Neurocomputational 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 Psychotherapy, RPTU Kaiserslautern-Landau, 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 are perpetuated by negatively biased feedback processing. Understanding the neurocomputational mechanisms of biased belief updating may help to counteract maladaptive beliefs. The present study uses functional neuroimaging to examine neural activity associated with prediction error-based learning in persons with major depression and healthy controls. We hypothesized that increased symptom burden is associated with negatively biased self-belief formation and altered neural tracking of social feedback. Results showed that a higher symptom burden was associated with forming more negative self-beliefs and more positive beliefs about others. This bias was driven by reduced learning from positive prediction errors in depression. Neural reactivity of the insula showed increased tracking of more negative self-related prediction errors. The interplay of increased neural responsiveness to negative feedback and reduced learning from positive feedback may contribute to 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[3]. Self-beliefs and beliefs about the world are a product of our learning history and are constantly updated in the face of incoming information that deviates from expectations[4], also called prediction errors[5,6]. In the case of depression, maladaptive self-beliefs often revolve around low self-efficacy, e.g., "I can't make it anyway"[7,8]. When the emphasis of this internal model is strong[9], one becomes insensitive to the context and may neglect contradictory positive feedback[10,11]. To understand the mechanisms of this maladaptive self-belief formation, we assessed the processes underlying ongoing self-belief formation in a computational modeling approach[12,13] in individuals diagnosed with depression and healthy control participants in a functional magnetic resonance imaging (fMRI) study.

Negative beliefs about oneself, the world, and the future are typical characteristics of depression, which are maintained by negatively distorted information processing[3,14,15]. Accordingly, a more negatively biased updating of beliefs about one's future, abilities, or social popularity has been observed in experimental settings in depression[16,17] or in association with depression-related symptoms like low self-esteem[12,18]. More recently, mental disorders have been conceptualized as an interplay of maladaptive prior internal models with a maladaptive response to prediction errors[9,19,20], which is the mismatch between prediction and actual outcome. The persistence of negative beliefs, in particular in depression, has been related to reduced updating after unexpected positive information[10]. This has been shown, for example, after receiving self-related positive performance feedback[21] or unexpectedly positive social reactions in imagined situations[22]. A potential cognitive mechanism that might hinder the change of negative beliefs in the face of positively disconfirming information is "cognitive immunization[21]." This reflects the reappraisal of disconfirming feedback, such that feedback is devalued and people's initial expectations are maintained. For example, people with depression could discount positive feedback by viewing it as an exception (e.g., "I was just lucky") or doubting its credibility (e.g., "She was only nice to me because she wants my help")[21,23]. In cognitive behavioral therapy, this process is referred to as cognitive distortion, "discounting the positive[24]."

The experience of prediction errors is also an affective phenomenon[25,26], which is of particular importance in the context of belief updating in depression. Negative affect is one of the central depressive symptoms[27], and the individual's current affective state shapes expectations[28], which thereby contribute to the formation of self-beliefs[29]. This, in turn, shapes how the individual feels in a particular situation, which results in a recursive influence of affect and belief updating on each other[29,30]. In this line, affective experiences during learning have been linked to biases in forming self-beliefs and neural processes[13]. In this sense, individuals show less flexibility in adjusting their beliefs in a positive direction when in a low mood[31,32]. Accordingly, the effectiveness of psychological interventions can increase

if, in addition to challenging maladaptive beliefs on a cognitive level, emotional meanings and in-session emotional experiences are also addressed[33]. This illustrates the importance of the interplay of affect and cognitive strategies for changing maladaptive self-beliefs.

The entanglement of forming self-beliefs and affective experiences also manifests on the neural systems level[13]. Several studies in healthy samples could show that valence-dependent tracking of prediction errors in the ventral striatum (VS) was more robust for self-related than other-related information[13] and that the neural prediction error signal is associated with biases in updating self-beliefs[34,35]. Dopaminergic regions such as the VS and also midbrain nuclei in the ventral tegmental area and substantia nigra (VTA/SN) are well described for coding prediction error signals[36,37]. However, regions linked to affective experience and motivational processes are also involved in prediction error processing. The insula, involved in emotional processing and body sensations as well as attention and action monitoring[38–40], shows modulated neural prediction error signaling, especially if they convey negative information[41–45]. Also, the amygdala, which is linked to emotional learning[46,47], is involved in tracking prediction errors[43,48,49]. Accordingly, the neural prediction error signals for self-related information in the anterior insula, the amygdala, and in the VTA/SN were associated not only with biased processes of self-belief formation but also with current affective states, which led to the idea of "affected beliefs[13]."

In the context of depression, most studies focused on the neural processing of reward prediction errors and behavioral adaption following rewarding information unrelated to a self-belief (i.e., reward learning[50]). In line with the persistence of negative beliefs against conflicting positive information, many studies reported reduced reward learning in depression[42,51–53]. However, other studies found no support for different reward learning[41,54–56] or prediction error-dependent fluctuation in state happiness between individuals with and without depression, indicating intact reward processing[57]. At the neural systems level, studies found reduced reward-related prediction error signaling in the ventral striatum[42,48,52,55] and other brain regions, such as the ventral tegmental area[42], anterior cingulate cortex, and hippocampus[48,55,58]. However, effects were also mixed, and other studies reported unaltered reward prediction error signaling in depression[41,57]. In response to negative prediction errors, some studies reported increased activity[59], while others did not[41,42].

Concerning more specific updating of self-related beliefs regarding one's future, depression was associated with greater tracking of negative prediction errors in the right inferior parietal lobule and inferior frontal gyrus[17]. In individuals with social anxiety, which is also characterized by negative self-beliefs[60] and often comorbid with depression[61,62], activity in the insula mediated the effect of negative social feedback on self-belief updates[63]. Hypersensitivity to negative information in the insula has also been reported for depression[64] and has been linked to the way emotions are processed in this condition[65,66]. Moreover, the induction of negative affect increased reactivity to negative information in the insula[67].

While prediction errors are relevant for changing beliefs and learning from prediction errors has been widely studied in depression, the link to the formation of self-beliefs is still missing. The current study builds on previous work by applying a well-established computational approach of trial-by-trial self-belief formation, the Learning Of Own Performance (LOOP) task[12,13,40,68], to a clinical sample with depression. To understand the mechanisms of how people arrive at their maladaptive self-beliefs, we aimed to test whether 1) depression is related to biased updating of self-beliefs, as opposed to updating beliefs about another person, and 2) whether this is underpinned by altered neural prediction error processing in individuals diagnosed with depression. To test these hypotheses, we followed a two-step procedure for our analyses. After a group comparison approach with participants diagnosed with depression vs. healthy controls, we followed a transdiagnostic dimensional approach following the recommendations of the Research Domain Criteria guidelines[56,69]. In line with negative self-beliefs and distorted information processing in depression, we expected more negatively biased processes of belief formation, specifically if the information is related to the self. As we previously showed associations between prediction error processing and biased self-belief formation and affect in the insula, VTA/SN, and amygdala using the same paradigm [13], we expected depression-related alterations in these regions. In line with other studies on reward learning and depression, we included the ventral striatum as a region of interest (ROI). The results show that a higher symptom burden is associated with forming more negative self-beliefs and more positive beliefs about others. Neural activity of the insula showed increased tracking of more negative self-related prediction errors but no difference in tracking positive prediction errors.

Results

Measuring neural processes of belief formation.

Participants diagnosed with depression (*n*=35) and healthy control participants (*n*=32) completed the Learning Of Own Performance (LOOP) task[12,13] in the MRI (for sample characteristics, see Supplementary Table 1). In the LOOP task, participants were asked to estimate specific attributes of objects (e.g., the height of buildings or the weight of animals). By incorporating manipulated performance feedback, participants were led to form novel beliefs about their own (Self) or another person's (Other) estimation abilities. The other person allegedly performed the task simultaneously in an adjacent room outside the MRI. At the beginning of each trial, a screen displayed the estimation category of the upcoming trial. It indicated whether participants had to perform a Self-trial or an Other-trial, followed by a rating of expected performance. On a trial-by-trial basis, participants received manipulated performance feedback that was more positive in one condition and negative in the other, allowing for an assessment of learning biases during belief formation (Fig. 1A). This resulted in four feedback conditions: Agent (Self vs. Other) x Ability (High Ability vs. Low Ability, Fig. 1B).

