How underconfidence is maintained in anxiety and depression

Sucharit Katyal[1], Quentin JM Huys[1,2,3], Raymond J Dolan[1,2], & Stephen M Fleming[1,2,4]

[1] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, UCL

[2] Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, UCL

[3] Mental Health Neuroscience Department, Division of Psychiatry, UCL

[4] Department of Experimental Psychology, UCL

Corresponding author:

Sucharit Katyal

s.katyal@ucl.ac.uk

Abstract

Individuals with anxiety and depression exhibit chronic metacognitive biases such as underconfidence. The origin of such biases is unknown. In two large general population samples (N=230 and N=278), we studied metacognition both locally, as confidence in individual task instances, and globally, as longer run self-performance estimates, while quantifying the impact of feedback valence on confidence. Global confidence was sensitive to both local confidence and feedback valence – more frequent positive (negative) feedback increased (respectively decreased) global confidence. Feedback valence impacted confidence in a domain-general fashion and also led to shifts in affective self-beliefs. Notably, global confidence was more sensitive to low (vs. high) local confidence in individuals with greater transdiagnostic anxious-depression symptomatology, despite sensitivity to feedback valence remaining intact. Together, our results reveal a mechanistic basis for chronic underconfidence in anxious-depression rooted in distorted interactions between local and global metacognition, while elucidating a method to restore confidence through targeted feedback.

Introduction

Computational and cognitive approaches in neuroscience and psychiatry have made great strides in understanding how humans perceive and represent their environment. A significant component of human mental activity, however, involves how we think about ourselves. For instance, metacognitive beliefs about our skills and abilities have a pervasive impact in educational and clinical settings, affecting people's decisions about whether to pursue new activities[1]. In experimental studies of metacognition, a notably robust link has been established between transdiagnostic symptoms of anxiety and depression (AD), and domain-general underconfidence in performance. These relationships are observed in both "local" confidence on individual trials[2–4] and "global" confidence measured through long-run self-performance estimates (which we refer to as global SPEs[5,6]). Conversely, remission from depression symptoms, through either therapy or antidepressants, ameliorates underconfidence[7]. However, previous studies examining the link between metacognition and symptoms have been descriptive, and a mechanistic understanding of why confidence distortions in AD persist despite good performance remains elusive.

One promising route to understanding the source of metacognitive biases in AD is to unpack how confidence is formed. For instance, Wittmann et al.[8] found that false feedback on performance in an ambiguous task influenced people's global confidence estimates. Similarly, Rouault et al.[9,10] found that providing true performance feedback, compared to a no-feedback condition, increased global SPEs, despite objective task performance remaining unaffected. More broadly, negative affect reduces confidence on an unrelated task[11], whereas positive affect increases confidence[12]. Manipulating monetary payoffs also increase and decrease local confidence when reward expectations are high and low respectively[13]. Similarly, manipulating prior beliefs about performance on an upcoming task, either through false feedback or expected task difficulty, can decrease or increase local confidence on subsequent instances of a similar task[14]. Finally, local confidence is higher on trials immediately following positive compared to negative feedback[15], suggesting that feedback may shape longer-timescale changes in global confidence. More generally, confidence may act as an internal reinforcement

signal[16,17] in the absence of external feedback (or reward) that is weighed in conjunction with external feedback when forming global self-performance estimates[9].

However, whether and how confidence formation differs in people with high AD symptoms remains unexplored. One attractive hypothesis is that a negative relationship between confidence and AD symptoms may be grounded in a tendency for depressed individuals to incorporate less positive and more negative information when making future predictions, particularly about themselves[18–22]. Recent work has found such a self-related negativity bias in AD individuals in relation to learning about expectations of future adverse life events[21,23] and when updating performance expectations from feedback[22]. Such biases could predispose depressed individuals to overweigh individual instances of negative feedback and/or low local confidence when forming global SPEs. In turn, local and global metacognitive evaluations may not only inform each other, but potentially shape abstract self-beliefs such as self-esteem[6,10,24], which by definition transcends individual tasks or cognitive domains. It is therefore also important to establish whether asymmetries in global confidence formation transfer across different tasks and/or influence more distal self-beliefs. Such transfer would be consistent with global confidence reflecting a slowly changing cognitive state with its own dynamics, akin to mood[25].

