Emotional Egocentricity Bias is modulated by implicit expectations of interpersonal emotional contingencies and perceptual noise

Vassilis Kotsaris[1], Dimitris Bolis[2], Ruben Azevedo[1]

[1]School of Psychology, University of Kent, UK

[2] Laboratory for Autism and Neurodevelopmental Disorders, Center for Neuroscience and Cognitive Systems UniTn, Instituto Italiano di Tecnologia, Rovereto, Italy

corresponding author:

Ruben Azevedo

Keynes College, School of Psychology, University of Kent, Canterbury, Kent CT2 7NP

Email: r.a.teixeira-azevedo@kent.ac.uk

Abstract

Emotional Egocentricity Bias (EEB) refers to the tendency to project one's own emotional state onto others. While previous research has demonstrated EEB in multiple paradigms, its underlying mechanisms remain unclear. Across two studies, we used a novel dual-task paradigm to examine how fluctuations in expected interpersonal emotional contingencies (IEC) and perceptual ambiguity shape EEB. In each trial, participants underwent an implicit emotion induction through a roulette game and subsequently categorized an ambiguous facial expression. Experimental blocks varied in the probability of emotion congruency between self and other. Behavioural results showed that implicit congruency expectations modulated EEB as accuracy was highest for congruent trials in neutral and congruent blocks but reversed in incongruent blocks, indicating implicit adaptation to IEC. Interestingly, higher perceptual noise improved performance and amplified contextual effects, suggesting that EEB is jointly shaped by interpersonal predictive processing and sensory noise. To model individual learning of IEC, we employed a Hierarchical Gaussian Filters (HGF) computational model, revealing that participants updated their beliefs about IEC in a volatility-sensitive manner and that decisions were primarily based on posterior beliefs. Heart rate acceleration following outcomes was linked to belief updates, suggesting that arousal influences socio-emotional learning. These findings show that EEB reflects context-sensitive inferences shaped by internal states and perceived uncertainty and highlight the role of interoception in adaptive emotion recognition. This work adds to our understanding of EEB offering insights for future studies on embodied emotion perception in dynamic social contexts.

Keywords:

emotional egocentricity bias, interoception, Hierarchical Gaussian Filter, Bayesian learning, predictive processing, emotion recognition, interpersonal contingencies

Introduction

Understanding others' emotions is fundamental for effective social interactions (Morrison & Bellack, 1981). Yet, we do not have direct access to others' affective states and need to infer them based on the available information. To achieve this, the brain may integrate sensory cues (e.g. facial expressions, prosody) with contextual information (e.g. prior social knowledge) and internal states (e.g. our own momentary feelings and bodily signals) to arrive at a "best guess" of others' affective state; a process and outcome often characterized by uncertainty (FeldmanHall & Shenhav, 2019, Ondobaka et al, 2017). Interestingly, when the observer's affective state differs from that of the other, these affective inferences can be biased to align with the observers' own emotions. This phenomenon is known as Emotion Egocentricity Bias (EEB; Steinbeis & Singer, 2014; Von Mohr et al., 2021) and has been linked to mood-congruency effects (Forgas, 2017; Trilla et al., 2021) and affective realism (Anderson et al., 2012). While EEB has been demonstrated across numerous tasks (Riva et al., 2016; Sevi et al, 2022; Steinbeis & Singer, 2014; Trilla et al., 2021; Weigand et al, 2021), its computational and psychophysiological mechanisms have not yet been explored.

One prominent mechanism proposed to explain how we understand others' emotional states is embodied simulation, wherein observing another's emotions automatically activates matching neural, emotional, and somatic states in the observer (Gallese, 2014; Singer & Lamm, 2009). This theory suggests that we innately use our own emotional experiences as a reference to interpret others' emotions. An alternative perspective argues that these mirroring effects are not innate but arise through associative learning (Heyes, 2018). According to the Learned Matching Hypothesis, repeated exposure to similar social situations results in the formation of associations between observed exteroceptive cues (e.g., facial expressions) and bodily and emotional responses on the observer. These learned

associations enable observers to activate congruent affective states when perceiving others' emotional cues. Crucially, the reverse effect can also occur: an experienced affective state could be projected onto the other as the most probable hypothesis about the observed affective state. As these bidirectional links between observer and observed emotions are shaped through learning, they remain flexible and sensitive to contextual variability and dynamic interpersonal emotion contingencies (IEC; Barsalou, 2013; Heyes, 2018; Kilner et al. 2007). Consequently, EEB could be influenced by probabilistically shaped expectations regarding emotional congruency between oneself and others. For example, feeling happy might create an expectation that others in a similar context will display happiness as well, thus promoting EEB by, for example, interpreting ambiguous or neutral expressions as happy. However, what happens when such interpersonal contingencies are not fixed or aligned with our past experience? To the best of our knowledge, no study so far has investigated contextual modulations of EEB. Can (implicit) learning of IEC influence EEB?

The predictive processing (PP) framework provides a complementary perspective on how we interpret social and emotional information (Ondobaka et al., 2017; Seth & Friston, 2016), casting perception as an active, probabilistic inference process driven by prior expectations (Friston & Kiebel, 2009). Within this framework, higher-level beliefs generate predictions about incoming input, which are updated based on prediction errors - discrepancies between expected and observed outcomes. These updates are modulated by the estimated precision/reliability (inverse uncertainty) of both prior beliefs and sensory evidence. Applications of PP principles to social learning have emphasized the importance of quantifying different sources of uncertainty in dynamic environments (Behrens et al., 2008; Diaconescu et al., 2014; Sevgi et al., 2020). A prominent computational tool in this domain is the Hierarchical Gaussian Filter (HGF), a Bayesian learning model that accommodates belief

updating across three distinct, yet coupled, nested levels (Mathy et al., 2011; 2014). The first level quantifies irreducible uncertainty, corresponding to the inherent ambiguity of sensory input. In the context of EEB, it reflects uncertainty about the emotional state of others relative to one's own affective experience. The second level models estimation uncertainty, that is, the probability of a congruent emotional outcome given the observer's internal state and contextual cues. At the third level, volatility estimates quantify the beliefs about how rapidly cue-outcome contingencies, such as IEC, change over time. By applying the HGF in the context of EEB, we can formally examine how individuals learn about dynamic changes in IEC and how this learning process is modulated by individual 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, especially under uncertain conditions (Damasio, 1996; Dunn et al., 2010; Pfeifer et al., 2017). Importantly, emotion inferences also build on interoception, shaping both affective experience and social cognition (Grynberg & Pollatos, 2015; Seth & Friston, 2016; Shah et al., 2017). In the context of EEB, a recent study by Von Mohr and colleagues (2021) investigated how interoception may affect EEB by synchronizing stimulus presentation to different phases of the cardiac cycle (systole/diastole). They showed 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 suggest a significant role for physiological activity and interoceptive processing in EEB, which remains largely unexplored.