As in previous studies, participants adjusted their expected performance ratings according to the feedback over time, i.e., they formed novel beliefs about their own and another person's ability (Fig. 1C, Supplementary Note 1, Supplementary Table 2). To describe the belief formation process, we modeled the changes in participants' expected performance through updates from prediction errors[70]. The central variable of interest here is the learning rate, i.e., the extent to which prediction errors are weighted to make an update. Consistent with our previous studies, the winning model included distinct learning rates for positive (LR+) and negative (LR-) prediction errors, separately for Self and Other, allowing a valence- and agent-specific description of belief updating (factors Prediction error valence [PE Val] and Agent; for a more detailed description of this model and the whole model space, see Methods section and Supplementary Table 3. For model comparison results in the total sample and the subsamples, see Supplementary Note 2 and Supplementary Table 4; for correlations of model parameters, Supplementary Fig. 10). This model was selected for further analysis. To measure biased learning, we aggregated the learning rates into a Valence Bias Score VBS = (LR+ - LR-)/ (LR+ + LR-), separately for Self and Other, as in previous studies[12,13]. We assessed self-reported symptom burden with a general measure of depressive symptoms[71] as well as a more specific measure of the cognitive symptoms of depression[72]. To address symptom burden in a broader context, we additionally assessed symptoms of social anxiety[73] and self-esteem[74] as transdiagnostic markers, already proven to bias self-belief formation behavior[12]. Variability in the severity of self-reported symptoms was well explained by a single principle component (83% variance explained) reflecting general depressive and socially anxious psychopathology, akin to other recent advances [75–77], see Methods section and Fig. 1E).

Negativity bias in self-belief formation in individuals with and without depression.

Both groups showed a negativity bias in forming novel self-beliefs, which replicated previous studies using this task[12,13,68]. This bias means that learning rates were lower with positive than negative prediction errors. It was not present when participants formed beliefs about the other person's ability (PE_Val x Agent interaction: $t(195)=-3.29$, $p<.001$, negativity bias: post-hoc $LR[-]_{Self}$ vs $LR[+]_{Self}$; $t(66)=4.78$, $p<.001$). Between 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_Val x Agent x Group interaction: $t(195)=.18$, $p=.861$; Fig. 1C and D, Supplementary Table 5). This suggests that people diagnosed with depression did not differ from healthy people when forming novel beliefs in response to the feedback received. The negativity bias in both groups, which may be due to rather negative prior beliefs regarding their estimation abilities and negative affect elicited by the performance context of the setting[13], demonstrates the susceptibility to biases in self-related information processing.

Biased belief updating is linked to symptom burden.

When addressing biased belief formation and symptom burden from a dimensional perspective, we found that learning biases were associated with self-reported symptoms.

With higher scores in psychopathology, participants learned more negatively about themselves (VBS_{Self}) but more positively about others (VBS_{Other}, Agent x Psychopathology interaction: $t(65)=-2.03$, $p=.047$, Fig. 1F, Supplementary Table 6). This relative devaluation of one's performance in comparison to another person has been described to play an essential role in both the etiology and maintenance of depression[78–80].

In the absence of group differences during belief formation, this effect suggests that within-group variance is particularly important. While many studies of depression only examine subclinical depression in healthy samples, we were able to look at the distribution of symptom burden in both groups. Our findings suggest that the association of negatively biased processes of belief formation with increased symptom burden was mainly driven by the clinical sample (VBS_{Self}: $r_{MDD}=-.43$, $r_{CON}=.05$, difference $z=-1.99$, $p=.047$; Supplementary Note 4, Supplementary Fig. 1). This highlights the importance of including individuals with more pronounced symptoms when examining biases in belief formation in depression.

Disregard of positive information in association with symptom burden.

Within the clinical sample, symptom burden was more strongly linked to a lower learning rate for positive pediction errors than to a higher learning rate for negative prediction errors (MDD: $rLR[+]_{Self}=-.51$, $rLR[-]_{Self}=.03$; difference of absolute values: $z=2.1$, $p=.036$, Fig. 1G, Supplementary Fig. 2, for more detailed correlations with symptoms scores, see Supplementary Fig. 3). This implies that people with more severe symptoms made fewer belief updates after positive prediction errors, which may drive the association of symptom burden and biased learning.

Neural tracking of negative prediction errors in depression and association with affective experience.

We examined the neural processing of self-related prediction errors to further understand the biased formation of self-beliefs. Individuals who updated their beliefs more positively (positive learning rate; LR[+]) also had stronger activity in the VTA/SN ROI after experiencing positive prediction errors (categorical PE_Val effect, Spearman correlation between averaged parameter estimates within ROI $r=.33$, $p=.006$). Individuals who updated their beliefs more negatively (negative learning rate; LR[-]) had stronger activity in the VTA/SN, insula, and amygdala ROIs during negative prediction errors (Spearman correlations within ROIs $rs\geq.27$, $ps\leq.029$, see Supplementary Fig. 5). This demonstrates the link of neural prediction error signals and behavioral updating of self-beliefs.

When comparing individuals with depression and healthy controls, a significant Group x PE Valence interaction in the right ventral and posterior insula indicated an imbalance of negative relative to positive prediction errors signaling in the clinical sample, but not in healthy controls (continuous PE effect, $ps\leq.030$ family-wise [FWE] corrected at peak level within ROIs, Fig. 2, Supplementary Table 8, for baseline activations of continuous PE effect see Supplementary Table 7). We found no difference between study groups in the association of neural activity with more positive self-related prediction errors (with $p<0.05$

FWE-corrected at peak level for all ROIs). This suggests that during the formation of novel self-beliefs, surprisingly positive feedback is processed similarly in individuals with and without depression. However, compared to control participants, individuals with depression showed stronger activity in the right dorsal, ventral, and posterior insula as well as in the VTA/SN with more negative prediction errors (two-sample t-tests, $ps \leq .046$ within ROIs, see Supplementary Table 8). This is in line with a generally stronger tracking of negative compared to positive prediction errors in the insula (Supplementary Fig. 6) and suggests that individuals with depression are more sensitive to negative prediction errors at the neural level. In contrast, the processing of positive prediction errors is unaltered.

In addition, stronger activity in the same regions (right dorsal, ventral, posterior insula) with more negative prediction errors was accompanied by decreased happiness (Spearman correlation between averaged parameter estimates within all three ROIs $rs \geq -0.33$, $ps \leq .029$; Supplementary Fig. 7). This finding of reduced positive affect during the processing of negative prediction errors is of particular importance in the context of affective disorders, as it is included in various psychological models of depression as a sustaining element in the vicious circle[81–84]. Stronger neural responsivity to more negative prediction errors and more negative affective responses could, therefore, influence each other and maintain the generally lower positive affect in individuals with depression during task performance (Supplementary Table 9).

Neural tracking of negative prediction errors is linked to biased belief formation and symptom burden.

In line with the results from the group comparison, symptom burden was associated with stronger responses to more negative prediction errors within the right dorsal and posterior insula (Spearman correlation with Psychopathology score within ROIs $rs \geq .26$, $ps \leq .037$). Stronger responses in the right dorsal insula with more negative prediction errors were also associated with a more negatively biased self-belief formation (VBS_{Self} $r=-.24$, $p=.046$), replicating previous findings[13]. Neither levels of symptom burden nor the Valence Bias Score were significantly associated with neural activity in any of the ROIs with more positive prediction errors (Supplementary Fig. 9). These results suggest that a stronger tracking of negative prediction errors in parts of the insula accompanies more biased belief updating and higher symptom burden.

Discussion

In the present study, we used neurocomputational methods to examine the mechanisms that underlie the altered processing of self-related feedback in the formation of maladaptive self-beliefs in depression. To promote a transdiagnostic dimensional approach, we recruited a naturalistic clinical sample of individuals diagnosed with depression, including comorbidities of social anxiety, and examined the symptom burden dimensionally. Our data suggest that the combination of stronger neural reactivity of the insula during self-related

negative prediction errors, together with greater neglect of self-related positive prediction errors during the updating process, might play a key role in the maintenance of maladaptive self-beliefs in depression. Our results provide new insights into cognitive distortions and the persistence of negative self-beliefs as important characteristics of the psychopathology of depression.

Behaviorally, the results replicate previous findings of a negativity bias in the formation of novel self-beliefs[12,13,68], both for individuals with depression and healthy controls. This bias is absent when participants form beliefs about others, highlighting the specificity of the valence bias for processing self-related information[13]. While other studies report a positivity bias when updating self-beliefs about one's future[35,85,86], personality[87], intelligence, or appearance[88,89], we find a negativity bias, which aligns with other studies using task-specific performance feedback[90–92]. This suggests that contextual effects, such as whether study participants learn about an ability vs. their personality, might impact the direction of learning biases. Besides the effects of the context, prior beliefs and confidence might be linked to the formation and revision of self-beliefs.

Although the dimensional approach suggests an association between depressive symptoms and biased belief formation, the simple group comparison demonstrates that individuals with depression are not fundamentally different from control participants in terms of how they form novel self-beliefs in this performance context. While some studies report less positive self-belief updating in samples with depression[16,21], other studies, particularly those that employ reward learning tasks, also do not find group differences in reward learning rates[41,54,56]. In addition to contextual effects like task selection, the sample's composition may be an important factor related to the extent of learning bias and differences between groups. The present findings indicate that, even in a depressive state, people can acquire novel beliefs about their abilities in a way that does not categorically distinguish them from people without this diagnosis.

