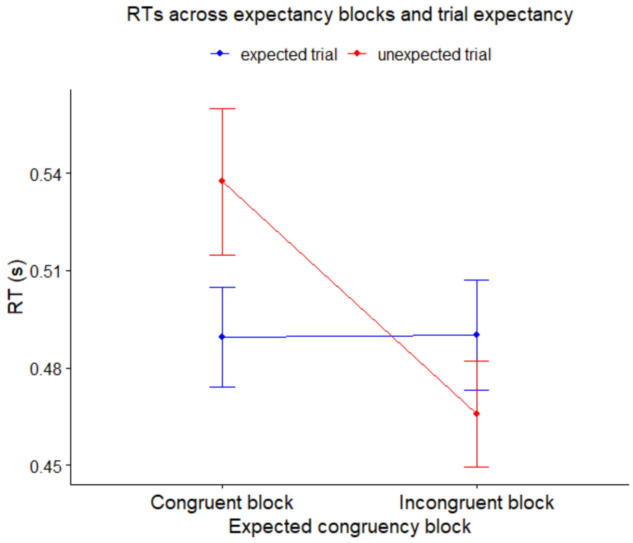


Figure B.2. Study 2 reaction time results. Mean RT across expected congruency blocks for trials with expected and unexpected emotional expressions in the context each block.

B.3.

Model	State/parameter	Mean	Variance
HGF3	κ	1	0
	ω2	-1	16
	ω3	-2	16
	μ2	0	0
	σ2	0.1	0
	μ3	1	0
	σ3	1	0
	ζ	48	10
	w	0.5	10
HGF2			
	κ	1	0
	ω2	-1	16
	ω3	-2	0
	μ2	0	0
	σ2	0.1	0
	μ3	1	0
	σ3	1	0
	ζ	48	10

Table B.3. Perceptual and response models configurations. The table presents the initial mean and variance configurations for all parameters and states for the HGF with two and three level. All parameters/states are part of the response model except ζ and w, which are part of the response models. The parameter w is the weighting factor included in the response model where prior and posterior beliefs are combined.

RW0	RW	HGF2	HG3	Prior	Posterior	Combined
-132.184	-131.493	-132.369	-132.403	-132.403	-131.777	-132.751
-168.432	-9011.31	-166.196	-41.803	-41.803	-161.726	-161.861
-155.924	-160.718	-190.231	-155.627	-155.627	-146.859	-147.006
-160.107	-8357.05	-157.064	-159.335	-159.335	-153.821	-152.324
-121.428	-122.526	-121.649	-121.372	-121.372	-116.536	-117.095
-153.82	-152.73	-153.189	-153.255	-153.255	-128.979	-131.678
-167.085	-8835.43	-164.486	-164.576	-164.576	-150.975	-151.472
-167.412	-166.727	-167.193	-167.271	-167.271	-155.968	-156.428
-146.875	-165.494	-151.612	-149.287	-149.287	-146.389	-146.121
-150.756	-151.24	-150.251	-150.329	-150.329	-148.962	-149.701
-168.982	-167.527	-168.164	-168.312	-168.312	-158.13	-160.321
-165.151	-163.663	-167.116	-164.28	-164.28	-157.869	-161.292
-162.685	-163.571	-162.044	-162.093	-162.093	-89.1499	-160.638
-167.869	-168.184	-166.923	-166.807	-166.807	-162.461	-163.006
-164.543	-164.09	-163.916	-163.847	-163.847	-150.839	-151.707
-165.914	-163.979	-165.117	-165.346	-165.346	-179.105	-144.819
-157.598	-158.751	-160.704	-157.16	-157.16	-139.707	-142.098
-164.823	-178.771	-164.985	-164.779	-164.779	-629.611	-153.638
-165.104	-172.13	-164.052	-163.948	-163.948	-162.517	-163.035
-151.567	-178.522	-150.713	-150.493	-150.493	-144.202	-147.647
-156.282	-11875	-154.844	-154.629	-154.629	-253.423	-154.826
-166.501	-209.51	-165.942	-165.991	-165.991	-139.791	-140.175
-164.958	-164.35	-164.208	-164.046	-164.046	-162.129	-161.971
-168.451	-165.569	-168.162	-99.4571	-99.4571	-153.269	-154.103
-161.42	-161.469	-161.357	-161.068	-161.068	-151.198	-152.157
-153.936	-153.889	-153.325	-153.256	-153.256	-146.12	-147.105
-166.121	-198.965	-165.217	-165.267	-165.267	-150.633	-151.185
-168.087	-166.946	-28.5926	-167.381	-167.381	-141.792	-149.685
-169.715	-8406.27	-169.613	-170.49	-170.49	-140.757	-142.021
-102.289	-103.356	-101.501	-101.618	-101.618	-122.254	-102.627
-162.321	-182.281	-162.353	-161.564	-161.564	-144.449	-143.945
-156.946	-156.495	-156.328	-156.419	-156.419	-136.143	-136.887
-162.604	-162.888	-162.011	-161.516	-161.516	-147.923	-148.782
-164.127	-180.244	-163.427	-163.404	-163.404	-147.739	-148.08
-159.299	-157.772	-181.73	-158.588	-158.588	-134.098	-135.887
-157.822	-157.456	-157.52	-157.58	-157.58	-145.636	-145.861
-164.227	-162.288	-163.737	-163.761	-163.761	-145.295	-148.672
-147.482	-146.139	-146.843	-21.0553	-21.0553	-139.71	-139.48
-163.734	-8835.95	-160.888	-160.584	-160.584	-153.803	-154.394
-158.807	-183.822	-158.811	-158.025	-158.025	-150.259	-151.138
-162.649	-163.25	-161.199	-160.899	-160.899	-141.099	-158.887
-164.429	-163.168	-163.385	-163.402	-163.402	-155.078	-157.605
-164.975	-178.727	-164.408	-164.407	-164.407	-146.675	-147.486

