Aberrant insula activity to negative and reduced learning from positive prediction errors as mechanisms underlying maladaptive self-belief formation in depression

Nora Czekalla[1,2], Alexander Schröder[1,2], Annalina V Mayer[1,2], Janine Stierand[1,2,3], David S Stolz[1,2], Tobias Kube[4], Christoph W. Korn[5], Ines Wilhelm-Groch[2,3], Jan Philipp Klein[2,6], Frieder M Paulus[1,2], Sören Krach[1,2]*, Laura Müller-Pinzler[1,2]*

1 Social Neuroscience Lab, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany.

2 Center of Brain, Behavior and Metabolism (CBBM), University of Lübeck, Lübeck, Germany.

3 Sleep and Mental Health, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany.

4 Department of Clinical Psychology and Experimental Psychopathology, Goethe University Frankfurt, Germany

5 Section Social Neuroscience, Department of General Psychiatry, University of Heidelberg, Heidelberg, Germany

6 Department of Psychiatry, Psychosomatics and Psychotherapy, University of Lübeck, Lübeck, Germany

Corresponding authors:

Nora Czekalla: n.czekalla@uni-luebeck.de

Laura Müller-Pinzler: la.muellerpinzler@uni-luebeck.de

Sören Krach: soeren.krach@uni-luebeck.de

Social Neuroscience Lab, Department of Psychiatry and Psychotherapy, Ratzeburger Allee 160, 23562 Lübeck,

Running title: Maladaptive Belief Formation in Depression

Keywords: depression, computational modeling, neuroimaging, self-belief updating, cognitive immunization, affect.

Abstract

Maladaptive self-beliefs are a core symptom of major depressive disorder. These beliefs are perpetuated by a negatively biased integration of self-related feedback. Understanding the neurocomputational mechanisms of biased belief updating may help to counteract maladaptive beliefs and the maintenance of depression. The present study uses a belief-updating task and functional neuroimaging to examine the neurocomputational mechanisms associated with self-related feedback processing in individuals with major depression and matched healthy controls. We hypothesized that increased symptom burden in depression is associated with negatively biased self-belief updating and altered neural tracking of social feedback. Our findings show that depression is related to reduced incorporation of unexpected positive feedback, with higher symptom burden alongside heightened insula reactivity to unexpected negative feedback. The interplay of increased neural responsiveness to negative feedback and the reduced learning from positive feedback provide new insights into cognitive distortions in depression and may explain the persistence of maladaptive self-beliefs and, thus, the maintenance of depression.

Introduction

Maladaptive, mostly negatively biased, and highly rigid self-beliefs are prominent in various mental disorders[1]. Humans act on their beliefs as if they reflect reality[2], which makes maladaptive beliefs a key maintaining factor of mental disorders[1]. These beliefs are a product of our learning history and are constantly updated in the face of incoming information that deviates from expectations[2–4]. Maladaptive beliefs in depression often center on a pessimistic view of oneself, like one's ability or one's future[1]. When the emphasis of this internal model is strong[5], one becomes insensitive to the external context, such as contradictory positive feedback information[6]. To understand the computational mechanisms of maladaptive self-belief formation[7,8], we assessed individuals with depression and matched healthy controls in a functional magnetic resonance imaging (fMRI) study. We used the Learning Of Own Performance (LOOP) task[7–9] with trial-by-trial performance feedback to make participants form novel beliefs about their abilities. On their way to their new self-beliefs, we investigated the brain's response to unexpected positive or negative performance feedback and measured accompanying affective states.

Negative self-beliefs are typical characteristics of depression and are maintained by negatively distorted information processing[10]. Accordingly, more negative belief-updating about one's future, abilities, or social popularity has been observed in experimental settings in depression[11,12]. More recently, mental disorders have been conceptualized as an interplay of maladaptive prior internal models with a maladaptive response to prediction errors[5,13], which is the mismatch between the expectation of any future event and the actual outcome. The persistence of negative beliefs in depression has been related to reduced updating after positive prediction errors[6], which means higher insensitivity to the context when incoming information is better than expected. Cognitive immunization against this unexpected

positive information is a cognitive mechanism that hinders the change of negative beliefs[6,14]. It reflects a reappraisal of unexpected feedback in a way that the initial expectation is maintained, for example, by devaluing positive feedback as an exception or by questioning its credibility.

The experience of prediction errors is linked to the experience of affect[15]. The current affective state can influence the updating of self-beliefs[8,16,17], which, in turn, contributes to how the individual feels. This results in a recursive influence of affect and belief-updating on each other[16,17]. Accordingly, the experience of more negative and less positive affect during learning has been linked to negatively biased updating of self-beliefs[8,18]. Depression is characterized by predominantly negative affect[19]. Correspondingly, the negative learning history is particularly present in this pervasive mood disorder, which makes a negative interpretation of incoming information appear especially plausible[6]. This negative interpretation bias can impact the integration of incoming information and trigger particularly strong emotional responses[20] in line with the negative interpretation, which further contributes to negative affect. When negative affect is experimentally induced, individuals show reduced flexibility in adjusting their beliefs in a positive direction[14].

The entanglement of negatively biased belief updating and affective experiences also manifests on the neural systems level[8]. In the context of depression, most studies focused on the neural processing of reward prediction errors and behavioral adaptation following rewarding information unrelated to a self-belief[21]. In line with the persistence of negative beliefs against conflicting positive information, many studies reported reduced reward learning[22] and reduced reward-related prediction error signaling in the ventral striatum and other brain regions, such as the ventral tegmental area, anterior cingulate cortex, and hippocampus in depression[23]. However, other studies found no support for altered reward

learning[24,25] or prediction error signaling in depression[25,26]. Similarly, in response to negative prediction errors, some studies reported increased activity[27], while others did not[25].

When processing self-related feedback, in particular, individuals with depression exhibit heightened sensitivity to negative feedback, such as social rejection[28]. When receiving negative feedback, they show increased activity in the amygdala, anterior cingulate cortex, left anterior insula, and left nucleus accumbens[28–30]. This heightened neural reactivity is associated with more negative self-perceptions, negatively biased information processing[29], and lower self-esteem[28]. Altered insula activity in individuals with depression has been discussed as a neural correlate of more painful emotions[31] and maladaptive emotion regulation in depression[29]. Additionally, inducing a depressed mood in healthy participants increases insula reactivity to negative information[32]. In individuals with social anxiety, which is also characterized by negative self-beliefs[33] and often comorbid with depression, activity in the insula mediated the effect of negative social feedback on self-belief updates[34]. These findings highlight the crucial role of emotion-related brain regions in processing self-related feedback, setting the stage for examining how such neural responses contribute to the formation of maladaptive self-beliefs in depression.

While prediction error processing has been widely studied in depression using reinforcement learning tasks, much less is known about its role in the formation of self-beliefs[16]. The current study builds on previous work by applying a well-established computational approach of trial-by-trial self-belief formation, the LOOP task[7–9], to a clinical sample with depression. To understand the mechanisms by which people arrive at their maladaptive self-beliefs, we aimed to test whether 1) depression is related to biased updating of self-beliefs about one's ability in response to positive and negative prediction errors, and 2) whether processes of belief formation are underpinned by altered neural

prediction error processing in depression. To investigate associations with prediction error processing in the brain, regions of interest (ROIs) in the insula and amygdala were defined based on our previous study with the same paradigm[8]. Based on other studies on reward learning and depression, we included the ventral striatum as an additional ROI. In line with negative self-beliefs and distorted information processing in depression, we expected more negatively biased self-belief formation accompanied by an unbalanced neural processing of positive and negative prediction errors. As we previously linked prediction error processing mapped in the insula and amygdala to biased self-belief formation and affect[8], we expected depression-related alterations in these regions.