A variety of different paradigms have been utilized to study EEB. Typically, these involve the induction of emotionally congruent or incongruent states between participants and

observed targets, followed by judgements of both self and other's affective states (Trilla et al., 2021; Steinbeis & Singer, 2014; von Mohr et al., 2019). Common emotion manipulation methods include monetary games or synchronous multisensory stimulation. However, these approaches often explicitly prompt participants to differentiate between their own and others' emotional states, which is likely to require explicit active disengagement from self-related perspectives, potentially confounding the observed EEB effects. In other words, it is challenging to discern whether EEB primarily reflects self-other distinction difficulties or affective projection processes. 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, induced using autobiographical recall and audio-visual clips. The emotion recognition task involved the categorization of ambiguous facial expressions as happy or sad. The results showed that participants were more likely to judge faces as happy after the happiness induction. While this approach allows investigating EEB in a more implicit way, by avoiding explicit cues of self-other comparisons, it still emphasizes participants' own emotions during induction.

Here, we developed a novel dual-task paradigm to assess implicit EEB by minimizing explicit self-other distinction cues. In each trial, participants first engaged in an emotion induction task involving a simplified roulette game, where successful bets are rewarded and unsuccessful ones penalized, to evoke positive and negative emotions, respectively. Immediately after the game's outcome, participants were presented with ambiguous facial expressions, i.e. morphings of sad and happy expressions, and were asked to perform binary emotional categorizations. Crucially, these two tasks were presented as unrelated, reducing the likelihood of explicit associations between participants' own emotional states and their judgments about others' emotions, thereby enabling a more naturalistic assessment of EEB.

Across two studies, we examined the role of implicit expectations of IEC in the modulation of EEB. In Study 1, we implemented the dual-task paradigm online to test whether contextual IEC and perceptual noise flexibly shape EEB, and to characterise trial-wise learning with Hierarchical Gaussian Filters (HGF). In Study 2, we ran the same task in the laboratory to test whether the behavioural and computational effects generalise across settings and to examine how autonomic arousal (i.e. cardiac activity) covaries with IEC learning. In both studies, the task was performed under three different block conditions with varying levels of expected IEC (neutral, expected congruency and expected incongruency) by manipulating the probability that the emotional states induced by the game matched the subsequently presented ambiguous facial expressions. We also manipulated perceptual uncertainty by adding pixelated noise to the facial expressions.

Firstly, we expected, for both studies, accuracy in the emotion recognition (ER) task to be higher for congruent (vs. incongruent) trials in the neutral block, indexing a baseline tendency towards EEB. We further predicted that accuracy would be shaped by IEC and visual noise: accuracy should be higher on congruent trials in blocks with expected congruency and on incongruent trials in blocks with expected incongruency, reflecting implicit learning of IEC; and overall performance should decline under high visual noise. In Study 2, given existing evidence on interoception in emotion and decision-making, we additionally hypothesised that trial-wise physiological arousal would covary with participants' adaptability to dynamically changing contingencies, such that lower arousal levels would be associated with higher learning rates and reducing EEB. Finally, across both studies we explored whether individual differences in self-reported interoceptive sensibility, alexithymia, and empathy modulated HGF parameters and their coupling to autonomic responses.

Methods

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.

Stimuli

Facial stimuli were based on colour photographs from the FACES database (Ebner, Riediger, & Lindenberger, 2010), depicting 10 actors (5 female, 5 male) displaying happy and sad expressions. Images were converted to greyscale and morphed along a happy-sad continuum using the Morpheus Photo Morpher software (https://www.morpheussoftware.net/) to create morphs at 20/80, 30/70 and 40/60% proportions of each emotion. In a separate pilot study, 15 participants (age range 18 - 44 years, median age 20, 8 females) performed a 2-alternative forced-choice task, categorising each morph as "happy" or "sad". For each model, responses were fitted with psychometric functions to identify morph levels at which images were judged as happy (or sad) with 60% probability. This procedure yielded subjectively calibrated morph levels that were ambiguous yet slightly biased towards one emotion. For the main task, we selected for each

model two ambiguous expressions (one biased towards happiness and one towards sadness) at these 60/40 levels. Each image was then pixelated to create a corresponding high-noise version. In total, 40 stimuli (10 models × 2 morph levels × 2 noise levels) were used in the emotion perception task.

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 (Fig. 1). The roulette game was used to induce positive and negative affect: participants made binary decisions (in some trials this was black/red in other trials even/odd number) to bet on where the ball would stop and were then shown whether they had won or lost. This feedback served as the emotion induction manipulation (Steinbeis & Singer, 2014). After an interval of 3s, participants were shown an 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 explicit associations between the roulette's outcome and the emotion categorization task, participants were told that the study was examining emotion perception under uncertainty using an irrelevant, yet demanding, secondary task to increase their cognitive load. This allowed us to study the preconscious influence of one's own emotional state on judgments of others' emotions.

The probability of winning or losing was fixed at 50%, controlled on a trial-by-trial basis by the experiment algorithm, but this was not disclosed to participants. To enhance engagement, participants were told that each win gave them 3 points, and each loss took 2

points from their score, with accumulated points increasing their chances of winning a £20 voucher via raffle tickets. To further motivate participants, a “special roulette” trial was included three times per experiment, offering five times the usual reward (15 points). Access to these trials was framed as contingent on performance relative to previous participants.