In the present study, we sought to determine how global confidence formation is affected by anxious-depression symptoms. We additionally investigated how systematically manipulating performance feedback impacts global confidence formation, and whether the influence of feedback on confidence generalises across two distinct cognitive domains (perception and memory). These questions were addressed first in an exploration sample (Exp 1, N = 230), and then in a replication sample (Exp 2, N = 278) with preregistered hypotheses and analysis plan (osf.io/7xfqw). In Experiment 2, we also tested whether our performance feedback

manipulation impacted broader facets of affective self-evaluation. A detailed description of each research question is provided at the beginning of the Methods section.

Results

Each experiment involved six blocks of gamified perception and memory tasks designed to measure performance and confidence in distinct cognitive domains (Figure 1). On each trial, participants were required to choose the correct response (the higher density colour on the perceptual task; the stimulus that was in a previous array on the memory task; see Figure 1 for details) before rating their confidence in their response on a continuous scale. A continuous staircase procedure targeting ~71% correct performance ensured performance equalised across both subjects and tasks. After completing baseline blocks of the task(s) without feedback, participants completed "intervention" blocks where they received intermittent feedback on their performance (delivered by an "auditor" who participants were told would occasionally check their performance). Depending on the block type, the auditor was rigged to appear more often on correct trials (positive feedback blocks) or incorrect trials (negative feedback blocks); see Methods for details. Intervention blocks were interleaved with test blocks without feedback. At the end of each block, participants provided a global self-performance estimate (SPE) on a sliding scale.

Figure 1A depicts our key measures and manipulation at the block level – each trial was followed by local confidence reports with occasional (veridical) performance feedback and a global SPE report after the trial block was completed. Participants were randomised to one of eight groups which differed according to the order of the intervention blocks they received (positive or negative first), whether the intervention was delivered on the perception or memory task, and whether the subsequent test blocks were in the same or different domain as the intervention blocks (Figure 1B; see Supplementary Figure 1 for Exp 2). Figure 1C illustrates one sample trial for each of the two tasks. Exp 2 was a replication of Exp 1 with

minor differences to the design (see Methods). In what follows, we report the effects from both Experiments together, noting any inconsistencies between datasets.