Results

Participants diagnosed with major depressive disorder (MDD; *n*=35) and healthy control (CON) participants (*n*=32) completed the LOOP task[7,8] in the MRI (for sample characteristics, see Supplementary Table 1a-c). In this task, participants estimated attributes of objects in two out of four different estimation categories (e.g., heights of buildings). On each trial, they received manipulated performance feedback relative to an alleged reference group (see Methods). The feedback referred to their own (Self) or another person's estimation abilities (Other, Fig. 1A), who allegedly performed the task in an adjacent room. By linking estimation categories either with mostly "better than expected" or mostly "worse than expected" feedback, participants were led to form novel ability beliefs. The emotional response to self-related feedback was assessed with two ratings of happiness, pride, and embarrassment during the task. To describe the belief formation process, we modeled the changes in participants' expected performance through updates from prediction errors for each participant. Consistent with our previous studies[7,8,35], the winning model for both the depression and healthy control groups showed higher learning rates for negative compared to positive prediction errors (PE) in forming novel self-beliefs. This negativity bias was not present when participants formed beliefs about the other person's ability (PE-Valence x Agent interaction: $t(195)=-3.29$, $p<.001$; Supplementary Table 2; negativity bias: post-hoc $\alpha_{Self/PE-}$ vs $\alpha_{Self/PE+}$; $t(66)=4.78$, $p<.001$; complete model space in Supplementary Table 3, for model selection, see Supplementary Note 1, Supplementary Table 4).

No learning differences between individuals with depression and healthy controls

When comparing learning rates between the two groups, there was no difference in the overall extent of learning (main effect Group $t(65)=-.30$, $p=.764$) and no difference in the extent of the negativity bias (PE-Valence x Agent x Group interaction: $t(195)=.18$, $p=.861$; Fig.

154 1B and C, Supplementary Table 2). This is in line with the analysis of the expectation ratings
155 in a model-agnostic approach (Supplementary Note 2, Supplementary Table 5). It suggests
156 that people diagnosed with depression do not differ from healthy controls when forming
157 novel beliefs in response to self-related performance feedback.

158 **Disregard of positive feedback with higher symptom burden**

159 We additionally assessed self-reported symptom burden with a general measure of
160 depressive symptoms (BDI[36]) and a specific measure of cognitive symptoms (ATQ[37]), as well
161 as a measure of social anxiety (SIAS[38]) and self-esteem (SDQ-III[39]) as transdiagnostic markers,
162 already shown in association with biased self-belief formation before[7]. Variability of
163 symptoms was well explained by a single principal component (83% variance explained),
164 reflecting general depressive and socially anxious psychopathology, akin to other recent
165 approaches[40–42] (see Fig. 1D and Methods section).
166 Within the clinical sample, symptom burden was associated with reduced self-belief
167 updating following positive prediction errors (MDD: positive learning rate $r\alpha_{Self/PE+}$=-.51,
168 p=.002; for comparison, negative learning rate $r\alpha_{Self/PE-}$=.03; difference of absolute values:
169 z=2.1, p=.035; Fig. 1E, Supplementary Fig. 1. For more detailed correlations with symptom
170 scores, see Supplementary Fig. 2). This indicates that individuals with higher symptom
171 burden integrate positive feedback less, while learning from negative feedback remains
172 unaffected.

173 --- Fig. 1 about here ---

174

175 **Imbalanced prediction error signaling in depression during feedback processing**

Next, we examined the neural processing of self-related positive and negative PEs as a basis of self-belief formation (continuous PE effect). When comparing individuals with depression and healthy controls, a significant Group x PE-Valence interaction in the right anterior and posterior insula indicated an imbalance of negative relative to positive prediction error signaling in the clinical sample but not in healthy controls (anterior: p=.045, posterior: p=.030, family-wise [FWE] corrected at peak level within ROIs, Fig. 2, Supplementary Table 6, for baseline activations of PE effects, see Supplementary Table 7). This effect was not present in the amygdala and ventral striatum ROIs. There was no group difference for neural activity with more positive self-related prediction errors. For negative prediction errors, post-hoc t-tests suggest a stronger activity for individuals with depression than healthy controls (two-sample t-tests within ROIs, anterior: p=.059, posterior p=.014). The findings suggest that individuals with depression are more sensitive to negative prediction errors at the neural level (for general feedback effects of Self/Other, PE+/PE- feedback categories, see Supplementary Fig. 3, for whole brain analysis of negative compared to positive PE processing [continuous PE effect], see Supplementary Fig. 4).

Subsequent exploratory analysis within the clinical sample suggests that activation in these same regions is associated with reduced positive affect experienced during the task (for affect ratings in both groups, see Supplementary Table 8). In negative prediction error trials, stronger activation corresponds to lower pride (Spearman correlation between averaged parameter estimates within insula ROIs, right-anterior: r=-.32 [p=.058], posterior: r=-.49 [p=.003], left-anterior: r=-.36 [p=.034], posterior: r=-.48 [p=.004], Fig. 2, Supplementary Fig. 5). Correlations with happiness show a similar negative trend but do not reach significance. Notably, for positive prediction errors, stronger activation also aligns with decreased happiness, following the same pattern (right-posterior: r=-.39 [p=.019], left-posterior: r=-.41

200 [*p*=.014]). This pattern may suggest that deviations in either direction are processed
201 negatively. Correlations with symptom burden within the clinical sample show the same
202 directional pattern as the group comparison, indicating that higher symptom burden is
203 associated with greater activation following negative prediction errors, although not
204 statistically significant (negative PE trials: insula ROIs *r*=.23-.31 [*p*=.07-.18]; no meaningful
205 correlations for positive PE trials: *r*=-.01-.03 [*p*=.79-.97]). Generally, while participants receive
206 recurrent performance feedback, insula activation in negative prediction error trials is
207 correlated with the learning behavior as indicated by the negative learning rates (see
208 Supplementary Fig. 6).
209 Together, these behavioral and neural findings indicate that individuals with depression
210 show impaired integration of positive feedback, paralleled by heightened neural sensitivity
211 to negative feedback, highlighting a valence-specific imbalance in self-related learning
212 mechanisms.
213

214 --- Fig. 2 about here ---

Discussion

Maladaptive self-beliefs are a core feature of depression, yet the neurocomputational mechanisms underlying their formation and persistence remain unclear. Here, we examined the role of prediction error processing during belief formation using a computational modeling approach. Our findings suggest that individuals with depression exhibit reduced incorporation of positive self-related prediction errors with higher symptom burden alongside heightened insula reactivity to negative self-related prediction errors, potentially reinforcing maladaptive self-beliefs. These results provide new insights into cognitive distortions in depression and the neural mechanisms sustaining negative self-perceptions.