In the dimensional approach, the results indicated that individuals with higher depressive and social anxiety-related symptom burden exhibited more negatively biased self-belief formation. In contrast, they formed their beliefs about others in a positively biased manner. This aligns with other studies linking biased self-belief updating to symptom burden[16,22,53,64,93]. The diminished value one places on one's accomplishments relative to others, along with the affective response to this self-esteem-debilitating social comparison, has been identified as a symptom-maintaining factor in depression[79].

In the absence of group differences, this effect highlights the importance of considering within-group variance in symptom burden. Diagnostic categories summarize a heterogeneous symptom constellation and have been criticized as non-distinct and not reflecting the continuous nature from at-risk through mild to severe symptoms[94,95]. Furthermore, comorbidities are common[96]. This means that heterogeneity in symptoms must be considered in clinical samples, especially in naturalistic samples, including comorbidities, as in the current study. A dimensional approach with a naturalistic sample

considers the continuous nature of psychopathology and increases ecological validity[97]. The results indicate that the correlation between symptom burden and negatively biased self-belief formation was mainly driven by the clinical sample. Therefore, it is necessary to include individuals with clinically relevant symptom burden. Also, within the facets of psychopathology, those targeting specific symptoms such as negative thoughts or self-esteem were more likely to yield associations with biased learning, which may provide evidence for focusing on symptoms rather than categories.

In the clinical sample, the bias in belief formation is driven by reduced learning from positive self-related prediction errors. One explanation for this finding can be that reduced attention is allocated to positive feedback[98,99]. An alternative explanation is that cognitive mechanisms are actively involved in devaluing positive feedback, resulting in a cognitive immunization against it[100]. Future studies may employ experimental manipulations to modulate feedback value and attentional allocation as well as measure the evaluation of feedback and the attentional focus (potentially using eye-tracking) to further disentangle these possibilities. Addressing these aspects will provide a more nuanced understanding of the underlying cognitive processes contributing to the observed bias in self-belief formation. Overall, these results suggest that sensitivity to context decreases with increasing symptom burden, possibly due to an overemphasis on an already established negative internal model. This means that less is learned from positive prediction errors, and beliefs are biased towards prior belief models in a confirmatory way[101,102].

This finding may have important clinical implications for psychotherapy. Rather than directly challenging the negative self-concepts of patients with depression (e.g., by cognitive restructuring), they can be supported in becoming more context-sensitive so that positive feedback that contradicts the internal model can be incorporated. This might be done by practicing skills that help to become less reactive to established negative self-beliefs, e.g., defusion in Acceptance and Commitment Therapy (ACT[103]) or detached mindfulness in Metacognitive Therapy[104] and more sensitive to the present context, e.g., mindfulness skills in ACT or mindfulness-based cognitive therapy[105]. To update a maladaptive belief, it may be necessary for a therapist to thoroughly examine the meaning attributed to new experiences and identify potential cognitive distortions, such as retrospectively devaluing positive experiences. These mechanisms can be discussed with patients as a factor that maintains a negative self-image. Second, patients may be empowered that meeting the diagnostic criteria for a particular disorder does not necessarily mean that they are fundamentally different in their ability to form new self-beliefs. This can be important, for example, when introducing new antidepressant activities where new self-beliefs need to be built.

On the neural system level, we found that the insula is associated with stronger tracking of negative prediction errors in the clinical sample compared to the control group. There was no group difference for positive prediction errors. Tracking of negative prediction errors was also associated with symptom burden. The stronger neural response to negative prediction errors is in line with a general neural sensitivity to negative (social) prediction errors in the insula[13,106,107], as well as a stronger insula reactivity specifically in samples with depression[64]

or social anxiety[63,108]. This effect could be explained by an attentional bias in individuals with depression[98,109] in an affect-congruent way[28], with either a faster orientation[99] or longer maintenance of attention to negative feedback[110,111]. Since the insula reactivity to more negative prediction errors is associated with less positive affect during task performance in the present experiment, the detected group difference might also reflect stronger emotional responsivity in individuals with depression. This would be in line with the idea of affected beliefs[13] and corresponds to the generally reduced positive affect in individuals with depression during the task. Accordingly, earlier studies found stronger reactions to negative emotional stimuli in individuals with depression together with reduced cognitive emotion regulation strategies like reappraisal[112] and, therefore, include affective responses in models of depression[84].

The stronger neural and possibly affective reaction to more negative self-related feedback can also be discussed in the context of emotional schemas and their role in shaping emotional responses. Emotional schemas are formed by connecting highly arousing emotional events[6] to the resulting emotional reactions[113]. They can be triggered by learned signs associated with the events, resulting in automatic, rapid, and exceptionally intensive emotional responses[113]. Depression is more likely related to early maladaptive schemas rooted in a learning history of negative experiences, such as social devaluation in response to failure[114,115]. Therefore, individuals may have emotional schemas that are particularly sensitive to negative self-related feedback, which results in a stronger emotional response and, thus, an overall lower positive affect during task performance. This stronger reaction could be an expression of a stronger emphasis on these previously established negative internal models, making it more difficult to process external stimuli that contradict them. To better understand task-specific negative affective reactions to self-related feedback, trial-by-trial fluctuations in happiness or other task-relevant self-conscious emotions like pride[116] or embarrassment[40] should be measured in future studies.

Clinical implications of the neurocomputational findings: When patients are particularly sensitive to negative feedback, in addition to psychoeducation about emotions and emotional schemas, the finding of increased activation in the insula can be used to contextualize and validate the patient's emotional experiences. Subsequently, as mentioned above, skills to reduce emotional reactivity can be trained. Our findings of uncompromised processing of positive prediction errors in the VS or VTA/SN are in line with several other studies in depression, which show unaltered VS activity in reward learning tasks. Together with the above finding of similar learning in both groups, our results question the widespread hypothesis of altered VS reactivity and prediction error learning deficits in depression[51,117,118]. The results suggest that the neural processing of positive self-related prediction errors as a basis for forming novel positive self-beliefs is unaltered, at least at this initial stage of processing in depression.

How can we interpret the reduced belief updating in the clinical sample following positive prediction errors while the neural tracking of positive prediction errors is unaffected?

A possible explanation is timing: the neural response is measured when feedback is given. The belief update may occur at a later stage, somewhere in between feedback, feedback processing, interpretation, and indication of the new expectation. As such, belief updates measured in behavior may represent an amalgam of several trials. In this case, cognitive appraisal mechanisms might have come into play. We also assume that stronger negative affective reactions in individuals with depression further promote cognitive distortions like cognitive immunization and, thus, influence self-ratings even in positive conditions.

The focus of the learning task used is on the formation of relatively novel beliefs[119]. We selected estimation entities in which most people do not have strong prior beliefs regarding their ability. Although participants diagnosed with depression had lower overall self-esteem before the task, their prior beliefs about their estimation abilities were not significantly different. So, we can look at the underlying mechanisms of how people arrive at their beliefs and show that depressive symptomatology has an impact already at this stage. Although established, internal models can potentially influence the process of belief updating, the little prior experience that people typically have, e.g., in estimating the weight of animals, and thus the lower confidence in their self-beliefs prior to the task, is different from stable maladaptive beliefs that have been formed over time and continuously reinforced by symptom-maintaining behaviors. To experimentally address the revision of established beliefs to better map psychotherapy processes, future studies could use tasks that address more global negative self-beliefs with high confidence levels. Alternatively, the initial belief formation task could be followed by another session to challenge established beliefs again. Also, negative feedback during social interaction could be used to map the negative learning experience of many people with depression more closely, which might activate maladaptive self-beliefs more strongly. This could further improve the translational nature of future work.

In conclusion, the present study sheds light on the neurocomputational mechanisms that contribute to maladaptive self-belief formation in depressive psychopathology. The results emphasize a heightened neural response to more negative self-related feedback in the insula and a potentially more emotional response that, together with less consideration of positive feedback, promotes more biased self-belief formation in depression (Fig. 3). The study highlights the importance of accounting for within-group variance and symptom burden in a clinically relevant symptom 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, 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 goal in recruiting the clinical sample 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); 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. 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. Twenty-eight were outpatients, seven inpatients. Twenty-seven of the outpatients were in a day treatment program for depression (8 weeks, weekday mornings to afternoons) with multi-professional therapy. All patients received psychotherapy, and n = 28 received additional pharmacotherapy. Measurement day was, on average, 2.4 weeks (SD = 2.0) after admission to the clinic. When recruiting the control group, group-wise pairing of age and gender distribution was aimed at. 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 1) were recruited at the university campus of Lübeck or via public notices. The exclusion criterion was mental illness or neurological disease. In a telephone interview before the study, the subjects were screened for symptoms of depression, and in the second part, a general screening for other mental illnesses was conducted. Screenings were adapted from a structured interview[120]. Two individuals had to be subsequently excluded despite screening due to clinically relevant BDI scores (> 35). We initially recruited 85 participants in total but had to exclude 18 participants. In the clinical sample, we excluded five due to technical problems, four due to premature termination of the study, and two participants who did not attentively complete the task until the end (e.g., ratings indicated that they stopped responding). In the control group, we excluded three participants who did not believe the cover story of the task and/ or did not attentively complete the task; one was excluded due to technical problems, one due to a premature termination of the study, and two due to high BDI scores as mentioned above.