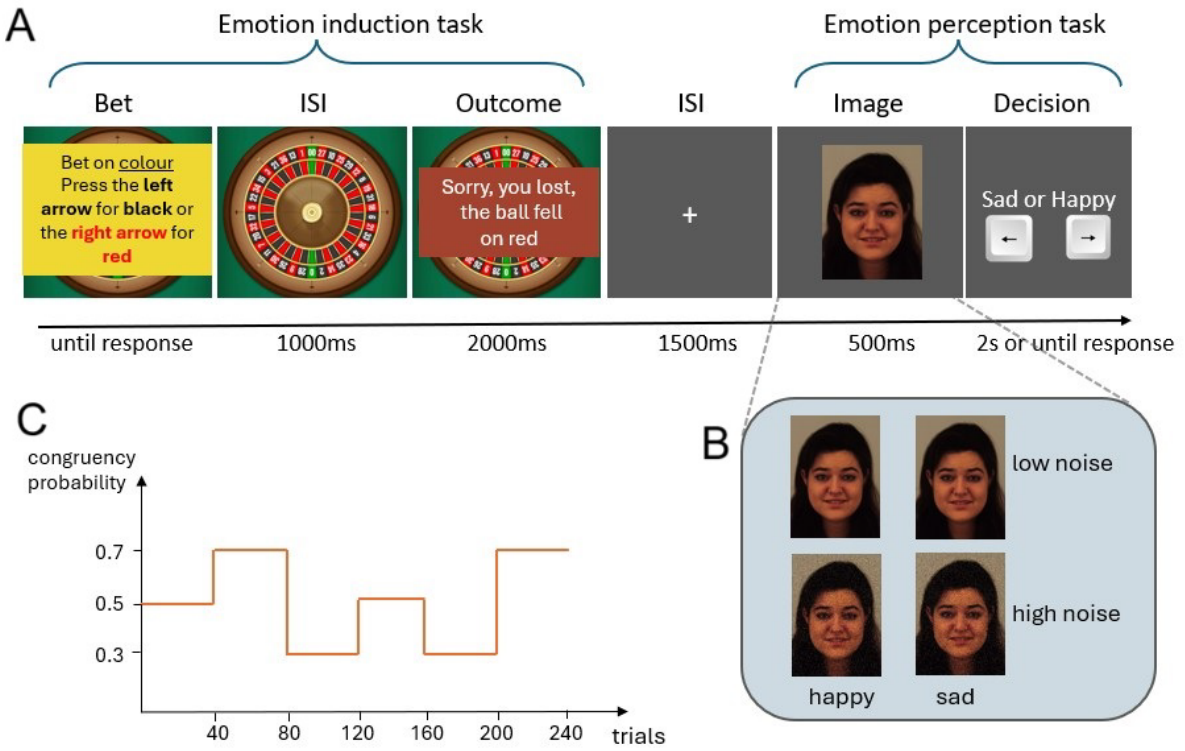


Figure 1. Experimental design. A) Example trial of the dual-task paradigm. In the emotion induction phase participants bet either on colour or on number and then see the outcome of the roulette game. Afterwards, they are briefly presented an ambiguous image and decide whether it was happy or sad. B) The 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 blocks varying on probability of expected congruency.

To manipulate the expectancy of IEC, we created three block types: expected congruency (CB), expected incongruency (IB), and neutral (NB). In the CB, there was a 70% probability of seeing a happy face after a win and a sad face after a loss (Fig. 1C). In the IB, these

probabilities were reversed (i.e., 30% for congruent pairings), while in the NB, there was a 50% probability for either expression after each outcome. The experiment comprised six blocks in the order: NB, CB, IB, NB, IB, CB. Each block contained 40 trials (total: 240 trials). In half the trials per block, faces were presented with high visual noise (pixelation), randomly intermixed, yielding two noise levels (Fig. 1B). To avoid stimulus-driven expectancy effects, the marginal probabilities of sad and happy faces were identical and constant across all blocks.

The task started with 4 practice trials. Participants were given two self-paced breaks. Each trial lasted approximately 7 s, and the total task duration was about 35 min. At the end, participants completed seven debriefing questions to assess their experiences during the task and to probe for explicit awareness of any associations between the two tasks (i.e., roulette outcome and subsequent emotion categorization; see Supplementary Material S5).

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/).

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) which includes 37 questions assessing the different ways people process and pay attention to their bodily

sensations. Specifically, it consists of 8 subscales: noticing, not-distracting, not-worrying, attention regulation, emotional awareness, self-regulation, body listening, and trust. 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 scale measures Alexithymia, i.e. the difficulty in identifying and expressing own emotions, 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.

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.

Physiological measurements

We recorded skin conductance responses (SCR) and cardiac activity using the BIOPAC MP36 (https://www.biopac.com/). For the SCR, two electrodes were attached to the middle and index finger of the participants' left hand (this data was not analysed due to poor data quality for several participants). For the ECG recording, we used a lead II chest configuration, where electrodes were placed below the left and right collarbones and on the left lower back. ECG signals were sampled at 2000Hz.

To assess cardiac responses to game outcomes, we calculated the interbeat interval (IBI) within a window beginning 0.5 seconds after outcome presentation and ending 4 seconds later. IBIs were interpolated between R-peaks to provide a continuous estimate of cardiac period. As an index of event-related autonomic arousal, we quantified peak cardiac acceleration as the lowest IBI within this window, normalised by dividing this value by the baseline IBI (average IBI during the 1 second preceding outcome presentation).

Statistical analysis

Model-free analysis

The main DV was accuracy in the emotion categorization task. To examine how the experimental conditions influenced responses, we conducted 3 × 2 × 2 within-subjects ANOVAs on both RTs (Supplementary Material S1) and accuracy, with Block Type (Neutral Block [NB], Congruent Block [CB], Incongruent Block [IB]), Trial Congruency (congruent, incongruent), and Noise (high, low) as factors. Post hoc pairwise comparisons were Bonferroni-corrected.

HGF – model-based analysis

We used the HGF to model the learning processes of IEC (Mathys et al., 2011; 2014). The HGF is a generative model that approximates a Bayesian observer that dynamically tracks the hidden states of its environment, an approach described as 'observing the observer' (Daunizeau et al., 2010). The HFG models trial-by-trial belief updating across different representational hierarchies, with each state evolving as a Gaussian random walk, and the step size (variance) at each level governed by the state above (Fig. 2). 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: $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}$ (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) encodes 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 reflecting irreducible, perceptual uncertainty of interpersonal emotional states. In the absence of perceptual noise, the mapping of input to first level beliefs can be deterministic, yet here we applied a stochastic mapping with perceptual noise captured by a fixed parameter alpha (α) representing the variance of the noise on the input.

The second level (x_2) represents the inferred probability of seeing a facial expression congruent with the outcome of the game, and is linked to the first via a Bernoulli distribution: $p(x_1|x_2) = \text{Bernoulli}(x_1; s(x_2))$, where s(x) is a sigmoid function $s(x) = 1/(1 + \exp(-$

x)). The evolution of x_2 is governed 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 resulting update equations have the following form describing how beliefs (posterior means) at each level of the hierarchy (i) are proportional to the prediction error (δ_{i-1}) and to a precision ratio:

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

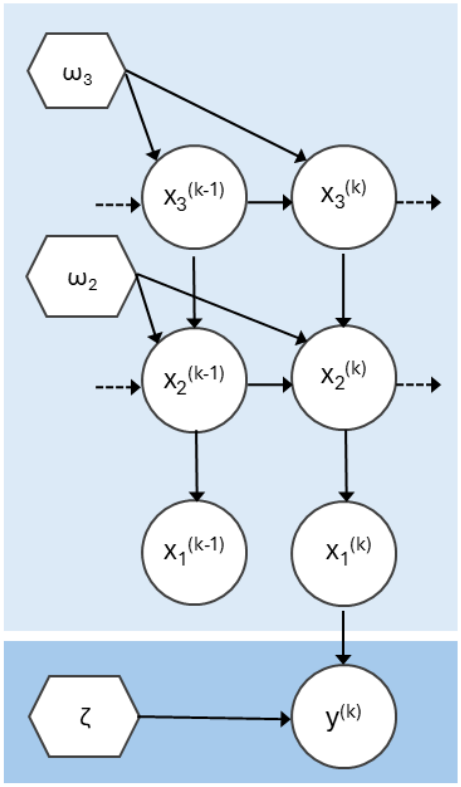
The prediction error describes the difference between the lower-level prediction $\hat{\mu}_{i-1}^{(k)}$ before seeing the input and the posterior expectation $\mu_{i-1}^{(k)}$ afterwards. This term is weighted by a ratio of the precision of the prediction of the lower level $\hat{\pi}_i^{(k)}$ before seeing