Both, the depression and the healthy control group show a self-specific negativity bias in belief updating. Our finding aligns with recent reward-learning studies of other groups that also do not find different learning patterns in depression and healthy controls[24,25] and extends it to the field of self-related learning. This observation challenges the notion of fundamental differences in basic learning mechanisms[43]. A key factor in understanding the discrepancy between earlier studies that specifically target self-beliefs and do find group differences, and our study, may be the task's unique design. Our task demands active task performance, creating the perception of control over feedback. Such task engagement can elicit specific self-related motivations, such as the motivation to improve one's ability and receive positive feedback[44] or the motivation to avoid failure[45]. Fear of failure can shift the focus toward negative feedback as a form of threat monitoring[46]. Additionally, we selected a domain with little prior experience to encourage novel belief formation, potentially leading to low confidence and a general sense of uncertainty during task performance. We consistently observed a negativity bias in belief updating, also in healthy individuals. It suggests that the transition from a healthy level of negative to a maladaptive self-belief

239 formation may be subtle, becoming more evident in dimensional analyses rather than
240 categorical group comparisons.
241 In the clinical sample, self-belief updates following positive prediction errors were reduced
242 with increased symptom burden. One explanation is that positive feedback simply is less
243 salient and therefore might draw fewer attention[47]. Alternatively, cognitive mechanisms
244 could actively contribute to the devaluation of positive feedback, resulting in a cognitive
245 immunization against it[48]. These findings suggest that sensitivity to context information,
246 such as social feedback, diminishes as symptom burden increases, potentially due to an
247 overemphasis on an entrenched negative internal model. As a result, learning from positive
248 prediction errors is reduced, and beliefs become biased toward prior models in a
249 confirmatory manner[49].

250 On the neural system level, we observed stronger tracking of negative prediction errors in
251 the insula in the clinical sample, despite no group differences in behavior. The stronger
252 neural response to negative prediction errors aligns with a general neural sensitivity to
253 negative (social) prediction errors in the insula[8], as well as heightened insula reactivity in
254 depression[50] and social anxiety[34,51]. Given that insula reactivity to negative prediction errors
255 was linked to reduced positive affect during task performance, the observed group
256 difference may reflect heightened emotional responsivity in depression. This corresponds
257 to the overall reduced positive affect in individuals with depression[19] and underscores the
258 notion of affected beliefs[8] - the entanglement of beliefs and affect.

259 The stronger neural and possibly affective reaction to more negative self-related feedback
260 can also be discussed in the context of emotional schemas and their role in shaping
261 emotional responses. Emotional schemas are formed in highly arousing emotional events
262 and can be triggered by learned signs associated with the events. This results in automatic

263 and exceptionally intensive emotional responses[20]. Depression is more likely related to early
264 maladaptive schemas rooted in an aversive learning history, such as experienced social
265 devaluation in response to failure[52]. Therefore, individuals may have emotional schemas that
266 are particularly sensitive to negative self-related feedback, which results in a stronger
267 emotional response and, thus, an overall lower positive affect during task performance. In
268 turn, the stronger reaction could indicate that the corresponding negative learning history
269 is particularly present, so that the emphasis on a negative internal model is strong, making
270 new information that deviates in the positive direction seem more implausible. The reduced
271 positive updating found on the behavioral level may support this notion, though there was
272 no such effect for positive prediction errors in the brain. However, the brain's response is
273 measured when the feedback is given, while the updating of the belief likely occurs at a later
274 time, somewhere between feedback, feedback processing, interpretation, and indication of
275 the new expectation. Given the alternating order of the positive and negative ability
276 conditions, belief updates measured in behavior may represent an amalgam of several trials.
277 If the reaction to negative feedback in the brain is strong, cognitive appraisal mechanisms
278 might come into play and influence self-belief updating even following positive prediction
279 errors. To better understand task-specific negative affective reactions to self-related
280 feedback, trial-by-trial fluctuations in emotions should be measured in future studies.

281 Our findings of unaltered positive prediction error processing in the VS align with some prior
282 research in depression, which has reported intact VS activity during reward learning
283 tasks[25,26]. Coupled with the above-mentioned similar learning behavior in both groups, our
284 results challenge the hypothesis of altered VS reactivity and prediction error learning deficits
285 in depression[22,53]. Instead, the findings suggest that the neural processing of positive self-

286 related prediction errors as a basis for forming positive self-beliefs remains intact in
287 depression, at least at this initial stage of processing.

288 What can be learned from the current findings for the therapeutic practice? The findings can
289 offer insights into addressing maladaptive beliefs therapeutically. As several aspects are
290 already present in existing psychotherapy practices, our findings offer new perspectives into
291 bridging the gap between research results and therapeutic interventions. Rather than
292 directly challenging the negative self-concepts of patients with depression (e.g., by
293 cognitive restructuring), they can be supported in becoming more context-sensitive so that
294 positive feedback that contradicts the internal model can be incorporated. This might be
295 done by practicing skills that help to become less emotionally reactive to established
296 negative self-beliefs, e.g., with the idea of defusion from Acceptance and Commitment
297 Therapy (ACT[54]), the detached mindfulness concept from Metacognitive Therapy[55], or
298 increasing sensitivity to the present context, as e.g., being implemented in mindfulness skills
299 in ACT or mindfulness-based cognitive therapy[56]. To update a maladaptive belief, it may be
300 necessary for a therapist to thoroughly examine the emotional and cognitive meaning
301 attributed to new experiences, support the understanding and regulation of intense
302 maladaptive emotional reactions, and identify potential cognitive distortions, such as
303 retrospectively devaluing positive experiences. These mechanisms can be discussed with
304 patients as a factor that maintains and continuously corroborates a negative self-image.

305 In conclusion, this study provides insight into neurocomputational mechanisms that
306 contribute to maladaptive self-belief formation in depression. Our findings highlight a
307 heightened neural response to more negative self-related feedback in the insula, potentially
308 reflecting a more affectively charged processing style. Alongside reduced consideration of
309 positive feedback with increasing symptom burden, it may promote negatively biased self-

belief formation in depression in the long run. In the absence of fundamental deficits in learning behavior in depression, this effect underscores the importance of accounting for within-group variance and symptom burden within a clinically relevant range. Insights from the interaction between cognitive processes, affective experiences, and symptom burden towards forming maladaptive self-beliefs can be of great importance for future therapeutic intervention strategies.

Methods

Participants.

The study was approved by the ethics committee of the University of Lübeck (AZ 18-066), was carried out in accordance with the ethical guidelines of the American Psychological Association, and all participants gave written informed consent. All participants were adults above the age of 18 and fluent in German. The clinical sample (n=35, 9 females, aged 20–55 years; M=34.23; SD=10.54) was recruited from the Department of Psychiatry and Psychotherapy at the university clinic Lübeck. The goal in recruiting was to obtain a naturalistic sample with variance in depressive and social anxiety symptoms. Inclusion criteria were a diagnosis of depression (F32.1/2, F33.1/2; given by medical or psychological therapists at the clinic); comorbid symptoms of social anxiety (F40.1, F60.6) were preferentially included but were not a necessary inclusion criterion. Exclusion criteria were acute suicidality, an acute psychotic state, schizophrenia, schizotypal disorder or delusional disorder, bipolar disorder, primary diagnosis of substance abuse or substance dependence, narcissistic, histrionic, or borderline personality disorder, and neurological disease. Subjects of the control group (n=32, 10 females, aged 18–55 years; M=34.25; SD=10.23, no group differences in age or gender distribution p>.8, Supplementary Table 1a) were recruited at the university campus of Lübeck or via public notices. The exclusion criterion was mental illness or neurological disease. Individuals interested in the study were screened for symptoms of depression through a telephone interview, and the second part included a general screening for other mental illnesses adapted from a structured interview (SCID; for additional sample descriptions, see Supplementary Note 3, Supplementary Table 1).