Learning Of Own Performance task.

The Learning Of Own Performance (LOOP) task allows participants to gradually learn about their own or another person's purported ability to estimate various properties through performance feedback after each trial. The task was previously introduced and validated in behavioral and fMRI studies on healthy participants[12,13,68]. Two participants were invited to

participate in the study covered as an experiment on cognitive estimation; one performed the task in the scanner, and the other as a behavioral study (data not included). When a second participant was unavailable, a confederate was presented and allegedly placed in the adjacent room to participate in the behavioral study. The participants were informed that they would take turns with the other individual, either actively performing the task themselves (Self condition) or observing the other person's performance (Other condition). The task involved estimating different attributes in four estimation categories: house height, animal weight, vehicle distance, and food quantity. After each estimation trial, participants received manipulated performance feedback. In two estimation categories, the feedback was related to one's own performance and, in the other two, to the other person's estimation performance. One of the two estimation categories for each agent was associated with predominantly positive feedback (High Ability condition), while the other was linked to mostly negative feedback (Low Ability condition). Estimation categories were counterbalanced between Ability conditions and Agent conditions. This resulted in four feedback conditions (Agent [Self vs. Other] x Ability [High vs. Low]) with 20 trials each. The trials of all conditions were intermixed in a fixed order with a maximum of two consecutive trials of the same condition. Performance feedback presented after every estimation trial indicated the participant's or the other person's current estimation accuracy as percentiles compared to an alleged reference group of 350 former participants (e.g., "You are better than 94% of the reference participants"; see Fig. 1). The feedback was defined by a sequence of predetermined prediction errors concerning the participants' current beliefs about their abilities. The current belief was calculated as the average of the last five performance expectation ratings per category, which started at 50% before participants rated their performance expectations. This procedure led to varying feedback sequences between participants while ensuring that the prediction errors remained mostly independent of the participants' performance expectations and that negative and positive prediction errors were relatively equally distributed across groups and conditions (MDD Self: mean positive PE = 14.2, SD = 1.5 (mean frequency = 20.4); mean negative PE = −12.8, SD = 1.8 (mean frequency = 19.1); MDD Other: mean positive PE = 13.6, SD = 1.3 (mean frequency = 19.0); mean negative PE = −14.0, SD = 1.8 (mean frequency = 20.3), CON Self: mean positive PE = 13.9, SD = 1.4 (mean frequency = 20.2); mean negative PE = −13.2, SD = 1.4 (mean frequency = 19.3); CON Other: mean positive PE = 13.8, SD = 1.7 (mean frequency = 18.7); mean negative PE = −13.5, SD = 1.2 (mean frequency = 20.7), no group differences in mean PE for all four PE conditions, all $p > 0.2$). At the beginning of each trial, a cue was presented indicating the estimation category (e.g., height) and the agent assigned for that trial (e.g., You). Participants were then asked to provide their expected performance for that specific trial on a scale with the same percentiles used for feedback. To enhance motivation and encourage honest responses, participants were informed, as part of the cover story, that accurate expected performance ratings would be rewarded with up to 6 cents per trial, that is, the closer their expected performance rating matched their actual feedback percentile, the more money they would receive. After each performance expectation rating, the estimation question was presented for 10 seconds. During the estimation phase, continuous response scales under the pictures defined a range of plausible answers for each question.

Participants indicated their responses by moving a pointer on the response scale using an MRI-compatible computer mouse. Subsequently, feedback was presented for 3 seconds (Fig. 1A). Jittered inter-stimulus-intervals were presented following the cue (mean: 4*TR (0.992 s), range: 2–6*TR), estimation (mean: 4.5*TR, range: 2.5–6.5*TR), and feedback phase (mean: 6*TR, range: 4–8*TR) for the fMRI task with jitters distributed in a uniform distribution with steps of 0.5*TR. During the task, participants rated their current levels of embarrassment, pride, happiness, and stress/ arousal on a continuous scale ranging from not at all (coded as 0) to very strong (coded as 100). Emotion ratings were presented four times, each following a trial of one of the four experimental conditions (Self/Other, High/Low). The two happiness ratings following Self-trials were averaged to receive a measure of a general affective experience following self-related feedback. All stimuli were presented using MATLAB (Release 2015b, The MathWorks, Inc.) and the Psychophysics Toolbox (Brainard, D. H. [1997]. The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436). The fMRI task was completed in two separate approximately 25-minute sessions with a short break in between.

Before starting the experiment, demographic data and a self-esteem personality questionnaire (Self-Description Questionnaire-III, SDQ-III subscale scores[74]) were retrieved. Following the task, participants completed questionnaires addressing the symptom burden, including Beck's depression inventory (BDI-V[71]), Automatic Thought Questionnaire (ATQ[72]), and the Social Interaction Anxiety Scale (SIAS[73]).

Statistical analysis.

Behavioral data analysis and modeling.

First, we checked for group differences in prior beliefs (Supplementary Note 5, Supplementary Table 10) and other setting characteristics at the start (Supplementary Table 11). Then, we ran a model-free analysis on the participants' expected performance ratings for each trial. We employed a linear mixed model with a maximum likelihood estimation. The model included the factors Ability (High vs. Low) x Agent (Self vs. Other) x Group (MDD vs. CON) and Trial (20 trials) as continuous predictors. Intercept, Ability, Agent, Trial, and Group were modeled as fixed effects; the intercept was additionally modeled as a random effect (Supplementary Note 1, Supplementary Table 2).

After the model-free analyses, dynamic changes in performance expectation ratings were modeled using prediction error delta-rule update equations (adapted Rescorla–Wagner model[121]). The model space and the procedure of model fitting and selection have been implemented and further developed in our previous studies[12,13]. For the learning models the following equation was used (EXP = Performance expectation rating, FB = feedback, PE = prediction error, α = learning rate):

$$EXP_{t+1} = EXP_t + \alpha\, PE_t; \text{ while } PE_t = FB_t - EXP_t$$

The model space consisted of three main models, each with different assumptions regarding biased updating behavior when forming novel beliefs (for model space, see Supplementary Table 3). The simplest learning model (Unity Model) employed a single learning rate for all conditions for each participant, thus assuming no learning biases. The Valence Model included separate learning rates for positive (LR+) and negative prediction errors (LR-) across both ability conditions, suggesting that the valence (positive vs. negative) of prediction errors influences belief formation. The Ability Model incorporated distinct learning rates for each ability condition, indicating context-specific learning. In addition, learning rates were either estimated separately for Self vs. Other (Models 4, 5, 6, and 7) or across Agent conditions (Models 1, 2, and 3). The Valence Model with separate learning rates for Self vs. Other (Model 5), which was the winning model in our first studies[12], was further extended by adding a weighting factor. This factor reduced the learning rates when feedback values approached the extremes of the scale (percentiles close to 0% or 100%, Model 7), assuming that participants would perceive extreme feedback to be less likely than average feedback[122]. Since many variables encountered in everyday life approximately follow a normal distribution where extreme values are less probable, we assigned the relative probability density of the normal distribution to each feedback percentile value. A weighting factor *w* was fitted for each individual, indicating how strongly the relative probability density reduced the learning rates for feedback further away from the mean. The Weighted Valence Model was the winning in our last study[13]. The normal decay (ND) weighted by the weighting factor *w* was introduced in the learning models in the following way:

$$EXP_{t+1} = EXP_t + \alpha\, PE_t\, (1 - w\, ND)$$

The initial beliefs about the own and the other participant's performance (EXP_1) were estimated as free parameters separately for Self and Other in both Ability conditions, resulting in four additional model parameters. To compare whether the participants' performance expectation ratings can be better explained in terms of prediction error learning compared to assuming stable values within each ability condition, we included a simple Mean Model with a mean value for each task condition (Model 8).

Model fitting.