the input divided by the precision of the current belief $\pi_i^{(k)}$, 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, but beliefs about how IEC change over time could be volatile nevertheless. To test model complexity, we compared a 3-level HGF (with volatility estimation) to a 2-level HGF (HGF2) without a volatility level (by fixing prior and variance $\kappa=0$). We also tested the widely used reinforcement learning model of Rescorla Wagner (RW) where, similarly to HGF, prediction errors drive belief updating, but with a fixed learning rate (Rescorla and Wagner, 1972). We can thereby examine whether a simpler model with a constant learning rate or a more complex and adaptive model provide a better account of participants' learning behaviour.

Model	Hidden states	Learning mechanism	Volatility inference	Free perceptual parameters
RW0	V only	No learning	no	–
RW	V only	Fixed α	no	α
HGF2	x_1, x_2	Adaptive	no	ω_2
HGF3	x_1, x_2, x_3	Adaptive	yes	ω_2, ω_3

Table 1. Summary of perceptual/learning models. Comparison of the four candidate learning models tested: RW0 (no-learning), RW (fixed-rate Rescorla–Wagner), HGF2 (two-level Hierarchical Gaussian Filter with adaptive learning), and HGF3 (three-level HGF including volatility inference). Models differ in representational depth, flexibility of learning rate, and whether volatility is estimated. RW0 ⊂ RW ⊂ HGF2 ⊂ HGF3 illustrates increasing hierarchical complexity. Free perceptual parameters capture distinct components of the learning process: α determines the fixed learning rate in RW, ω_2 reflects tonic log-volatility of belief updates in HGF2/3, and ω_3 encodes meta-volatility driving changes in volatility estimates in HGF3.



x_1	Outcome uncertainty: binary belief about the emotional congruency of the facial expression – irreducible uncertainty about the current trial
x_2	Estimation/information uncertainty: estimated probability of seeing an emotion congruent with one's own induced state – how expected congruency is learned
x_3	Volatility uncertainty: estimated rate of change of x_2 over time - belief about how stable (or volatile) the context is
y	Observed responses
ω_2	Tonic log-volatility: how quickly one thinks congruency contingencies can change – governs variance of belief updates at level 2
ω_3	Meta-volatility: how quickly one thinks the volatility itself can change – governs variance of belief updates at level 3
ζ	Response model decision parameter: How deterministically one uses their beliefs to decide

Figure 2. Hierarchical Gaussian Filter (HGF) architecture for modelling emotion congruency beliefs under uncertainty. The three-level perceptual model (light blue) and the response model (dark blue) together simulate belief updates and response generation. In our task, the lowest representation level x_1 encodes irreducible uncertainty about interpersonal emotion congruency, with binary outcomes coded as 1 (congruent) and 0 (incongruent); x_2 represents the estimated probability of congruent expressions (learned expected congruency), and x_3 tracks the volatility of changes in this expectation. Circles denote model states that evolve over time, while hexagons refer to the fixed parameters (ω_2, ω_3). $x_1^{(k)}$, $x_2^{(k)}$, $x_3^{(k)}$, are the hidden states of the social environment at each time point (k). $x_2^{(k)}$ and $x_3^{(k)}$ values are generated at each time point based on the values of the previous time point (k-1) of the HGF levels, and the learning parameters ω_2 and ω_3, where ω_2 captures the tonic log-volatility of x_2 and ω_3 the meta-volatility of x_3. Responses ($y^{(k)}$) are generated from $x_1^{(k)}$ through the response model, with decision parameter ζ determining the stochasticity of choice.
HGF3 differs from HGF2 in that it includes the volatility level x_3, enabling learning-rate adaptation to environmental change, whereas HGF2 contains only x_1–x_2 and therefore learns without volatility inference.

Response models

We modelled participants' binary responses using the unit-square sigmoid function (Iglesias et al., 2013), which maps model-derived beliefs onto response probabilities. Specifically, the probability of choosing a congruent response was given by:

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

where $\hat{\mu}_1$ is the expected probability of observing a congruent facial expression. The free parameter ζ captures decision noise or exploration tendency, with higher values indicating more deterministic, less exploratory choices. In the formula above, $\mu_i(k)$ denotes the posterior mean at level *i* after observing trial *k*, whereas $\hat{\mu}_i(k)$ is the prior (prediction) for trial *k* before the outcome. The posterior on trial *k* becomes the prior for trial *k+1*.

To examine how participants combined perceptual and expectation-based information, we specified two variants of the response model. In this task, participants could base their judgments on ambiguous visual cues, prior expectations shaped by the IEC manipulation, or both. Accordingly, one model variant used the predicted probability at the first perceptual level ($\hat{\mu}_1^{(k)}$) as input, while a second variant used the posterior belief at the first level ($\mu_1^{(k)}$). Comparing these models allowed us to assess whether participants' responses were more strongly influenced by immediate sensory evidence or by their updated expectations.

Model inversion

We performed approximate Bayesian inversion for each model by estimating the maximum-a-posteriori (MAP) estimates of model parameters for each participant. This was implemented using the HGF toolbox (version 2.1), specifying parameter priors and providing trial-wise input sequences. For the main perceptual model, we used the tapas_hgf_binary_pu function from the TAPAS toolbox, with optimization carried out via the quasi-Newton algorithm (Frässle et al., 2021). The objective was to maximise the log-joint posterior density of all perceptual and response parameters, conditioned on the observed data and the generative model.

Model comparison

Formal model comparison was conducted in two steps. Firstly, we compared four alternative perceptual models: a two-level HGF (HGF2), a three-level HGF (HGF3), a standard Rescorla-Wagner (RW) model, and a non-learning control model (RW0). In the RW0 model, the learning rate was fixed at zero, effectively treating all input as noise and simulating the absence of learning; this model served as a baseline to test whether participants could learn IEC contingencies (Table 1). After identifying the best-fitting perceptual model, we combined it with two response models based on the unit-square sigmoid family, differing in whether they received posterior or predicted beliefs as input. This process yielded a total of eight models tested across the two steps.

This two-step procedure was adopted because the response models were nested within the perceptual architecture - they simply re-parameterised the mapping from beliefs to choices without altering the belief-update process itself. This means the response models could be evaluated conditionally on the best-fitting perceptual model without needing to refit the full model space factorially. As such, this approach provides a principled and interpretable simplification equivalent to a full factorial model comparison, but with reduced computational cost.