Learning Of Own Performance (LOOP) task.

The LOOP task assesses how participants learn about their own or another person's purported estimation ability through trial-by-trial performance feedback. It was covered as a cognitive estimation experiment. The task was previously validated in behavioral and fMRI studies on healthy participants[7,8,35]. Two participants were invited, one performing the task in the scanner and the other completing a behavioral version (data not included). When a second participant was unavailable, a confederate was introduced as the behavioral participant. Participants alternated between two trial types: actively performing the estimation task (*Self*-trial) or observing another person's performance (*Other*-trial). The task involved estimating attributes in four categories: house height, animal weight, vehicle distance, and food quantity. Feedback was manipulated such that, for each agent, one category was paired with predominantly positive feedback (*High Ability*), while the other was paired with mostly negative feedback (*Low Ability*), resulting in four feedback conditions: *Self-High*, *Self-Low*, *Other-High*, and *Other-Low* (20 trials each). Each feedback condition was linked to a specific estimation category. Trials of all conditions were intermixed in a fixed order, preventing more than two consecutive trials of the same condition. Feedback, presented as percentiles relative to an alleged reference group of 350 former participants, was determined by a sequence of predefined PEs based on participants' evolving ability belief. This was calculated as a moving average of the last five performance expectations per category, initialized at 50%. This approach ensured that PEs remained largely independent of performance expectations and that negative and positive PEs were relatively equally distributed across groups and conditions (see Supplementary Note 4).

Each trial began with a cue indicating the estimation category and agent. Participants then rated their performance expectation (in percentiles) and were incentivized with up to 6

365 cents per trial for accurate predictions of their performance. The estimation question was
366 presented for 10 s, with responses made on a continuous scale within a predefined plausible
367 range for each question. In *Other*-trials, participants also provided estimates to ensure
368 comparable mental effort across both agent conditions. Subsequently, feedback was
369 presented for 3 seconds (Fig. 1A). In *Other*-trials, the other's feedback was visible, while their
370 performance expectations and answers to the estimation questions remained private.
371 Jittered inter-stimulus-intervals were presented following the cue (mean: 4*TR [0.992 s],
372 range: 2–6*TR), estimation (mean: 4.5*TR, range: 2.5–6.5*TR), and feedback phase (mean:
373 6*TR, range: 4–8*TR) with jitters distributed in a uniform distribution with steps of 0.5*TR.

374 To assess affective responses, participants rated their current levels of embarrassment,
375 pride, happiness, arousal, and tiredness on a continuous scale (0-100) four times, each
376 following one trial per feedback condition. The two ratings following *Self*-trials were
377 averaged to measure affective responses to self-related feedback. The task was
378 implemented using MATLAB (Release 2015b, The MathWorks, Inc.) and the Psychophysics
379 Toolbox (Brainard, D. H. [1997]. The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436). The
380 task comprised two ~25-minute sessions with a short break in between.

381 **Statistical analysis.**

382 *Behavioral data analysis and modeling.*

383 *Model-agnostic analysis.*

384 First, we checked for group differences in prior estimation ability beliefs (Supplementary
385 Note 5, Supplementary Table 1b), other baseline characteristics like estimation experience
386 and importance, and social perceptions of the other (Supplementary Table 1c). Then, we ran
387 a model-agnostic analysis on the participants' expected performance ratings for each trial.
388 We employed a linear mixed model with a maximum likelihood estimation. The model

389 included the factors Ability (High/Low) x Agent (Self/Other) x Group (MDD/CON) and Trial
390 (20 trials) as a continuous predictor. Intercept, Ability, Agent, Trial, and Group were modeled
391 as fixed effects; the intercept was additionally modeled as a random effect (Supplementary
392 Note 2, Supplementary Table 5).

393 *Computational modeling*.

394 Dynamic changes in performance expectation ratings were modeled using PE delta-rule
395 update equations (adapted Rescorla–Wagner model[57]). The model space and the procedure
396 of model fitting and selection have been implemented and further developed in our
397 previous studies[7,8]. The model space included three main approaches: the Unity Model
398 (single learning rate), the Valence Model (distinct learning rates for positive vs. negative
399 prediction errors), and the Ability Model (distinct learning rates for ability conditions). A
400 Mean Model served as a baseline, assuming stable values. Learning rates were estimated
401 either separately for Self vs. Other or across both agents (for detailed model space
402 description, see Supplementary Note 6, Supplementary Table 3). The winning model was a
403 Valence Model with separate learning rates for Self vs. Other. It was further extended by
404 adding a weighting factor that reduced the learning rates when feedback values
405 approached the extremes of the scale (percentiles close to 0% or 100%), assuming that
406 participants would perceive extreme feedback to be less likely than average feedback[58].
407 Since many variables encountered in everyday life approximately follow a normal
408 distribution where extreme values are less probable, we assigned the relative probability
409 density of the normal distribution to each feedback percentile value. A weighting factor w
410 was fitted for each individual, indicating how strongly the relative probability density
411 reduced the learning rates for feedback further away from the mean. The Weighted Valence
412 Model was the winning in our last study[8]. The normal decay (ND) weighted by the weighting
413 factor w was introduced in the learning models in the following way:

414 (EXP=Performance expectation rating, FB=feedback, PE=prediction error, α=learning rate):

415 $$EXP_{t+1=}EXP_t + \alpha\ PE_t\ (1 - w\ ND);\ \text{while}\ PE_{t=}FB_t - EXP_t$$

416 The initial beliefs about the own and the other participant's performance (EXP_1) were
417 estimated as free parameters separately for Self and Other in both Ability conditions,
418 resulting in four additional model parameters.
419 For model fitting, we used RStan, employing Markov Chain Monte Carlo sampling to
420 estimate individual learning parameters. Across three chains, 2400 samples were drawn
421 after 1000 burn-in samples (thinned by 3). Convergence was assessed via $\hat{R}$ values[59], with
422 effective sample sizes (n_{eff}) > 1500 for most parameters and subjects. Posterior distributions
423 were summarized by their mean for group-level analysis.
424 For model selection, we used leave-one-out cross-validation (LOO[60]) with Pareto smoothed
425 importance sampling (PSIS) to assess the prediction accuracy of fitted models. $\hat{k}$ values
426 evaluated PSIS-LOO reliability, showing only a few trials with insufficient values and, thus,
427 potentially unreliable PSIS-LOO scores (Supplementary Table 4). Bayesian Model Selection
428 on PSIS-LOO scores was performed on the group level for the whole sample and both sub-
429 samples[61] (Supplementary Note 1). The winning model was validated by replicating the
430 model-agnostic analyses with its predicted data (see Supplementary Note 7; for model
431 parameter correlations, see Supplementary Fig. 7).

432 *Capturing symptom burden.*

433 For a dimensional perspective on depression, we assessed the severity of depressive
434 symptoms with the BDI and ATQ. To address symptom burden in a broader context, we
435 additionally assessed symptoms of social anxiety with the SIAS and self-esteem with the
436 SDQ-III (subscale scores) as we could show a relationship with belief updating behavior and
437 these two concepts in a previous study[7]. To reduce dimensionality, we ran a principal

438 component analysis on our questionnaire data using the *R* package '*psych*' (varimax rotation,
439 component scores based upon the structure matrix [default], William Revelle (2023). psych:
440 Procedures for Psychological, Psychometric, and Personality Research. Northwestern
441 University, Evanston, Illinois. R package version 2.3.3, https://CRAN.R-
442 project.org/package=psych). We extracted one component as a combined measure of
443 depressive and socially anxious psychopathology (Fig. 1E).