For model fitting, we used the R package *'RStan'* (Stan Development Team [2022]. RStan: the R interface to Stan. R package version 2.21.7. https://mc-stan.org/). It employs Markov chain Monte Carlo (MCMC) sampling algorithms. Each participant's learning parameters for all models of the model space were individually fitted. A total of 2400 samples were drawn after 1000 burn-in samples (overall 3400 samples thinned with a factor of 3) in three MCMC chains. Posterior parameter distributions were sampled for each participant. We assessed whether MCMC chains converged by inspecting $\hat{R}$ values for all model parameters[123]. Effective sample sizes n_{eff} of model parameters (effective number of independent draws from the posterior distribution) were typically greater than 1500 (for most parameters and subjects). To obtain a single parameter value per participant for the group statistics,

posterior distributions for all parameters for each participant were summarized by their mean.

Model selection.

To select the model that best described the course of our performance expectation ratings, pointwise out-of-sample prediction accuracy was estimated by approximating leave-one-out cross-validation (LOO[124]). LOOs were calculated for all fitted models separately for every participant using Pareto smoothed importance sampling (PSIS) with the log-likelihood from the posterior simulations of the parameter values[125]. We calculated $\hat{k}$ values, the estimated shape parameters of the generalized Pareto distribution, a measure of the reliability of the PSIS-LOO (see Supplementary Table 4 for sum LOO Scores, $\hat{k}$ values, and difference scores of PSIS-LOO in contrast to the winning model). Only a few trials resulted in insufficient parameter values for $\hat{k}$ and thus potentially unreliable PSIS-LOO scores. Bayesian Model Selection on PSIS-LOO scores was performed on the group level for the whole sample and both sub-samples[126] using MATLAB (Release 2019b, The MathWorks, Inc.). It provides the protected exceedance probability for each model (*pxp*), indicating the likelihood of a given model explaining the data better than all other models in the comparison set. The Bayesian omnibus risk (*BOR*) quantifies the posterior probability that model frequencies for all models are equal. To assess whether the winning model captured the effects in the behavioral data, we repeated the model-free analysis (Supplementary Note 1) with the data predicted by the winning model; here, the Weighted Valence Model (Supplementary Note 3).

Capturing symptom burden.

For a dimensional perspective on depression, we assessed the severity of depressive symptoms with the BDI-V as a general measure of depressive symptoms, as well as the ATQ as a more specific measure of the cognitive symptoms of depression. To address symptom burden in a broader context, we additionally assessed symptoms of social anxiety with the SIAS and self-esteem with SDQ-III subscale scores as a transdiagnostic measure since we could show a relationship with belief updating behavior and these two concepts in a previous study[12]. To reduce dimensionality, we ran a principal component analysis on our questionnaire data using the *R* package '*psych*' (varimax rotation, component scores based upon the structure matrix [default], William Revelle (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.3.3, https://CRAN.R-project.org/package=psych). We extracted one component as a combined measure of depressive and socially anxious psychopathology (Fig. 1E).

Statistical analyses of learning parameters.

To test whether the updating of self-beliefs was different between groups, learning rates of the winning models for positive (LR+) and negative prediction errors (LR-, factor Prediction error valence, PE_Val) and for Self and Other (factor Agent) were compared between the two groups in a PE_Val x Agent x Group linear mixed model. To associate learning biases with

symptom burden, we computed a Valence Bias Score for Self and Other that captured a bias in belief updating as a difference between the positive and negative learning rate, VBS=(LR[+] – LR[-])/ (LR[+] + LR[-])[12,13,127], and calculated an Agent x Psychopathology score linear mixed-effects model. For a more detailed understanding of the relationship between biased belief updating and symptom burden, we additionally performed Spearman correlations between the Psychopathology score and the Valence Bias Score as well as the two separate self-related learning rates LR[+] and LR[-] within the two sub-samples. We compared the correlation of the Psychopathology score and the Valence Bias Score between the MDD and CON groups to see which group drives the effect. Since we had a particular interest in the relationship between the Psychopathology score and the self-related positive learning rate as a potential measure of cognitive immunization against positive feedback, we additionally tested the absolute correlation parameter of LR[+] and Psychopathology against the one with LR[-] within the clinical sample (*r* package *'cocor'* that implemented to correlation comparison approach by Pearson and Filon, 1898; Diedenhofen, B. & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE, 10*(4): e0121945). Statistical tests were performed two-sided. Statistical analyses on the behavioral data were performed using *R* (R Core Team [2022]. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/).

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. In each of the two experimental sessions, 1516 functional volumes (min=1400, max=1724) were acquired on average using echo planar imaging (EPI, TR=0.992 s, TE=28 ms, flip angle=60°, voxel size= $3\times3\times3mm^3$, simultaneous multi-slice factor 4). A high-resolution anatomical T1 image was obtained for normalization purposes (voxel size $=1\times1\times1mm^3$, $192 \times 320 \times 320$ 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 analyses.

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 $2\times2\times2$ 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 vs. Other × PE_Val: positive [+] vs. negative [-]), weighted with PE magnitude per trials (continuous effect of the unsigned PE values), one regressor for the performance expectation rating phase, two for the estimation period for Self and Other, and one for the emotion rating phase. Parametric modulators were not orthogonalized; thus, each only explained their specific variance. Regressors were modeled with the duration as presented during the experiment (cue phase: 2.5s, performance expectation rating: individual reaction times with *M*=4.0s, *SD*=2,4, estimation phase: 10s, feedback phase: 3s, emotion rating phase: *M*=26.2s, *SD*=10.1). Six additional regressors were included to correct for head movement, and one regressor was included with a constant term for each of the two sessions.

On the second level, we first compared the brain activity for self- vs. other-related positive and negative prediction errors in the feedback phase in a flexible factorial design (regressor for the four conditions Agent: Self x Other, PE_Val: PE[+] x PE[-], Supplementary Fig. 4) and checked whether the brain activity in the positive self-related PE condition (categorical effect) correlates with $LR[+]_{Self}$ and for the negative PE condition with $LR[-]_{Self}$ within our predefined ROIs. For this, parameter estimates of all voxels within one ROI were extracted, averaged, and correlated.

Second, the tracking of positive and negative self-related prediction errors captured by the parametric modulator of PE magnitude (continuous effect) was compared between the MDD and CON group with a two-sample t-test within our ROIs (for the distribution of parameter estimates; see Supplementary Fig. 8). To test whether the difference between the tracking of positive and negative prediction errors is stronger in the clinical compared to the control sample, we additionally checked the PE Valence x Group interaction in a flexible factorial design within our ROIs. To assess correlations with the Psychopathology score, the Valence Bias Score, and the happiness ratings, parameter estimates of all voxels within one ROI were extracted and averaged as mentioned above.

The selection of regions of interest was guided by our previous fMRI study with a healthy sample using the same paradigm[13]. As the insula, with its suggested role as an integrative hub for motivated cognition and emotional behavior, has emerged as particularly relevant in this study, it was again defined as an ROI. We used unilateral insula ROIs with their dorsal, ventral, and posterior part as described in the three-cluster solution of Kelly and colleagues[128]. Other ROIs that were motivated by this study are in the amygdala (two unilateral ROIs from the AAL atlas definition in the WFU PickAtlas[129]) and an anatomically defined VTA/SN ROI, dopaminergic nuclei in the midbrain (probabilistic atlases of the midbrain; Adcock Lab[130,131]). Since altered prediction error processing in the ventral striatum has been discussed a lot in the context of depression, a functional ROI within the ventral striatum (according to the SPM anatomy toolbox) capturing the processing of prediction error valence was additionally defined as ROI. It derived from a family-wise error $p<.05$ corrected baseline effect of the continuous signed self-related prediction errors of our previous study[13].

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

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.