Bayesian model selection (BMS) was performed using the spm_bms function from the SPM toolbox, applying random-effects BMS at the group level. For each model, we extracted both expected posterior probabilities (the average probability that a given model generated the data for a randomly selected participant) and exceedance probabilities (the likelihood that a given model is the best among those tested). Model evidence was quantified using

participants' log model evidence (LME), allowing us to determine which model provided the best account of the observed data.

Regression analysis for individual differences

To examine whether interoceptive sensibility and alexithymia were related to learning and decision parameters, we pooled the samples from Studies 1 and 2 (N = 88) and ran separate multiple linear regressions for each parameter (ω_2, ω_3, ΔLR_2 and ζ). Each regression included the eight MAIA subscale scores and the TAS-20 total score as predictors, and a factor coding Study (online vs lab) as a covariate. In addition, for each trait–parameter association we computed Bayes factors using Bayesian linear regression.

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.

Results

Model-free analysis

Study 1 - Accuracy

We conducted the same analyses for the online and lab studies on the ER accuracy data. In the online study, the 3x2x2 (Block x Trial x Noise) within-subjects repeated measures ANOVA revealed a significant main effect of Block ($F(2,84) = 6.291$, $p = 0.003$, $\eta^2_p = 0.124$), with participants being less accurate in the incongruent block compared to the neutral one, as shown by post hoc pairwise comparisons ($t(1,42) = 4.010$, $p < 0.001$). No significant differences were observed between the other blocks ($ps > 0.05$). We found a significant main effect of Trial ($F(1,42) = 9.147$, $p = 0.004$, $\eta^2_p = 0.156$), with incongruent trials producing more errors than the congruent trials. The main effect of Noise was significant as well ($F(1,42) = 15.395$, $p < 0.001$, $\eta^2_p = 0.272$), with participants, surprisingly, being more accurate in the high noise trials.

The Block x Trial interaction ($F(1,42) = 29.706$, $p < 0.001$, $\eta^2_p = 0.424$) was also significant. Pairwise comparisons of congruent vs incongruent trials within each block indicated higher

accuracy for congruent trials in the neutral block ($t(85) = 3.930$, $p = 0.048$, Cohen's $d = 0.343$), reflecting a small baseline EEB effect. This suggests that, in the absence of contextual constraints, i.e. 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 substantial decrease in accuracy for incongruent trials (Figure 3A). Conversely, in the expected incongruency block, accuracy was higher for the incongruent trials compared to the congruent ones, with the difference approaching significance ($t(85) = -1.742$, $p = 0.054$, Cohen's $d = 0.246$). The increased tendency for egocentric responses in the expected congruency block (large effect size vs small effect size in the Neutral block) and the opposite pattern in the expected incongruency block, reveal contextual adaptation and implicit learning of expectancy.

The Block × Noise interaction was also significant ($F(2,84) = 14.77$, $p < 0.001$, $\eta^2_p = 0.259$; Figure 3B), as accuracy increased with higher noise only in the expected congruency block ($t(85) = 5.72$, $p < 0.001$, Cohen's $d = 0.77$). No significant difference in accuracy between noise levels was found in the incongruent ($t(85) = 0.97$, $p = 0.33$, Cohen's $d = 0.12$) or neutral block ($t(85) = 0.38$, $p = 0.70$, Cohen's $d = 0.05$). The Trial × Noise interaction was not significant ($F(1,42) = 0.78$, $p = 0.38$, $\eta^2_p = 0.024$).

Finally, the Block x Trial x Noise interaction was also significant ($F(2,84) = 11.106$, $p<0.001$, $\eta^2_p = 0.219$; Fig4A), and to unpack it, we ran simple-effects tests comparing Low- versus High-Noise accuracy within each Block × Trial cell. Only one contrast was significant after applying Bonferroni for multiple comparisons: in the Congruent Block-Incongruent trials condition, accuracy was substantially lower in low noise trials ($t(503) = – 6.63$, $p < 0.001$,

Cohen's d = –1.46). All other Noise comparisons were non-significant (ps > 0.073), indicating that perceptual noise selectively amplified egocentric biases in that condition.

Study 2 - Accuracy

The 3x2x2 (Block x Trial x Noise) within-subjects ANOVA on accuracy for the lab study was significant ($F(2,86) = 19.672$, $p < 0.001$, $\eta^2_p = 0.281$,) with results largely mirroring those of Study 1. There was a significant main effect of Trial ($F(1,43) = 15.796$, $p < 0.001$, $\eta^2_p = 0.245$), with incongruent trials producing more errors than the congruent ones. The main effect of Noise ($F(1,43) = 36.049$, $p < 0.001$, $\eta^2_p = 0.405$) was also significant, with higher accuracy in the high noise condition. The main effect of Block approached significance ($F(2,86) = 2.879$, $p = 0.060$, $\eta^2_p = 0.010$).

As in study 1, there was a significant Trial x Block interaction ($F(2,86) = 21.790$, $p < 0.001$, $\eta^2_p = 0.321$), driven by significant difference between congruent and incongruent trials in the expected congruency ($t(89) = 6.589$, $p < 0.001$, Cohen's d = 0.982) and Neutral ($t(89) = 3.151$, $p = 0.002$, Cohen's d = 0.518) blocks but with no difference in the expected incongruency block ($t(89) = 0.247$, $p = 0.805$, Cohen's d = 0.039) (Fig. 3C). Similar to study 1 findings, the Block x Noise interaction ($F(2,84) = 14.767$, $p < 0.001$, $\eta^2_p = 0.259$) was significant (Fig.3D), 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 (ps>0.665).

Similarly to study 1, the 3-way Block x Trial x Noise interaction was also significant ($F(2,86) = 19.672$, $p < 0.001$, $\eta^2_p = 0.281$; Fig.4B). Simple-effects comparisons between Noise conditions within each Block × Trial cell showed that only in the Congruent Block-Incongruent trials

condition, accuracy was significantly lower in low noise trials ($t(498) = –6.63$, $p < 0.001$, Cohen's $d = –1.77$). All other Noise comparisons were non-significant ($ps > 0.113$).