Table B.4. LME values for the perceptual and response models of study 1. The table presents the Log Model Evidence for each participant and model compared in study 1. The RW0, RW, HGF2, HG3 are the perceptual models compared and the Prior, Posterior, Combined are the response models compared.

RW0	RW	HGF2	HG3	Prior	Posterior	Combined
-168.074	-147.994	-159.509	-136.228	-154.633	-173.094	-154.272
-146.992	-142.056	-149.438	-173.176	-179.445	-157.681	-116.776
-145.36	-128.537	-172.97	-156.382	-192.672	-129.935	-159.049
-154.939	-154.184	-164.33	-149.688	-144.875	-144.236	-162.917
-174.941	-144.32	-152.308	-168.509	-177.25	-158.582	-147.889
-167.848	-180.577	-182.463	-145.166	-172.253	-172.318	-169.77
-179.462	-174.376	-181.792	-165.903	-157.785	-153.603	-163.692
-146.239	-200.034	-172.002	-147.201	-151.665	-140.982	-173.956
-157.351	-172.04	-184.663	-154.521	-143.341	-171.995	-173.819
-148.672	-166.765	-137.293	-157.626	-167.888	-159.349	-150.275
-168.872	-171.049	-149.064	-186.498	-157.4	-146.052	-145.743
-132.334	-143.279	-156.868	-151.029	-172.049	-143.308	-148.861
-132.746	-172.8	-143.716	-168.365	-164.995	-154.951	-161.685
-161.858	-171.381	-150.571	-177.973	-150.334	-148.236	-185.381
-176.659	-158.53	-173.111	-192.21	-150.948	-166.878	-174.854
-170.214	-172.744	-131.233	-156.622	-139.44	-167.326	-168.1
-159.787	-151.589	-158.538	-154.445	-154.306	-146.997	-158.894
-160.893	-165.212	-152.395	-188.209	-156.699	-149.664	-153.806
-169.915	-139.437	-144.629	-159.206	-178.964	-140.848	-146.995
-155.411	-148.885	-173.918	-170.503	-167.08	-151.314	-171.828
-166.136	-170.693	-173.338	-156.738	-167.773	-159.903	-138.584
-179.228	-147.929	-174.748	-186.696	-173.077	-159.92	-146.937
-164.274	-145.336	-160.471	-152.204	-115.33	-151.168	-157.672
-160.972	-145.906	-146.969	-165.333	-164.859	-156.965	-153.363
-144.994	-193.944	-173.722	-130.231	-130.243	-151.754	-161.41
-172.02	-167.705	-159.485	-180.248	-154.383	-154.812	-172.508
-158.623	-189.842	-163.754	-169.527	-148.679	-131.42	-163.932
-155.893	-134.827	-175.863	-171.819	-159.857	-162.369	-153.701
-192.002	-154.62	-172.92	-139.478	-155.723	-166.275	-178.449
-153.94	-163.8	-165.303	-138.365	-160.531	-177.473	-168.493
-174.377	-177.203	-147.557	-167.685	-169.676	-155.627	-170.202
-164.895	-170.513	-165.531	-155.464	-149.754	-177.344	-176.687
-122.82	-169.682	-159.076	-129.313	-170.962	-160.93	-126.431
-136.914	-148.272	-151.629	-157.401	-189.59	-155.875	-154.226
-147.595	-155.265	-162.646	-159.005	-157.524	-174.55	-160.316
-162.394	-140.884	-144.276	-190.287	-195.397	-153.591	-184.899
-172.866	-161.118	-144.793	-196.36	-140.336	-175.543	-160.548
-154.025	-194.554	-174.232	-140.474	-144.277	-167.355	-161.013
-163.372	-145.216	-160.161	-130.85	-129.598	-148.414	-174.458
-179.72	-144.646	-165.272	-156.092	-167.891	-166.727	-163.601
-160.327	-158.026	-185.243	-165.576	-201.034	-173.861	-161.78
-173.201	-150.969	-175.74	-168.684	-143.992	-158.508	-150.3
-143.37	-160.102	-149.653	-144.424	-163.515	-150.765	-150.1
-166.962	-158.679	-159.888	-132.147	-136.665	-185.394	-175.255