Figures and Legends

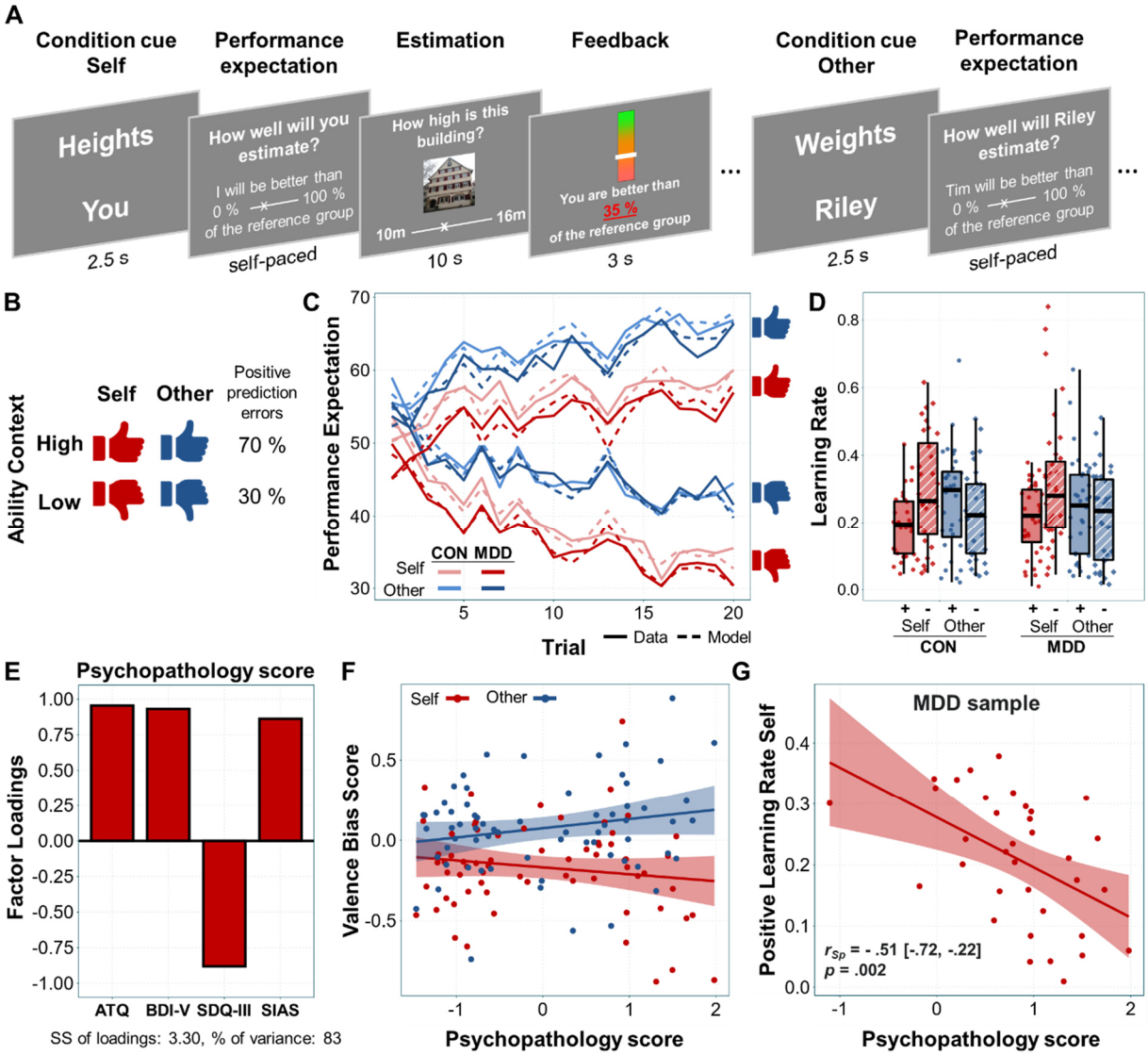

Figure 1. Trial sequence of the Learning Of Own Performance (LOOP) task, experimental conditions, and behavioral results. **A** Sequence of One Trial. 1.) Condition Cue: estimation category (e.g., weights) paired with either high or low ability feedback and agent, 2.) Rating of expected performance, 3.) Estimation question, 4.) Feedback. **B** Low and High Ability conditions of alleged estimation performance for Self and Other (Agent condition) were presented in pseudo-randomized order. Feedback with controlled prediction errors: High Ability: 70 %, Low Ability: 30 % of the trials with planned positive prediction errors (the actual percentage of positive/ negative prediction errors could slightly differ, e.g., if the feedback had been out of range). **C** 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 Prediction error valence captured the participants' behavior, as indicated by a close match of actual ratings and predicted values.

D Learning rates of the winning model show a bias towards increased updating in response to negative prediction errors (-) in contrast to positive prediction errors (+), specifically for self-beliefs in both groups. Colored bars indicate the first and third quartile 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). **E** 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[72], Beck's depression inventory, BDI-V[71], Self-Description Questionnaire-III, SDQ-III subscale scores[74], Social Interaction Anxiety scale, SIAS[73]). SS = sum of squares. **F** Valence Bias Score (LR[+] – LR[-])/ (LR[+] + LR[-]) as a function of the Psychopathology score and Agent. A significant Agent x Psychopathology interaction shows more negatively biased updating for Self and more positively biased updating for Other with more symptom burden. Lines for displaying purposes. **G** Correlation of the positive learning rate (LR[+]) and the Psychopathology score in the clinical sample. Spearman correlation coefficient, line for displaying purposes.

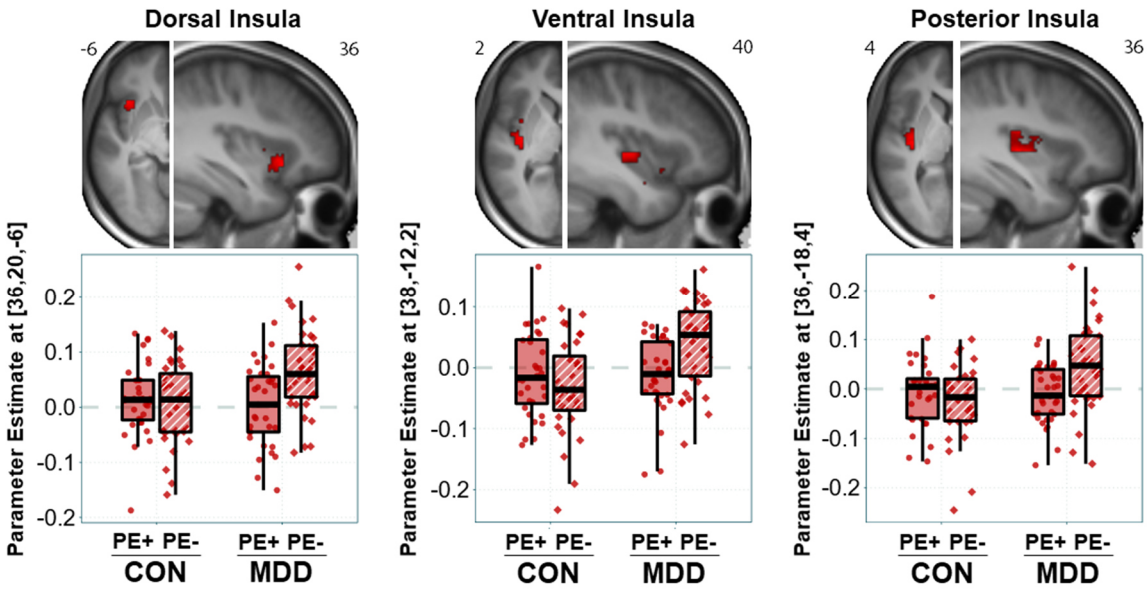

Figure 2. Differential insula tracking of prediction error valence in depression. The interaction (MDD[PE-]>MDD [PE+])>(CON [PE-]>CON [PE-]) shows a greater difference between PE- vs. PE+ in the MDD compared to the CON group (dorsal: t=2.90, p=.070, ventral: t=3.35, p=.029, posterior: t=3.17, p=.030, FWE corrected within ROIs at peak level). 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 three regions of interest in the right insula. Brain plot: uncorrected p<0.05 within ROI for display purposes; see Supplementary Table 8 for FWE corrected statistics.