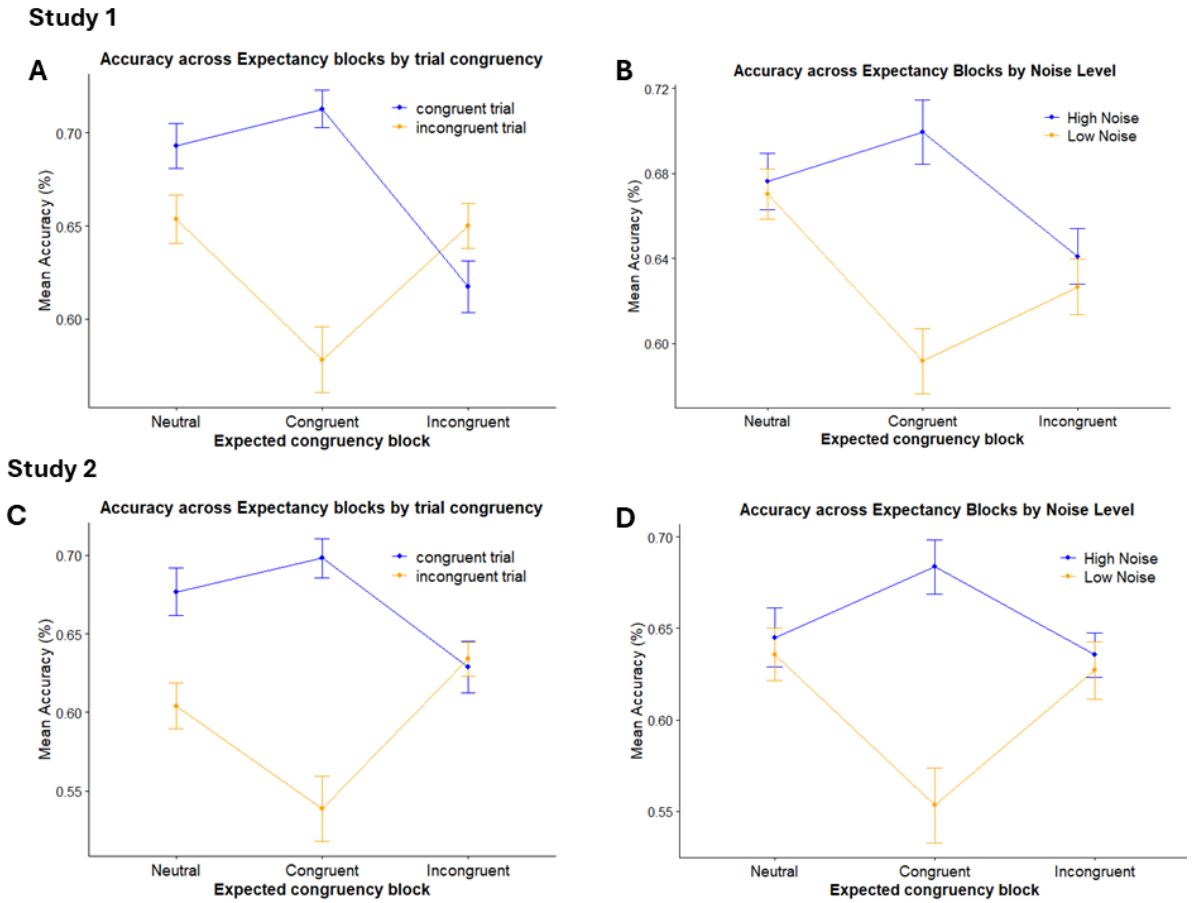


Figure 3. Accuracy across expectancy blocks, trial congruency, and stimulus noise in Studies 1 and 2. (A) Study 1: Accuracy by block and trial congruency (congruent vs. incongruent). Emotion congruency expectancy modulated emotion egocentricity bias (EEB), amplifying congruent trial accuracy when congruency was expected, and reducing it when incongruency was expected. (B) Study 1: Accuracy as a function of expected congruency block (neutral, congruent, incongruent) and stimulus noise (high vs. low). Accuracy was highest in the congruent block under high noise, suggesting increased reliance on contextual expectations when perceptual uncertainty is high.
(C) Study 2: Same as (A), showing similar modulation of EEB by contextual expectancy in a lab-based setting. (D) Study 2: Same as (B), replicating the interaction between perceptual noise and contextual congruency on emotion categorization accuracy.

Analyses on RTs yielded, for the most part, consistent results with those observed for accuracy in both studies, with the notable exception of quicker responses for congruent trials in incongruent blocks (see Supplementary Material S1). To explore the possible impact of RT differences across blocks we performed two types of speed-accuracy trade-off

analyses. Although slower responses were generally more accurate, we found no indication of context-dependent speed-accuracy trade-offs influencing accuracy (see Supplementary Material S2).

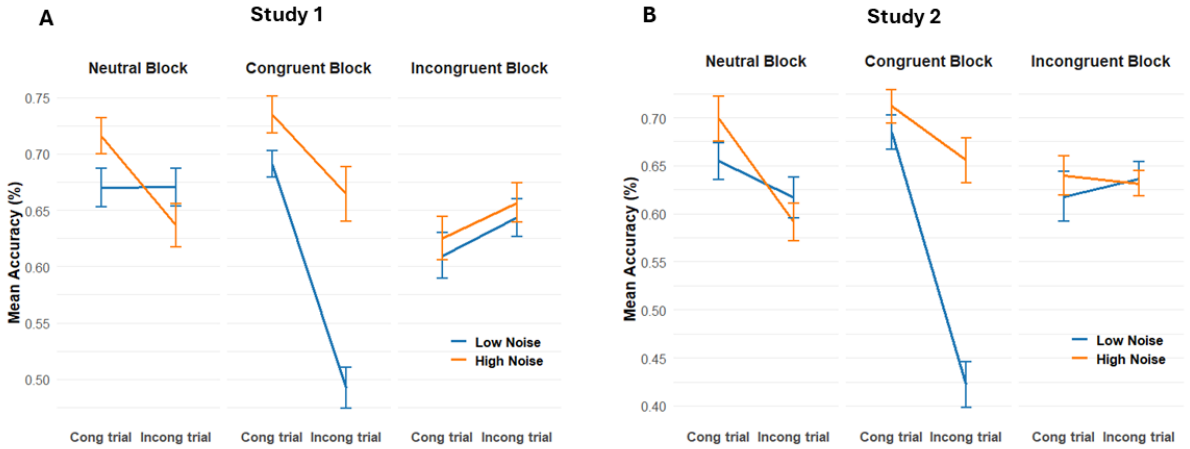


Figure 4. Three-way interaction between Congruency Block, Trial Congruency, and Noise Level on emotion recognition accuracy in Study 1 (A) and Study 2 (B). Each panel shows mean accuracy (± SE) for congruent and incongruent trials within each block (Neutral, Congruent, Incongruent), separately for low and high noise conditions. A robust interaction pattern emerged in both studies, with a pronounced drop in accuracy for incongruent trials under low noise in the Congruent block, suggesting heightened egocentric bias under stable congruency expectations and lower perceptual ambiguity. Accuracy was relatively stable across noise levels in the Incongruent and Neutral block.

Model-based analysis (HGF)

Bayesian Model Selection

Study 1

For the perceptual model comparison, the three-level HGF outperformed both the two-level HGF and the Rescorla-Wagner models, with an expected posterior probability of 0.77 and an exceedance probability of 1.0 (Fig. 5A). This suggests that participants' behaviour was best explained by a model incorporating volatility, indicating that they adapted their belief updating in response to changes in environmental volatility. In contrast, the zero-learning

model performed worst among all models (expected posterior probability = 0.03), providing strong evidence that participants were indeed learning the IEC contingencies.