444 *Statistical analyses of learning parameters.*

445 To test whether the updating of self-beliefs was different between groups, learning rates of
446 the winning models for positive and negative PEs (factor PE-Valence) and for Self and Other
447 (factor Agent) were compared between the two groups in a PE-Valence x Agent x Group
448 linear mixed model. For a more detailed understanding of the relationship between biased
449 belief updating and symptom burden, we additionally performed Spearman correlations
450 between the Psychopathology score and the self-related learning rates $\alpha_{Self/PE+}$ and $\alpha_{Self/PE-}$
451 within the clinical samples. Since we had a particular interest in the relationship between
452 the Psychopathology score and the self-related positive learning rate as a potential measure
453 of cognitive immunization against positive feedback, we additionally tested the absolute
454 correlation parameter of $\alpha_{Self/PE+}$ and Psychopathology against the one with $\alpha_{Self/PE-}$ within the
455 clinical sample (*r* package *'cocor'* that implemented to correlation comparison approach by
456 Pearson and Filon, 1898; Diedenhofen, B. & Musch, J. (2015). cocor: A Comprehensive
457 Solution for the Statistical Comparison of Correlations. *PLoS ONE, 10*(4): e0121945). Statistical
458 tests were performed two-sided. Statistical analyses on the behavioral data were performed
459 using *R* (R Core Team [2022]. R: A language and environment for statistical computing. R
460 Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/).

461

FMRI data acquisition.

FMRI data were collected at the Center of Brain, Behavior, and Metabolism at the University of Lübeck, Germany, using a 3 T Siemens MAGNETOM Skyra scanner (Siemens, München, Germany) with 60 near-axial slices. 1516 functional volumes (min=1400, max=1724) were acquired on average per session using echo planar imaging (EPI, TR=0.992 s, TE=28 ms, flip angle=60°, voxel size=3×3×3mm^3, simultaneous multi-slice factor 4). A high-resolution anatomical T1 image was obtained for normalization purposes (voxel size =1×1×1mm^3, 192 × 320 × 320 mm^3 field of view, TR=2.300 s, TE=2.94 ms, TI=900 ms; flip angle=9°; GRAPPA factor 2; acquisition time 6.55 min).

FMRI data analysis.

FMRI data were preprocessed and analyzed with Statistical Parametric Mapping 12 (SPM12, Wellcome Trust Centre for Neuroimaging, University College London). Field maps were recorded to obtain voxel displacement maps (VDMs) to correct for geometric distortions. EPIs were slice-time corrected, motion-corrected, and unwarped using the corresponding VDMs, co-registered with the T1 image, and normalized using the forward deformation fields as obtained from the unified segmentation of the anatomical T1 image. The normalized volumes were resliced with a 2×2×2 mm^3 voxel size and smoothed using an 8mm, full-width-at-half-maximum isotropic Gaussian kernel. Functional images were high-pass filtered at 1/384 to remove low-frequency drifts.

A two-level, mixed-effects procedure was implemented for the statistical analyses. On the first level, a fixed-effects GLM included four regressors for the cue conditions (Ability: High vs. Low × Agent: Self vs. Other), weighted with the performance expectation ratings per trial as parametric modulator, four regressors for the feedback conditions (Agent [Self /Other] × PE-Valence [PE+/PE-], categorical PE effect), weighted with PE magnitude per trials

486 (continuous effect of the unsigned PE values for each feedback condition), one regressor for
487 the performance expectation rating phase, two for the estimation period for Self and Other,
488 and one for the emotion rating phase. Parametric modulators were not orthogonalized;
489 thus, each only explained their specific variance. Regressors were modeled with the duration
490 as presented during the experiment (cue phase: 2.5s, performance expectation rating:
491 individual reaction times with *M*=4.0s, *SD*=2,4, estimation phase: 10s, feedback phase: 3s,
492 emotion rating phase: *M*=26.2s, *SD*=10.1). Six additional regressors were included to correct
493 for head movement, and one regressor was included with a constant term for each of the
494 two sessions.

495 On the second level, we first compared the brain activity for self- vs. other-related positive
496 and negative PEs in the feedback phase (categorical PE effect) in a flexible factorial design
497 (regressor for the four conditions Agent [Self /Other] × PE-Valence [PE+/PE-]) to test whether
498 results of our previous study[8] replicate (Supplementary Fig. 4). To compare the tracking of
499 positive and negative self-related PEs captured by the parametric modulator of PE
500 magnitude (continuous effect) between the MDD and CON group, we tested the PE-Valence
501 × Group interaction in a flexible factorial design within our ROIs. We compared activity
502 between MDD and CON separately for positive and negative PEs in post-hoc two-sample t-
503 tests. In a subsequent exploratory analysis, we correlated brain activity in self-related
504 positive and negative PE trials (categorical PE effect, extracted and averaged parameter
505 estimates of all voxels within one ROI) with $\alpha_{Self/PE+}$ and $\alpha_{Self/PE-}$, respectively, as well as the
506 Psychopathology score and the emotion ratings using Spearman correlation.

507 The selection of regions of interest was guided by our previous fMRI study with a healthy
508 sample using the same paradigm[8]. As the insula has emerged as particularly relevant in this
509 study, it was again defined as an ROI. We used unilateral insula ROIs as described in the

510 three-cluster solution of Kelly and colleagues[62], with their dorsal and ventral part combined
511 into an anterior ROI and a posterior part. Other ROIs motivated by this study are in the
512 amygdala (two unilateral ROIs from the AAL atlas definition in the WFU PickAtlas[63]). Since
513 altered PE processing in the ventral striatum has been discussed a lot in the context of
514 depression, a functional ROI within the ventral striatum (according to the SPM anatomy
515 toolbox) capturing the processing of PE valence was additionally defined as ROI. It derived
516 from a family-wise error $p<.05$ corrected baseline effect of the continuous signed self-related
517 PEs of our previous study[8].

518 FMRI results were family-wise error (FWE) corrected at peak level for the whole brain or
519 within the ROIs. Coordinates are reported in the MNI space. Anatomical labels of all resulting
520 clusters were derived from the SPM Anatomy toolbox, version 3.0 (Eickhoff, S. B., Stephan, K.
521 E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. [2005]. A new SPM toolbox for
522 combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage,
523 25(4), 1325-1335).

524

Data and Code availability

The behavioral data and code used to generate the behavioral analyses are available at https://osf.io/gwkp3/ (Identifier: DOI 10.17605/OSF.IO/GWKP3). For a quick overview, see *R Markdown/Quarto* output document. Neural data and analysis scripts, as well as *R Stan* scripts to estimate the computational models are available from the corresponding author upon reasonable request.

Acknowledgments

We are grateful to Rebecca Rocksien, Katharina Ohm, Ann-Kristin Schmidt, Jovana Lehmann-Grube, Alica Steinert, Leonie Christine Pasquier, and Yana Schwarze for their help with data collection. We thank Yana Schwarze for providing a Matlab analysis script to check the MRI motion parameters. The research was funded by the German Research Foundation (Temporary Positions for Principal Investigators: MU 4373/1-1; Sachbeihilfe KR 3803/11-1; 3803/14-1) and the Medical Department of the University of Lübeck (J21- 2018).