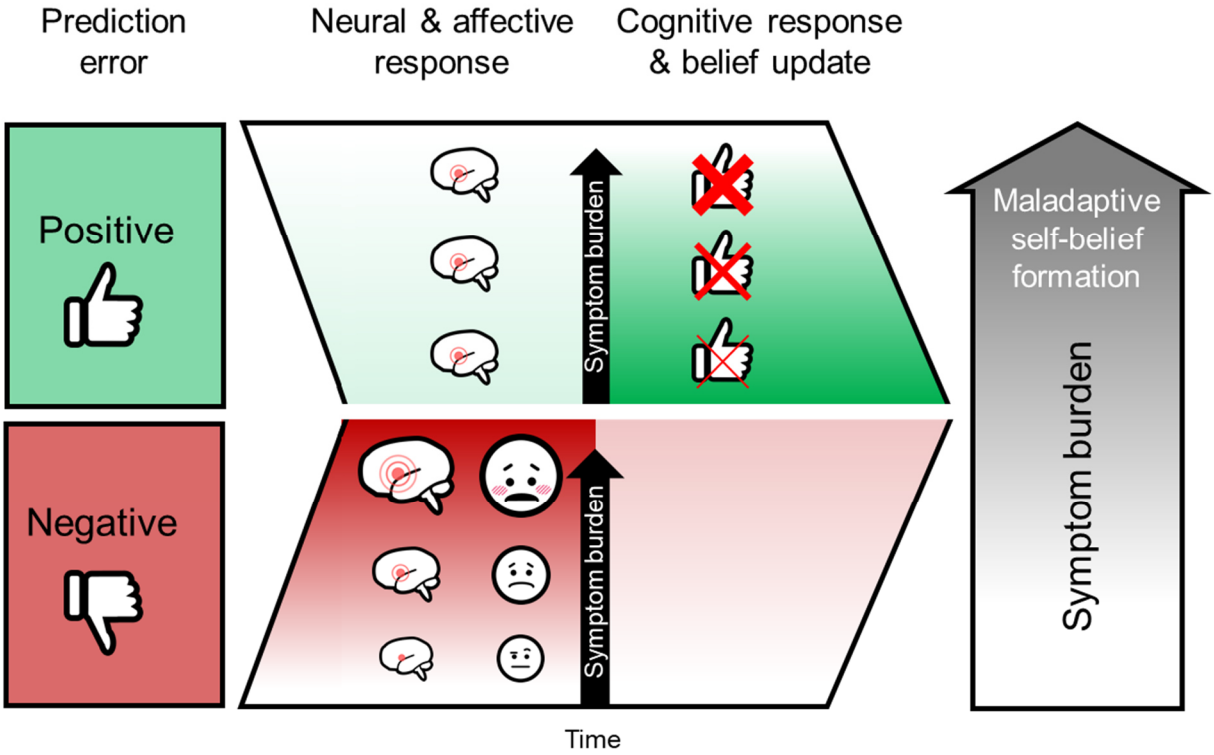

Figure 3. Schematic model of belief formation in relation to increasing depressive social anxiety symptom burden. Prediction error: Trial-by-trial positive and negative prediction errors were presented as social performance feedback in a belief updating task. Symptom burden: combined measure of Beck's depression inventory[71] as a general measure of depressive symptoms, Automatic Thought Questionnaire[72] as a measure of cognitive symptoms of depression, as well as self-esteem[74], and social anxiety[73] as transdiagnostic measures. Neural and affective response to feedback: Stronger activity within the insula with more negative prediction errors is associated with higher symptom burden. Stronger insula activity is also linked to less positive affect. Cognitive response and belief update: Fewer belief updates following positive prediction errors in the clinical sample might reflect the bias of cognitive immunization against unexpected positive feedback. Time: suggested interpretation of an initial neural and affective response during feedback presentation and possible downstream cognitive mechanisms with subsequent belief updating. All of this together might contribute to more biased beliefs with higher symptom burden.

References

1. Beck, A. T. *Cognitive Therapy and the Emotional Disorders*. (International Universities Press, 1976).
2. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**, 191–215 (1977).
3. Beck, A. T. *Cognitive Therapy of Depression*. (Guilford Press, 1979).
4. Markus, H. R. & Wurf, E. The dynamic self-concept: a social psychological perspective. *Annu. Rev. Psychol.* **38**, 299–337 (1987).
5. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 815–836 (2005).
6. Rouhani, N., Niv, Y., Frank, M. J. & Schwabe, L. Multiple routes to enhanced memory for emotionally relevant events. *Trends Cogn. Sci.* **27**, 867–882 (2023).
7. Bandura, A., Pastorelli, C., Barbaranelli, C. & Caprara, G. V. Self-efficacy pathways to childhood depression. *J. Pers. Soc. Psychol.* **76**, 258–269 (1999).
8. Hollon, S. D. & Kendall, P. C. Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. *Cognit. Ther. Res.* **4**, 383–395 (1980).
9. Clark, J. E., Watson, S. & Friston, K. J. What is mood? A computational perspective. *Psychol. Med.* **48**, 2277–2284 (2018).
10. 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).
11. Villano, W. J. & Heller, A. S. Depression is associated with blunted affective responses to naturalistic reward prediction errors. *Psychol. Med.* 1–9 (2024).

12. Müller-Pinzler, L. *et al.* Negativity-bias in forming beliefs about own abilities. *Sci. Rep.* **9**, 14416 (2019).

13. Müller-Pinzler, L. *et al.* Neurocomputational mechanisms of affected beliefs. *Commun Biol* **5**, 1241 (2022).

14. Beck, A. T. Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of general psychiatry* **9**, 324–333 (1963).

15. Beck, A. T. Thinking and depression: II. Theory and therapy. *Archives of general psychiatry* **10**, 561–571 (1964).

16. 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).

17. Garrett, N. *et al.* Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Frontiers in human neuroscience* **8**, 639 (2014).

18. Will, G.-J. *et al.* Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Transl. Psychiatry* **10**, 96 (2020).

19. Sterzer, P. *et al.* The predictive coding account of psychosis. *Biol. Psychiatry* **84**, 634–643 (2018).

20. Stephan, K. E. *et al.* Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front. Hum. Neurosci.* **10**, 550 (2016).

21. Kube, T., Rief, W., Gollwitzer, M., Gärtner, T. & Glombiewski, J. A. Why dysfunctional expectations in depression persist - Results from two experimental studies investigating cognitive immunization. *Psychol. Med.* **49**, 1532–1544 (2019).

22. Everaert, J., Bronstein, M. V., Cannon, T. D. & Joormann, J. Looking through tinted glasses: depression and social anxiety are related to both interpretation biases and inflexible negative interpretations. *Clin. Psychol. Sci.* **6**, 517–528 (2018).

23. Rief, W. *et al.* Expectancies as core features of mental disorders. *Curr. Opin. Psychiatry* **28**, 378–385 (2015).

24. Dobson, K. S. & Dozois, D. J. A. *Handbook of Cognitive-Behavioral Therapies, Fourth Edition*. (Guilford Publications, 2021).

25. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1833 (2017).

26. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016).

27. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th Ed., Text Rev.)*. (American Psychiatric Association Publishing, 800 Maine Avenue, SW, Suite 900, Washington, DC 20024, 2022).

28. Bennett, D., Radulescu, A., Zorowitz, S., Felso, V. & Niv, Y. Affect-congruent attention modulates generalized reward expectations. *PLoS Comput. Biol.* **19**, e1011707 (2023).

29. Bromberg-Martin, E. S. & Sharot, T. The value of beliefs. *Neuron* **106**, 561–565 (2020).

30. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* **6**, 6149 (2015).

31. Karnick, A. T., Bauer, B. W. & Capron, D. W. Negative mood and optimism bias: An experimental investigation of sadness and belief updating. *J. Behav. Ther. Exp. Psychiatry* 101910 (2024).

32. Kube, T., Kirchner, L., Gärtner, T. & Glombiewski, J. A. How negative mood hinders belief updating in depression: results from two experimental studies. *Psychol. Med.* **53**, 1288–1301 (2023).

33. Samoilov, A. & Goldfried, M. R. Role of emotion in cognitive-behavior therapy. *Clin. Psychol.* **7**, 373–385 (2000).

34. Kuzmanovic, B., Jefferson, A. & Vogeley, K. NeuroImage The role of the neural reward circuitry in self-referential optimistic belief updates. **133**, 151–162 (2016).

35. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* **14**, 1475–1479 (2011).

36. Diederen, K. M. J., Spencer, T., Vestergaard, M. D., Fletcher, P. C. & Schultz, W. Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron* **90**, 1127–1138 (2016).

37. Schultz, W. Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27 (1998).

38. Koban, L. & Pourtois, G. Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neurosci. Biobehav. Rev.* **46 Pt 1**, 71–84 (2014).

39. Touroutoglou, A., Hollenbeck, M., Dickerson, B. C. & Feldman Barrett, L. Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *Neuroimage* **60**, 1947–1958 (2012).

40. Müller-Pinzler, L. *et al.* Neural pathways of embarrassment and their modulation by social anxiety. *Neuroimage* **119**, 252–261 (2015).

41. 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).

42. Kumar, P. *et al.* Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology* **43**, 1581–1588 (2018).

43. Seymour, B. *et al.* Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat. Neurosci.* **8**, 1234–1240 (2005).

44. Waltz, J. A. *et al.* Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology* **34**, 1567–1577 (2009).

45. Mulej Bratec, S. *et al.* Cognitive emotion regulation enhances aversive prediction error activity while reducing emotional responses. *Neuroimage* **123**, 138–148 (2015).

46. Phelps, E. A. Emotion and cognition: insights from studies of the human amygdala. *Annu. Rev. Psychol.* **57**, 27–53 (2006).

47. Murray, E. A. The amygdala, reward and emotion. *Trends Cogn. Sci.* **11**, 489–497 (2007).

48. Kumar, P. *et al.* Abnormal temporal difference reward-learning signals in major depression. *Brain* **131**, 2084–2093 (2008).

49. McHugh, S. B. *et al.* Aversive prediction error signals in the amygdala. *J. Neurosci.* **34**, 9024–9033 (2014).

50. Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).

51. Admon, R. & Pizzagalli, D. A. Dysfunctional reward processing in depression. *Curr Opin Psychol* **4**, 114–118 (2015).

52. Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J. & Drevets, W. C. Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *Am. J. Psychiatry* **169**, 152–159 (2012).