Next, we combined the winning HGF3 model with two response models. Bayesian model selection indicated that the response model using posterior beliefs as input provided the best fit, with an expected posterior probability of 0.81 and an exceedance probability of 1.0, outperforming the models using priors (expected posterior probability = 0.19; Fig. 5B). This suggests that participants' decisions were primarily guided by their updated/posterior beliefs rather than by prior expectations alone.

Study 2

Consistent with Study 1, the three-level HGF (HGF3) again outperformed the two-level HGF, Rescorla-Wagner, and zero-learning models, with an expected posterior probability of 0.64 and an exceedance probability of 0.96 (Fig. 5C). For the response models, the Bayesian model comparison showed that the HGF3 combined with the response model using posteriors as input provided the best fit (expected posterior probability = 0.85, exceedance probability = 1.00), outperforming the model using priors only (expected posterior probability = 0.15; Fig. 5D).

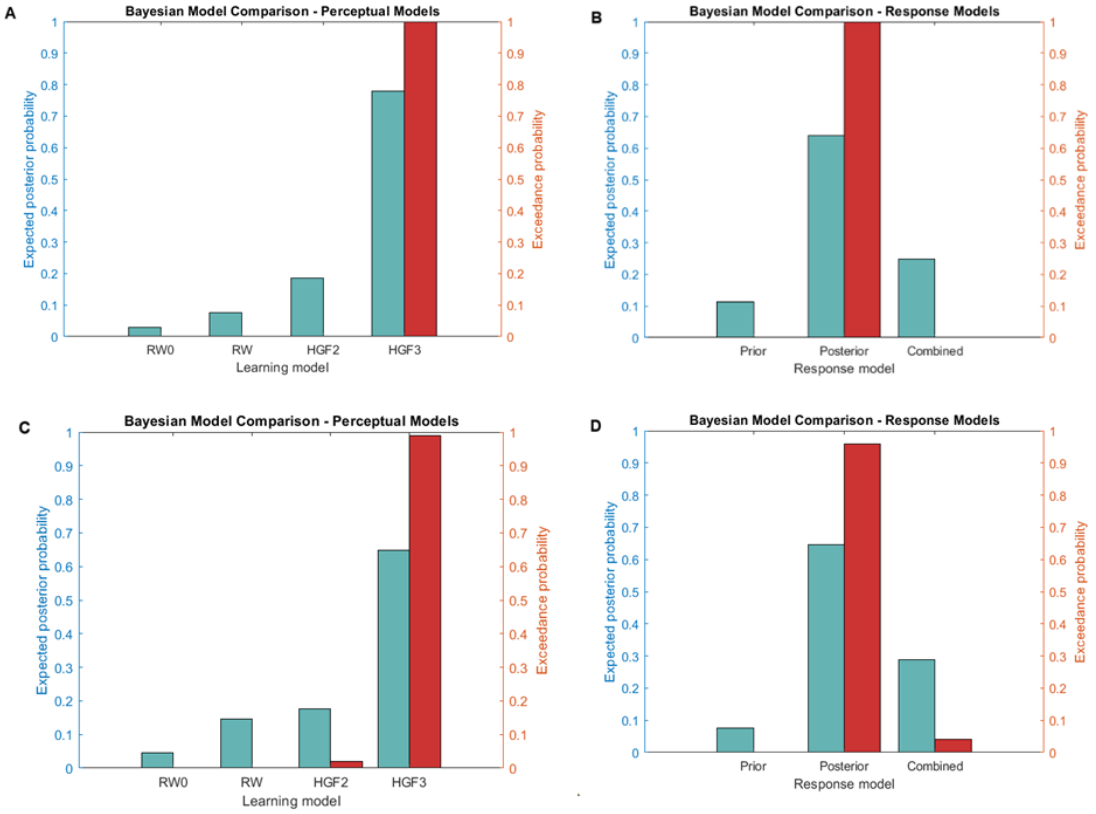


Figure 3. Bayesian model comparison for the learning and response models tested in study 1 (A,B) and study 2 (C,D). RW0 is the zero-learning Rescorla-Wagner model (α=0), RW is a Rescorla-Wagner model with stable learning rate, HGF2 does not include volatility representation, HGF3 is the typical HGF model 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. Results showed that the HGF3 model provided a better fit to the data, confirming that participants adapted their belief updating in response to changes in environmental volatility. The winning response model contained only posterior beliefs as input, revealing that participants' decisions were primarily guided by their updated beliefs rather than by prior expectations alone.

HGF and autonomic reactivity

To assess whether the changes in heart rate in response to the outcome of the roulette game were associated with trial-wise fluctuations in the learning rates, we conducted mixed-model analysis with the evoked HRA as DV and the level 2 and 3 learning rates as fixed effects. The results indicated level-2 learning rate as a significant predictor ($\beta = -0.014$, $t = -2.330$, $p = 0.020$), suggesting that lower physiological arousal was associated with increased

learning rate of IEC. The level-3 learning rate was not significant (β = -0.042, t = -1.393, p = 0.167).

In a similar analysis using belief trajectories at the second (μ_2^k) and third (μ_3^k) level of the HGF as predictors, level-3 beliefs were found to be significant (β = –0.03, t = –2.09, p = 0.037), indicating that lower perceived volatility was associated with stronger sympathetic responses. Notably, the significant predictors remained significant after including trial-wise RTs as covariates, suggesting that the observed associations between HRA and learning were not primarily driven by task engagement or alertness.

Finally, we explored how individual differences predicted the degree to which physiological arousal was associated with IEC learning. The regression on LR2 coefficients revealed Body Listening (β = 0.769, t = 2.425, p = 0.022) as significant predictor, with Trust also close to significance (β = 0.435, t = 2.026, p = 0.052); all other predictors were non-significant (ps > 0.170). For LR3 coefficients, Body Listening was again significant (β = 13.56, t = 2.59, p = 0.015), while other trait measures did not predict this relationship (ps > 0.20). Together, these findings suggest that individuals with a greater tendency to listen to and trust their bodily sensations exhibit a closer coupling between physiological arousal and learning of IEC.