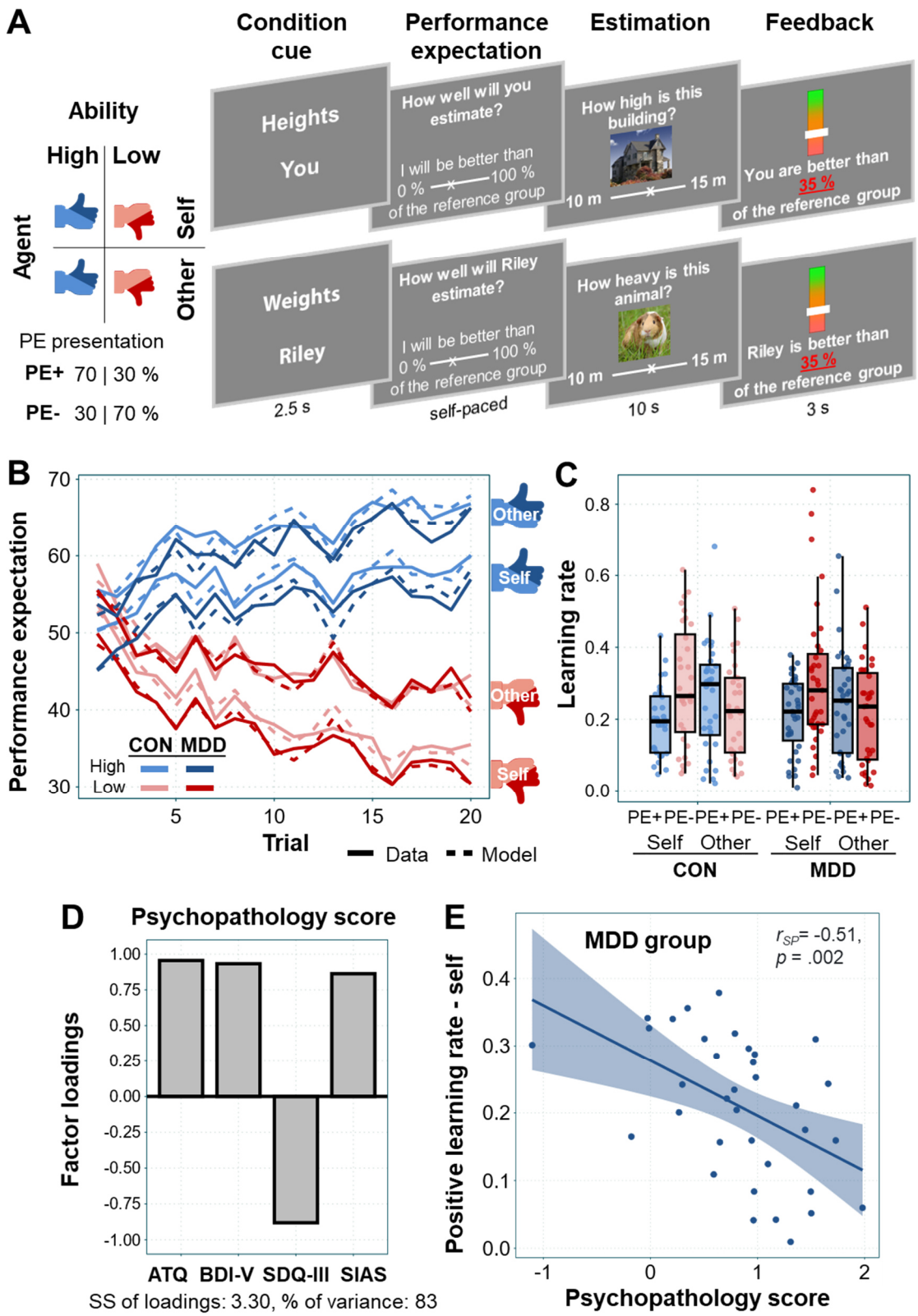
Author Contributions

L. M.-P., S. K., F. M. P., N. C. conceived and designed the experiments. N. C., A. S., J. S. collected the data, with the consultation of J. P. K. to recruit the clinical sample. N. C. analyzed the data under the supervision of L. M.-P., with the consultation of F. M. P.. N. C., A. S., A. V. M., T. K., F. M. P., S. K., and L. M.-P. discussed the data analyses and interpretation of the results. N. C. wrote the original draft of the manuscript. L. M.-P. and S.K. supervised and edited the paper. A. S., A. V. M., J. S., D. S. S., T. K., C.W.K., I. W.-G., J. P. K., F. M. P. reviewed and edited the paper.

Declaration of interests

The authors declare no competing interests.

547 **Figures and Legends**



548

549 **Figure 1.** Trial sequence of the Learning-Of-Own-Performance (LOOP) task, experimental
550 conditions, and behavioral results. **A** Low and High Ability conditions of alleged estimation

performance for Self and Other (Agent condition) were presented in pseudo-randomized order. Feedback with controlled PEs: High Ability: 70 %, Low Ability: 30 % of the trials with planned positive PEs (the actual percentage of positive/ negative PEs could slightly differ, e.g., if the feedback had been out of range). Sequence of one self- and one other-trial from left to right: At the beginning of each trial, a screen displays the estimation category of the upcoming trial and whether participants have to perform a self-trial or an other-trial. This is followed by a rating of the expected performance (performance expectation). Next, participants are asked to estimate the specific attributes of objects (estimation), which is followed by a manipulated performance feedback (feedback). **B** Predicted and actual ratings of expected performance for both groups over time. The ratings indicate that participants update their expected performance (solid lines) according to the feedback, thus forming a belief about the performance levels in the different estimation categories. The winning model's predicted values (dashed lines) with separate learning rates for Agent and PE valence captured the participants' behavior, as indicated by a close match of actual ratings and predicted values. **C** Learning rates of the winning model show a bias towards increased updating in response to negative PEs (PE-) in contrast to positive PEs (PE+), specifically for self-beliefs in both groups. Colored bars indicate the first and third quartiles of the data; the line marks the median. Whiskers extend from the upper/ lower box borders to the largest/ smallest data point at most 1.5 times the interquartile range above/ below the respective border. MDD=Major Depressive Disorder (n=35), CON=control group (n=32). **D** Factor loadings of the principal component analysis, by which one factor, a summary measure of symptom burden, was extracted from the questionnaire variables (Automatic Thought Questionnaire, ATQ, Beck's depression inventory, BDI-V, Self-Description Questionnaire-III, SDQ-III subscale scores, Social Interaction Anxiety scale, SIAS). SS=sum of squares. **E** Correlation of the positive learning rate ($\alpha_{Self/PE+}$) and the Psychopathology score in the clinical sample. Spearman correlation coefficient, line for displaying purposes.