53. Safra, L., Chevallier, C. & Palminteri, S. Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Comput. Biol.* **15**, e1007224 (2019).

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

55. Gradin, V. B. *et al.* Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* **134**, 1751–1764 (2011).

56. Brolsma, S. C. A. *et al.* Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. *Psychol. Med.* **52**, 303–313 (2022).

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

58. Chen, C., Takahashi, T., Nakagawa, S., Inoue, T. & Kusumi, I. Reinforcement learning in depression: A review of computational research. *Neurosci. Biobehav. Rev.* **55**, 247–267 (2015).

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

60. 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).

61. Adams, G. C., Balbuena, L., Meng, X. & Asmundson, G. J. G. When social anxiety and depression go together: A population study of comorbidity and associated consequences. *J. Affect. Disord.* **206**, 48–54 (2016).

62. Kessler, R. C., Stang, P., Wittchen, H. U., Stein, M. & Walters, E. E. Lifetime co-morbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychol. Med.* **29**, 555–567 (1999).

63. 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).

64. 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).

65. Sliz, D. & Hayley, S. Major depressive disorder and alterations in insular cortical activity: a review of current functional magnetic imaging research. *Front. Hum. Neurosci.* **6**, 323 (2012).

66. 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).

67. 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).

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

69. Insel, T. *et al.* Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* **167**, 748–751 (2010).

70. Lockwood, P. L. & Klein-Flügge, M. C. Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Soc. Cogn. Affect. Neurosci.* **16**, 761–771 (2020).

71. 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).

72. 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).

73. 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).

74. 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).

75. Hoven, M., Luigjes, J., Denys, D., Rouault, M. & van Holst, R. J. How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health* **1**, 337–345 (2023).

76. Will, G.-J., Rutledge, R. B., Moutoussis, M. & Dolan, R. J. Neural and computational processes underlying dynamic changes in self-esteem. *Elife* **6**, 1–21 (2017).

77. Seow, T. X. F., Rouault, M., Gillan, C. M. & Fleming, S. M. How local and global metacognition shape mental health. *Biol. Psychiatry* **90**, 436–446 (2021).

78. Swallow, S. R. & Kuiper, N. A. Social comparison and negative self-evaluations: An application to depression. *Clin. Psychol. Rev.* **8**, 55–76 (1988).

79. McCarthy, P. A. & Morina, N. Exploring the association of social comparison with depression and anxiety: A systematic review and meta-analysis. *Clin. Psychol. Psychother.* **27**, 640–671 (2020).

80. Weary, G., Elbin, S. & Hill, M. G. Attributional and social comparison processes in depression. *J. Pers. Soc. Psychol.* **52**, 605–610 (1987).

81. Lewinsohn, P. M. A behavioral approach to depression. *Essential papers on depression* 150–172 (1974).

82. Maier, S. F. & Seligman, M. E. Learned helplessness: Theory and evidence. *J. Exp. Psychol. Gen.* **105**, 3–46 (1976).

83. Gross, J. J. & Muñoz, R. F. Emotion regulation and mental health. *Clinical Psychology: Science and Practice* **2**, 151–164 (1995).

84. Smith, R., Alkozei, A., Killgore, W. D. S. & Lane, R. D. Nested positive feedback loops in the maintenance of major depression: An integration and extension of previous models. *Brain Behav. Immun.* **67**, 374–397 (2018).

85. Sharot, T. & Garrett, N. Forming Beliefs: Why Valence Matters. *Trends Cogn. Sci.* **20**, 25–33 (2016).

86. Vandendriessche, H. & Palminteri, S. Neurocognitive biases from the lab to real life. *Commun Biol* **6**, 158 (2023).

87. Korn, C. W., Prehn, K., Park, S. Q., Walter, H. & Heekeren, H. R. Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience* **32**, 16832–16844 (2012).

88. Eil, D. & Rao, J. M. The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics* **3**, 114–138 (2010).

89. Villano, W. J. *et al.* Individual differences in naturalistic learning link negative emotionality to the development of anxiety. *Sci Adv* **9**, eadd2976 (2023).

90. Ertac, S. Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *J. Econ. Behav. Organ.* **80**, 532–545 (2011).

91. Zamfir, E. & Dayan, P. Interactions between attributions and beliefs at trial-by-trial level: Evidence from a novel computer game task. *PLoS Comput. Biol.* **18**, e1009920 (2022).

92. Brotzeller, F. & Gollwitzer, M. Exploring Asymmetries in Self-Concept Change After Discrepant Feedback. *Pers. Soc. Psychol. Bull.* 1461672241232738 (2024).

93. Everaert, J., Bronstein, M. V., Castro, A. A., Cannon, T. D. & Joormann, J. When negative interpretations persist, positive emotions don't! Inflexible negative interpretations encourage depression and social anxiety by dampening positive emotions. *Behav. Res. Ther.* **124**, 103510 (2020).

94. Roefs, A. *et al.* A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behav. Res. Ther.* **153**, 104096 (2022).

95. Hyman, S. E. Psychiatric disorders: grounded in human biology but not natural kinds. *Perspect. Biol. Med.* **64**, 6–28 (2021).

96. Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R. & Walters, E. E. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62**, 617–627 (2005).

97. Morrison, K. H., Bradley, R. & Westen, D. The external validity of controlled clinical trials of psychotherapy for depression and anxiety: a naturalistic study. *Psychol. Psychother.* **76**, 109–132 (2003).

98. Joormann, J. & Quinn, M. E. Cognitive processes and emotion regulation in depression. *Depress. Anxiety* **31**, 308–315 (2014).

99. Sears, C. R., Newman, K. R., Ference, J. D. & Thomas, C. L. Attention to emotional images in previously depressed individuals: an eye-tracking study. *Cognit. Ther. Res.* **35**, 517–528 (2011).

100. Kube, T. Biased belief updating in depression. *Clin. Psychol. Rev.* **103**, 102298 (2023).

101. Klayman, J. Varieties of Confirmation Bias. *Psychology of learning and motivation* **32**, 385–418 (1995).

102. Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends Cogn. Sci.* **26**, 607–621 (2022).

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

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

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

106. Garrison, J., Erdeniz, B. & Done, J. Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews* **37**, 1297–1310 (2013).

107. Wächter, T., Lungu, O. V., Liu, T., Willingham, D. T. & Ashe, J. Differential effect of reward and punishment on procedural learning. *J. Neurosci.* **29**, 436–443 (2009).

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

109. Donaldson, C., Lam, D. & Mathews, A. Rumination and attention in major depression. *Behav. Res. Ther.* **45**, 2664–2678 (2007).

110. Koster, E. H. W., De Raedt, R., Goeleven, E., Franck, E. & Crombez, G. Mood-congruent attentional bias in dysphoria: maintained attention to and impaired disengagement from negative information. *Emotion* **5**, 446–455 (2005).

111. Eizenman, M. *et al.* A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry Res.* **118**, 117–128 (2003).

112. Joormann, J. & Gotlib, I. H. Emotion regulation in depression: relation to cognitive inhibition. *Cogn. Emot.* **24**, 281–298 (2010).

113. Greenberg, L. S. Emotion-Focused Therapy: A Clinical Synthesis. *Focus* **8**, 32–42 (2010).

114. Rezaei, M., Ghazanfari, F. & Rezaee, F. The role of childhood trauma, early maladaptive schemas, emotional schemas and experimental avoidance on depression: A structural equation modeling. *Psychiatry Res.* **246**, 407–414 (2016).

115. 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).

116. Stolz, D. S., Müller-Pinzler, L., Krach, S. & Paulus, F. M. Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature Communications 2020 11:1* **11**, 1–13 (2020).

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

118. Whitton, A. E., Treadway, M. T. & Pizzagalli, D. A. Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Curr. Opin. Psychiatry* **28**, 7–12 (2015).

119. Krach, S. *et al.* Examining self-belief formation through artificial beliefs. *PsyArXiv Preprints* (2024) doi:10.31234/osf.io/2y5tv.

120. Wittchen, H.-U., Zaudig, M. & Fydrich, T. *SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II. Handanweisung*. (Hogrefe, 1997).

121. 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).

122. 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).

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

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

125. Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T. & Winther, O. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *J. Mach. Learn. Res.* **17**, 3581–3618 (2016).

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

127. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience* **32**, 551–562 (2012).

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

129. 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).

130. Murty, V. P. *et al.* Resting state networks distinguish human ventral tegmental area from substantia nigra. *Neuroimage* **100**, 580–589 (2014).

131. Ian C. Ballard, Vishnu P. Murty, R. McKell Carter, Jeffrey J. MacInnes, Scott A. Huettel and R. Alison Adcock. Dorsolateral prefrontal cortex drives mesolimbic dopaminergic regions to initiate motivated behavior. *J. Neurosci.* **31**, 10340–10346 (2011).