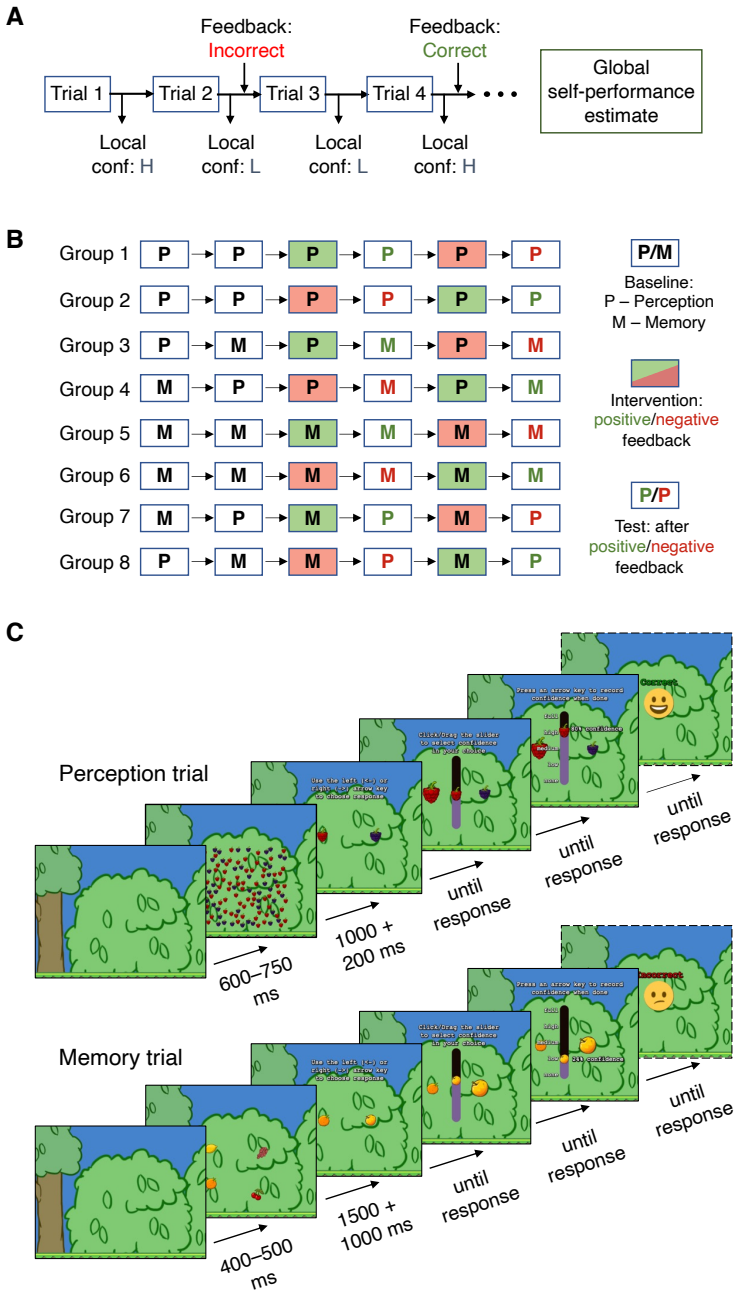


Figure 1. **A)** A typical task block where individual trials involving stimulus and decisions were followed by confidence reports and occasional (veridical) performance feedback. After completing the block, participants provided a self-performance estimate for that block. **B)** The eight groups of participants and the order in which they performed the perception (P) and memory (M) tasks in Exp 1. Participants within a group performed one or both tasks. A task sequence began with two baseline blocks followed by two sets of intervention blocks (with feedback), interleaved with test blocks (without feedback). For groups that performed both tasks, their order during baseline was randomised across participants. Unfilled boxes indicate the absence of feedback. Green filled boxes indicate positive feedback blocks (more feedback on correct than incorrect trials) and red blocks indicate negative feedback blocks (more feedback on incorrect than correct trials). The colour of the subsequent test block indicates whether feedback delivered on the preceding (intervention) block was positive (green) or negative (red). **B)** Single trials of the perception (top) and memory (bottom) tasks. For the perception task, participants attended the green bush in the middle where 121 stimuli of two different kinds of berries (red raspberries and dark purple blackberries) would appear at non-overlapping random locations for 1000 ms. Participants were asked to respond as to whether there were more raspberries or blackberries, and then used a slider to report their confidence. Feedback for correct or incorrect trials appeared with a defined frequency depending on the intervention block (positive or negative). For the memory task, a set of fruits appeared on the bush for 1500 ms. Participants were tasked to memorise the fruits and then decide between two options, which fruit was present in the set. Confidence reporting and feedback for the memory task was similar to the perception task.

Feedback impacts confidence without changing performance

Our asymmetric feedback manipulation led to systematic shifts in global SPEs, despite having no impact on actual performance on either of the two tasks (Exp 1: Figure 2A; Exp 2: Supplementary Figures 2). Specifically, positive feedback intervention blocks led to higher global SPEs than baseline blocks, whereas negative feedback intervention blocks led to lower global SPEs, resulting in a significant main effect of *Feedback type* in predicting intervention block global SPEs (Exp 1: positive – negative = 0.25 ± .01 (mean & SE), $t(228) = 23.56$, $p < .0001$; Exp 2: positive – negative = 0.28 ± .01, $t(282) = 26.53$, $p < .0001$). To control for potential order effects, we also modelled 3- and 2-way interactions between *Feedback type* (positive, negative), *Task* (perception, memory) and *Feedback order* (positive first, negative first), none of which were significant (all $p > .15$). Metrics of first-order performance did not differ between positive and negative feedback blocks: mean accuracy (Exp 1: $t(229) = -.18$, $p = .86$, Exp 2: $t(554) = .49$, $p = .62$); mean difficulty level (Exp 1: $t(229) = 1.20$, $p = .23$; Exp 2: $t(277) = -1.13$, $p = .26$; Supplementary Figures 3), and neither accuracy or difficulty showed interactions with *Task* (all $p > .19$).