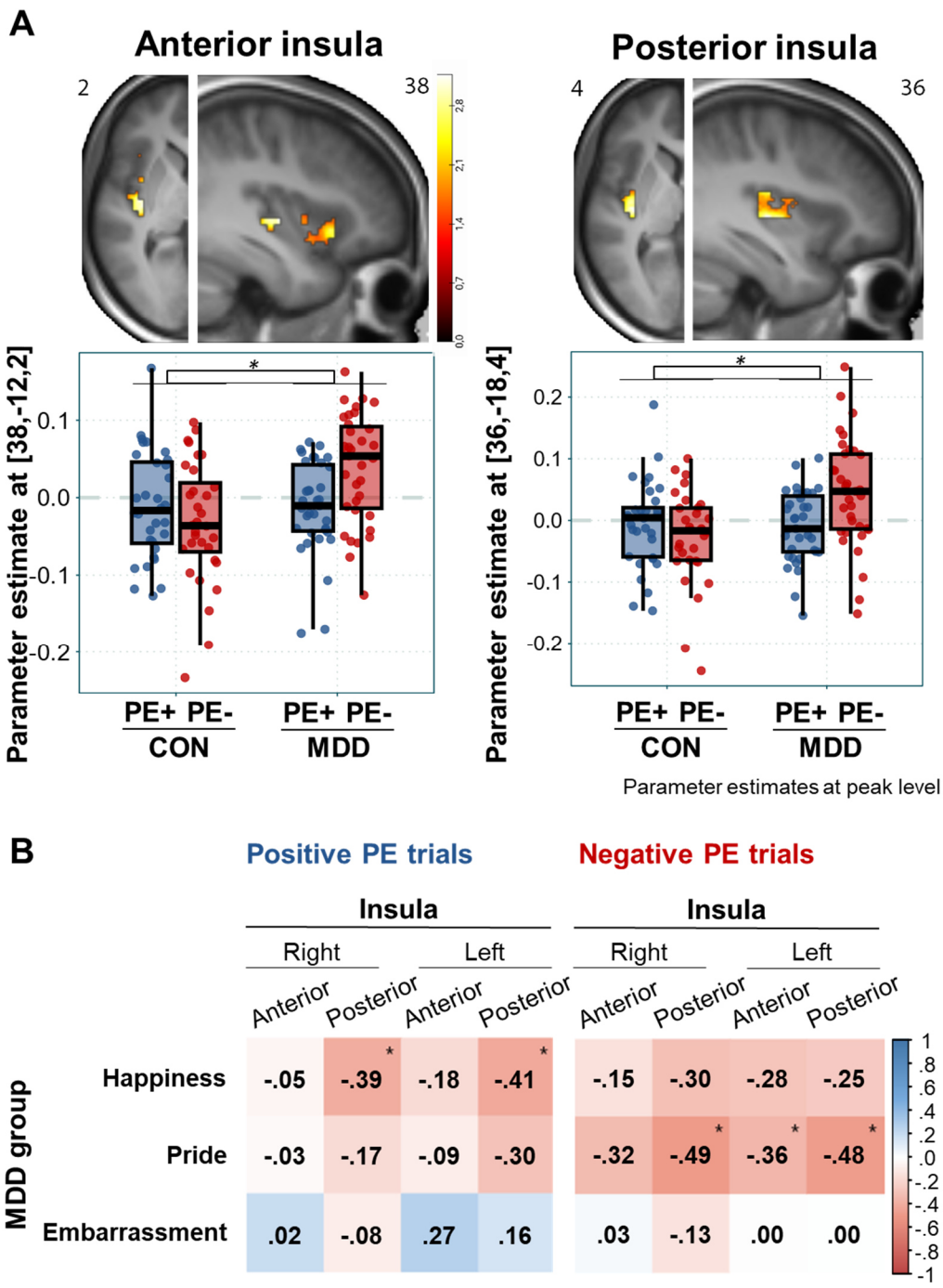


Figure 2. Differential insula tracking of PE valence and correlations with affective state. **A** The interaction (MDD[PE-]>MDD [PE+])>(CON [PE-]>CON [PE-]) shows a greater difference between PE- vs. PE+ in the MDD compared to the CON group (FWE corrected within ROIs at peak level; ROI selection guided by previous fMRI study[8]). CON=control group (n=32), MDD=Major Depressive Disorder (n=35). The plot depicts the parameter estimates of the Group [CON, MDD] x PE-Valence [PE+, PE-] interaction at peak level for the two regions of interest in the right insula. Brain plot: uncorrected $p<0.05$ within ROI for display purposes; see Supplementary Table 6 for FWE corrected statistics. **B** Spearman correlation between the affect ratings and brain activity within insula ROIs in response to positive and negative prediction errors (categorical PE effect) in the clinical sample. * $p < .05$.

References

1. Beck, A. T. *Cognitive Therapy of Depression*. (Guilford Press, 1979).
2. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**, 191–215 (1977).
3. Markus, H. R. & Wurf, E. The dynamic self-concept: a social psychological perspective. *Annu. Rev. Psychol.* **38**, 299–337 (1987).
4. Wittmann, M. K. *et al.* Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91**, 482–493 (2016).
5. Clark, J. E., Watson, S. & Friston, K. J. What is mood? A computational perspective. *Psychol. Med.* **48**, 2277–2284 (2018).
6. Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A. & Rief, W. Distorted cognitive processes in major depression: a predictive processing perspective. *Biol. Psychiatry* **87**, 388–398 (2020).
7. Müller-Pinzler, L. *et al.* Negativity-bias in forming beliefs about own abilities. *Sci. Rep.* **9**, 14416 (2019).
8. Müller-Pinzler, L. *et al.* Neurocomputational mechanisms of affected beliefs. *Commun Biol* **5**, 1241 (2022).
9. Müller-Pinzler, L. *et al.* Neural pathways of embarrassment and their modulation by social anxiety. *Neuroimage* **119**, 252–261 (2015).
10. Beck, A. T. Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of general psychiatry* **9**, 324–333 (1963).
11. Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R. & Dolan, R. J. Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol. Med.* **44**, 579–592 (2014).

612 12. Garrett, N. *et al.* Losing the rose tinted glasses: neural substrates of unbiased belief
613 updating in depression. *Frontiers in human neuroscience* **8**, 639 (2014).
614 13. Sterzer, P. *et al*. The predictive coding account of psychosis. *Biol. Psychiatry* **84**, 634–643
615 (2018).
616 14. Kube, T. & Korn, C. Induced negative affect hinders self-referential belief updating in
617 response to social feedback. *Emotion* (2024) doi:10.1037/emo0001426.
618 15. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum.
619 *Trends Cogn. Sci.* **20**, 15–24 (2016).
620 16. Bromberg-Martin, E. S. & Sharot, T. The value of beliefs. *Neuron* **106**, 561–565 (2020).
621 17. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood
622 instability. *Nat. Commun.* **6**, 6149 (2015).
623 18. Bottemanne, H. *et al.* How mood shapes belief updating bias in depression. *Cogn. Affect.*
624 *Behav. Neurosci.* 1–15 (2025).
625 19. American Psychiatric Association A. Diagnostic and statistical manual of mental
626 disorders: DSM-IV. **4**, (1994).
627 20. Greenberg, L. S. Emotion-Focused Therapy: A Clinical Synthesis. *Focus* **8**, 32–42 (2010).
628 21. Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin.*
629 *Neurobiol.* **18**, 185–196 (2008).
630 22. Admon, R. & Pizzagalli, D. A. Dysfunctional reward processing in depression. *Curr Opin*
631 *Psychol* **4**, 114–118 (2015).
632 23. Kumar, P. *et al.* Abnormal temporal difference reward-learning signals in major
633 depression. *Brain* **131**, 2084–2093 (2008).