Table B.5. LME values for the perceptual and response models of study 2. The table presents the Log Model Evidence for each participant and model compared in study 2. The RW0, RW, HGF2, HG3 are the perceptual models compared and the Prior, Posterior, Combined are the response models compared.

Appendix C: Chapter 4

HEPs and adaptive empathy

Cluster	Electrodes
fronto_central	FC1, FC2, Fz,Cz
right_frontal	FP2,F4,F8,FC6
left_frontal	FP1,F3,F7,FC5
right_centroparietal	CP2,CP6,C4,P4
left_centroparietal	CP1,CP5,C3,P3
right_tempoparietal	TP10,T8,P8
left_tempoparietal	TP9,T7,P7
occipital	O1,O2,Pz

Table C.6. Electrode clusters for HEP analysis. The table presents the 7 clusters of electrodes created for all HEP cluster-based permutation analyses.

Appendix D: Abbreviations

Abbreviation	Meaning
AET	Adaptive Empathy Task
AQ	Autism Quotient
DDM	Drift Diffusion Model
EC	Empathic Concern
ECG	Electrocardiogram
EEB	Emotion Egocentricity Bias
EEG	Electroencephalogram
EOG	Electrooculogram
EQ	Empathy Quotient
ERS	Emotion Regulation Strategy
HEP	Heart-Evoked Potential
HGF	Hierarchical Gaussian Filter
HRA	Heart Rate Acceleration
IAcc	Interoceptive Accuracy
IS	Interoceptive Sensibility
ICA	Independent Component Analysis
IEC	Interpersonal Emotion Contingencies
IRI	Interpersonal Reactivity Index
IT	Interaction Theory
IU	Intolerance of Uncertainty
LMH	Learned Matching Hypothesis
MAIA	Multidimensional Assessment of Interoceptive Awareness
PD	Personal Distress
PE	Prediction Error
POMDP	Partially Observable Markov Decision Process
PP	Predictive Processing
pwPE	Precision-weighted Prediction Error
RL	Reinforcement Learning
RW	Rescorla-Wagner Model
SCR	Skin Conductance Response
ST	Simulation Theory
STAI-T	State-Trait Anxiety Inventory - Trait
TAS-20	Toronto Alexithymia Scale

Table D.7. Abbreviations list