Replicating previous findings[9,26], mean local confidence (metacognitive bias) significantly predicted global SPEs (Exp 1: $t(402) = 5.91$, $p < .0001$; Exp 2: $t(488) = 4.35$, $p < .0001$). In contrast, mean accuracy did not predict global SPEs after effects of mean local confidence were taken into account (Exp 1: $t(446) = -.74$, $p = .46$; Exp 2: $t(451) = .52$, $p = .60$). The absence of any interaction or main effect of *Feedback order* in these analyses (all $p > .15$) led us to collapse over the two possible feedback orders in subsequent analyses.

As the trials on which feedback was given to participants were determined probabilistically, we next asked whether the actual proportion of positive/negative feedback trials received on each block influenced subsequent global SPEs. We observed that a higher proportion of positive feedback trials boosted global SPEs (Exp 1: estimate = .55 ± .04, t(696) = 15.54, p < .0001; Exp 2: estimate = .59 ± .04, t(854) = 14.51, p < .0001) whereas a higher proportion of

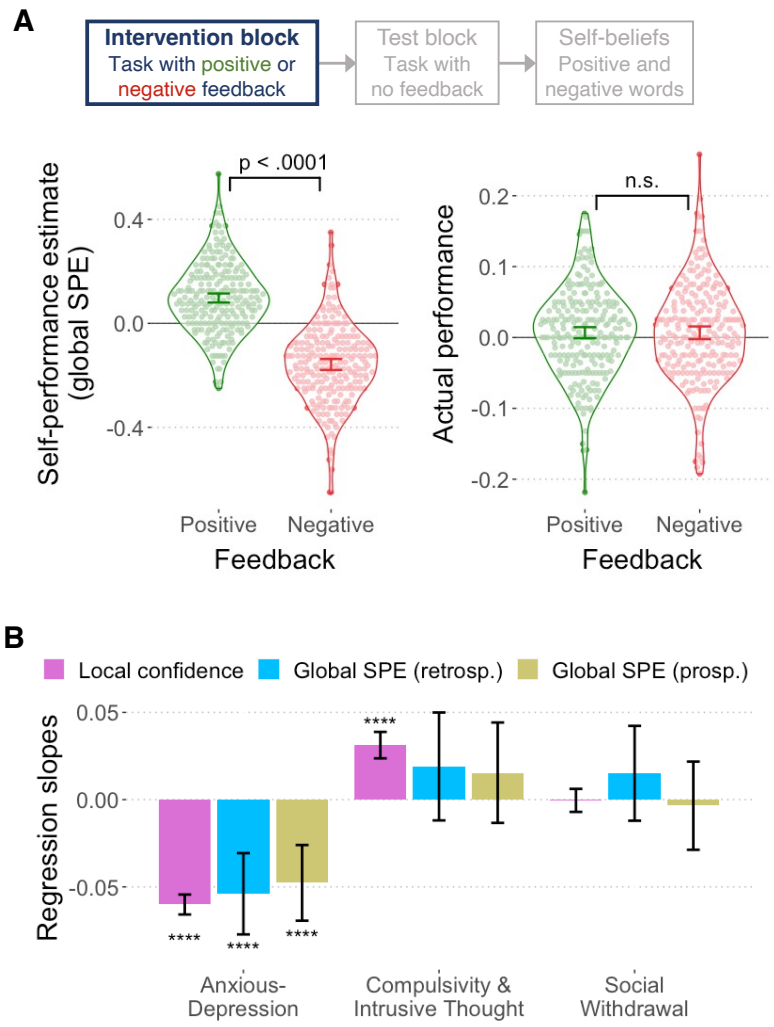


Figure 2. **A)** Intervention blocks had either more frequent positive (green) or negative feedback (red). Global confidence measured as self-performance estimates (global SPEs, left panel), and performance (mean accuracy; right panel) on intervention blocks in Exp 1 with baseline values subtracted out for each individual. Each dot is an individual participant. Error bars show 95% bootstrapped confidence intervals. **B)** Regression coefficients depicting the relationship in Exp 2 between transdiagnostic symptom axes and mean local confidence on individual trials, global SPEs measured retrospectively and global SPEs measured prospectively. ****p < .0001.