24. Rouhani, N. & Niv, Y. Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology* **236**, 2425–2435 (2019).

25. Rothkirch, M., Tonn, J., Köhler, S. & Sterzer, P. Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain* **140**, 1147–1157 (2017).

26. Rutledge, R. B. *et al.* Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry* **74**, 790–797 (2017).

27. Ubl, B. *et al.* Altered neural reward and loss processing and prediction error signalling in depression. *Soc. Cogn. Affect. Neurosci.* **10**, 1102–1112 (2014).

28. Kumar, P. *et al.* Increased neural response to social rejection in major depression. *Depress. Anxiety* **34**, 1049–1056 (2017).

29. Jankowski, K. F. *et al.* Feeling left out: depressed adolescents may atypically recruit emotional salience and regulation networks during social exclusion. *Soc. Cogn. Affect. Neurosci.* **13**, 863–876 (2018).

30. He, Z., Ao, X., Muhlert, N., Elliott, R. & Zhang, D. Neural substrates of expectancy violation associated with social feedback in individuals with subthreshold depression. *Psychol. Med.* 1–9 (2020).

31. Mutschler, I., Ball, T., Wankerl, J. & Strigo, I. A. Pain and emotion in the insular cortex: evidence for functional reorganization in major depression. *Neurosci. Lett.* **520**, 204–209 (2012).

32. Harlé, K. M., Chang, L. J., van 't Wout, M. & Sanfey, A. G. The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *Neuroimage* **61**, 32–40 (2012).

33. Heimberg, R. G., Brozovich, F. A. & Rapee, R. M. Chapter 24 - A Cognitive-Behavioral Model of Social Anxiety Disorder. in *Social Anxiety (Third Edition)* (eds. Hofmann, S. G. & DiBartolo, P. M.) 705–728 (Academic Press, San Diego, 2014).

34. Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D. & Arch, J. J. Brain mediators of biased social learning of self-perception in social anxiety disorder. *Transl. Psychiatry* **13**, 292 (2023).

35. Czekalla, N. *et al.* Self-beneficial belief updating as a coping mechanism for stress-induced negative affect. *Scientific Reports* **11**, 1–13 (2021).

36. Schmitt, M. *et al.* Messgüte des vereinfachten Beck-Depressions-Inventars (BDI-V) [Measurement quality of the simplified Beck Depression Inventory (BDI-V)]. *Diagnostica* **49**, 147–156 (2003).

37. Pössel, P., Seemann, S. & Hautzinger, M. Evaluation eines deutschsprachigen Instrumentes zur Erfassung positiver und negativer automatischer Gedanken [Evaluation of a German-language instrument for recording positive and negative automatic thoughts]. *Zeitschrift für Klinische Psychologie und Psychotherapie* **34**, 27–34 (2005).

38. Stangier, U., Heidenreich, T., Berardi, A., Golbs, U. & Hoyer, J. Die Erfassung sozialer Phobie durch die Social Interaction Anxiety Scale (SIAS) und die Social Phobia Scale (SPS) [The assessment of social phobia using the Social interaction anxiety scale (SIAS) and the social phobia scale (SPS)]. *Zeitschrift für Klinische Psychologie und Psychotherapie* **28**, 28–36 (1999).

39. Marsh, H. W. & O'neill, R. Self Description Questionnaire iii: The construct validity of multidimensional self-concept ratings by late adolescents. *J. Educ. Meas.* **21**, 153–174 (1984).

682 40. Hoven, M., Luigjes, J., Denys, D., Rouault, M. & van Holst, R. J. How do confidence and
683 self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental*
684 *Health* **1**, 337–345 (2023).
685 41. Will, G.-J., Rutledge, R. B., Moutoussis, M. & Dolan, R. J. Neural and computational
686 processes underlying dynamic changes in self-esteem. *Elife* **6**, 1–21 (2017).
687 42. Seow, T. X. F., Rouault, M., Gillan, C. M. & Fleming, S. M. How local and global
688 metacognition shape mental health. *Biol. Psychiatry* **90**, 436–446 (2021).
689 43. Satti, M. H. *et al.* Absence of systematic effects of internalizing psychopathology on
690 learning under uncertainty. *bioRxiv* 2025.05.12.653409 (2025)
691 doi:10.1101/2025.05.12.653409.
692 44. Taylor, S. E., Neter, E. & Wayment, H. A. Self-evaluation processes. *Pers. Soc. Psychol. Bull.*
693 **21**, 1278–1287 (1995).
694 45. Elliot, A. J. Approach and avoidance motivation and achievement goals. *Educ. Psychol.*
695 **34**, 169–189 (1999).
696 46. Shechner, T. & Bar-Haim, Y. Threat Monitoring and Attention-Bias Modification in
697 Anxiety and Stress-Related Disorders. *Curr. Dir. Psychol. Sci.* **25**, 431–437 (2016).
698 47. Joormann, J. & Quinn, M. E. Cognitive processes and emotion regulation in depression.
699 *Depress. Anxiety* **31**, 308–315 (2014).
700 48. Kube, T., Rief, W., Gollwitzer, M., Gärtner, T. & Glombiewski, J. A. Why dysfunctional
701 expectations in depression persist - Results from two experimental studies
702 investigating cognitive immunization. *Psychol. Med.* **49**, 1532–1544 (2019).
703 49. Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation
704 biases in reinforcement learning. *Trends Cogn. Sci.* **26**, 607–621 (2022).

50. Engelmann, J. B., Berns, G. S. & Dunlop, B. W. Hyper-responsivity to losses in the anterior insula during economic choice scales with depression severity. *Psychol. Med.* **47**, 2879–2891 (2017).

51. Heitmann, C. Y. *et al.* Neural correlates of anticipation and processing of performance feedback in social anxiety. *Hum. Brain Mapp.* **35**, 6023–6031 (2014).

52. Bishop, A., Younan, R., Low, J. & Pilkington, P. D. Early maladaptive schemas and depression in adulthood: A systematic review and meta-analysis. *Clin. Psychol. Psychother.* **29**, 111–130 (2022).

53. Russo, S. J. & Nestler, E. J. The brain reward circuitry in mood disorders. *Nat. Rev. Neurosci.* **14**, 609–625 (2013).

54. Hayes, S., Strosahl, K. & Wilson, K. *Acceptance and Commitment Therapy*. vol. 6 (New York: Guilford press, New York, 1999).

55. Wells, A. *Metacognitive Therapy for Anxiety and Depression*. (Guilford Press, 2011).

56. Segal, Z., Williams, M. & Teasdale, J. *Mindfulness-Based Cognitive Therapy for Depression, Second Edition*. (Guilford Publications, 2018).

57. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. in *Classical conditioning II: current research and theory* (eds. Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, New York, 1972).

58. Kube, T., Kirchner, L., Lemmer, G. & Glombiewski, J. A. How the Discrepancy Between Prior Expectations and New Information Influences Expectation Updating in Depression—The Greater, the Better? *Clin. Psychol. Sci.* **10**, 430–449 (2022).

59. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.* **7**, 457–511 (1992).

60. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).

61. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies - Revisited. *Neuroimage* **84**, 971–985 (2014).

62. Kelly, C. *et al.* A convergent functional architecture of the insula emerges across imaging modalities. *Neuroimage* **61**, 1129–1142 (2012).

63. Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289 (2002).