Regression analysis for individual differences

To examine individual differences, we ran multiple regressions in the pooled sample (N = 88) with MAIA subscales and TAS-20 scores predicting each perceptual and response parameter, including Study (online vs lab) as a covariate. Study was a significant predictor in any model. Overall, these models explained little variance ($R^2 \approx .11–.21$) and yielded only weak associations. For the level-2 learning parameter ω_2, the regression model showed that

higher scores on the Not-Distracting MAIA subscale were associated with higher ω_2 estimates (Not-Distracting: $\beta = 0.74$, 95% CI [0.13, 1.36], $p = 0.019$, $BF_{10} = .34$), whereas all other subscales and alexithymia showed no reliable relationship with ω_2 (*p*s ≥ .11). For the level-3 volatility parameter ω_3, only the Noticing subscale showed a small positive association ($\beta = 0.06$, 95% CI [0.01, 0.12], $p = 0.019$, BF_{10}=1.34). None of the questionnaire predictors was significantly related to the change in learning rate ΔLR_2. For the decision noise parameter ζ, only MAIA Emotional Awareness showed a small positive association (β = 0.45, 95% CI [0.01, 0.90], $p = 0.046$, $BF_{10} = 0.66$), with all other predictors non-significant. Overall, they indicate that evidence for trait–parameter associations is anecdotal at best (BF_{10} range = 0.34 - 1.32), and therefore, these effects should be interpreted as exploratory only.

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; Von 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 trials, 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. Notably, the Block-Congruency interaction pattern persisted when controlling for RT and was reproduced using BIS, indicating that the behavioural effects reflect genuine performance changes rather than speed–accuracy trade-offs. These results suggest 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 (Supplementary material S5).

Previous EEB studies typically required explicit self–other distinctions (Trilla et al., 2020; Sevi et al., 2020; Von Mohr et al., 2021), 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 (Supplementary Material S1). 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 constraints are considered critical in EEB and social cognition more generally. For instance, Steinbeis and Singer (2014) showed that EEB can be enhanced by time constraints, 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 constraints 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 examining how interoceptive signals are implicated in social-learning dynamics, with interoception being widely associated with 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. 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 et al., 2019). One alternative interpretation is that elevated HRA may reflect increased motivation or attentional capture following emotionally salient outcomes. However, a control analysis including response times from the emotion recognition task (Supplementary Material S4) showed that the relationship between HRA and learning remained significant when accounting for RTs, suggesting that the observed effects are not likely driven simply by task engagement or arousal-related alertness. Instead, this relationship could reflect that conditions of heightened arousal may promote more reliance on priors and reduced learning of modulations in IEC. We could further speculate that this autonomic reactivity was driven by feelings of surprise to the game outcome; however, our design did not allow for measuring such violations of expectations in a trialwise basis. It should be further noted that, since our design is correlational, it does not allow us to determine whether changes in arousal causally alter learning rates, whether learning dynamics drive autonomic responses, or whether both are jointly influenced by a third process such as outcome surprise or engagement. Interestingly, 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 ability to noticing but not be distracted by their bodily sensations, exhibited higher sensitivity to changes of IEC. Additionally, individuals with heightened awareness of bodily-emotional connections relied more consistently on trial-wise updated beliefs about IEC during emotion categorization. However, it should be noted that these effects were modest in size, and not robust enough to survive correction for multiple comparisons, with Bayes factors suggesting only anecdotal evidence for the presence of effects. Moreover, parameter-recovery analyses suggested very good recoverability for ω_2 and ζ but only moderate recovery for ω_3, implying that null or weak associations with ω_3, in particular, may partly reflect noisier parameter estimates (see Supplementary Material S8). Taken together, our data do not provide strong support for stable trait-level links between specific dimensions of interoceptive sensibility and learning dynamics, but they remain compatible with the broader idea that bodily awareness can modulate socio-emotional processing and EEB (Shah et al., 2017; Stoica & Depue, 2017; von Mohr et al., 2020). We therefore regard these patterns related to individual differences as preliminary and hypothesis-generating, pointing to candidate dimensions that future work with larger samples and preregistered models could test more rigorously.

Despite the contributions of this 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? Real-world social interactions involve richer, reciprocal cues and multiple overlapping goals. Future research could explore whether similar biases occur in more naturalistic interactions, where several contextual and

interpersonal cues compete for attention. For instance, this can be examined within more socially embedded or interactive emotion induction methods, such as by modulating interpersonal attunement in real-time interactions (Bolis et al., 2023). 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.

In this context another area for improvement would be enhancing participants' emotional engagement with the game to induce emotional changes more effectively. The roulette game may have elicited only moderate emotional engagement: post-task ratings were, on average, around the midpoint of the scale and we did not collect trial-by-trial emotion ratings. We avoided frequent ratings to minimise cognitive load and prevent trial-wise self–other emotion comparisons from contaminating EEB measures. As a result, our findings are best interpreted as showing that relatively subtle affective states are sufficient to modulate interpersonal contingency learning and physiological responses at the group level, rather than as evidence for effects induced by strongly felt emotions. Future work could increase affective salience (e.g., explicit feedback about monetary gains and losses) and include intermittent emotion ratings to track within-task emotional dynamics while limiting interference with the social-judgement aspect of the task. 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 (Supplementary Material S5).

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 are closely associated with 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.

Author Contribution Statement

VK and RA designed the study. VK conducted testing and data collection. VK and DB analysed the data. VK, DB and RA wrote the manuscript. All authors approved the final version of the manuscript.

Declaration of Interest

The authors declare no competing interests.

Acknowledgments

This research was conducted as part of my PhD at the School of Psychology, University of Kent, and was supported by a Graduate Teaching Assistant (GTA) scholarship awarded by the University of Kent.

Data and code availability

All behavioural data, model outputs, and analysis scripts supporting the findings of this study are openly available via the Open Science Framework (OSF) at the associated osf project. The repository includes raw and processed datasets, Hierarchical Gaussian Filter (HGF) configuration files, and scripts for statistical analyses. Supplementary materials, including analysis of reaction times, detailed model specifications and debriefing response summaries, are also provided. Further materials and clarifications are available from the corresponding author upon reasonable request.

References

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

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

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.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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

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.

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

Hein, T. P., de Fockert, J., & Ruiz, M. H. (2021). State anxiety biases estimates of uncertainty and impairs reward learning in volatile environments. *NeuroImage*, *224*, 117424.

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

Kumar, D., & Srinivasan, N. (2011). Emotion perception is mediated by spatial frequency content. *Emotion*, *11*(5), 1144.

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

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

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.

Morrison, R. L., & Bellack, A. S. (1981). The role of social perception in social skill. *Behavior therapy*, *12*(1), 69-79.

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

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

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.

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

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

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

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.

Sevi, L., Stantic, M., Murphy, J., Coll, M. P., Catmur, C., & Bird, G. (2022). Egocentric biases are predicted by the precision of self-related predictions. *Cortex*, *154*, 322-332.

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

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156(1)*, 81-96.

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

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

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

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.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature neuroscience*, *6*(6), 624-631.

Weigand, A., Trilla, I., Enk, L., O'connell, G., Prehn, K., Brick, T. R., & Dziobek, I. (2021). How much of me do I see in other minds? Modulating egocentricity in emotion judgments by tDCS. *Brain Sciences*, *11*(4), 512.

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