negative feedback trials reduced global SPEs (Exp 1: estimate = –.70 ± .04, t(689) = –17.12, p < .0001; Exp 2: estimate = –.80 ± .05, t(859) = –17.48, p < .0001).

Finally, we examined whether asymmetric feedback delivered on the perception task affected confidence in the memory task, and vice-versa. In both experiments we found evidence for a domain-general effect of intervention-block feedback on test block confidence (see Supplementary Material for a full analysis). In Exp 1, average local confidence was significantly higher on memory test blocks following positive compared to negative feedback perception blocks (Supplementary Figure 4). This was not the case for perception test blocks following memory intervention blocks. In Exp 2, we found a significant domain-general effect of feedback in both directions – from perception to memory, and from memory to perception.

Distortions in the formation of global confidence as a function of anxious-depression symptoms

We next asked how symptoms of depression and anxiety (measured using standardised questionnaires in Exp 1, and transdiagnostically in Exp 2) related to the formation of global confidence estimates. First, we replicated previous findings showing that heightened depression and anxiety symptoms are associated with lower average local confidence (metacognitive bias; [2]). In Exp 1, individuals with greater anxiety levels showed lower mean baseline local confidence (GAD-7, (t(440) = –2.65, p = .008; mini-SPIN, t(441) = –2.88, p = .004), although this was not the case for depression scores (PHQ-9, t(440) = –.79, p = .43). Baseline global SPEs were also negatively related to both depression (PHQ-9, t(443) = –3.45, p = .0006) and anxiety levels (GAD-7, (t(443) = –4.91, p < .0001; mini-SPIN, t(443) = –3.75, p = .0002). For Exp 2, we obtained transdiagnostic scores for each participant across the three transdiagnostic symptom axes, as identified by Gillan et al.[27]: anxious-depression (AD), compulsivity and intrusive thought (CIT), and social withdrawal (SW). We estimated scores along each axis using a reduced questionnaire battery developed by Hopkins et al.[28].

Replicating earlier work[2], we found that baseline local confidence had a significantly negative association with AD (t(20919) = –20.62, p < .0001), a significantly positive association with

CIT ($t(20920) = 8.08$, $p < .0001$) and no association with SW ($t(20919) = –.12$, $p = .90$; Figure 2B). Also consistent with recent work[5], we established that global SPEs were significantly negatively associated with AD ($t(538) = –4.51$, $p < .0001$), but not with CIT ($t(538) = 1.19$, $p = .23$) or SW ($t(538) = 1.09$, $p = .28$). In Exp 2, we also measured *prospective* global confidence, as self-performance estimates on the two tasks before participants performed a single trial of either task but after they were informed what the tasks would constitute. Similar to retrospective global SPEs, prospective SPEs were significantly negatively related to AD ($t(538) = –4.29$, $p < .0001$), but not CIT ($t(538) = 1.04$, $p = .30$) or SW ($t(538) = –.26$, $p = .79$).

We next sought to develop distinct computational accounts of cognitive distortions that might impact global confidence formation in high AD individuals (Figure 3A; in Supplementary Material we report how our model parameters vary with changes in the CIT axis). These accounts make different predictions for how individuals with high and low AD symptoms would generate global SPEs by modelling potential asymmetries in sensitivity to both positive vs. negative feedback, and high vs. low confidence. Regression parameters relating symptom scores to a variety of learning asymmetries were recovered in simulation (Supplementary Figure 5), underscoring a robustness to our approach. To visualise the impact of these asymmetries on qualitative data patterns, we simulated the different models and quantified their impact on global SPEs following intervention blocks with feedback (baseline-corrected) and both baseline and test blocks without feedback (not baseline-corrected).

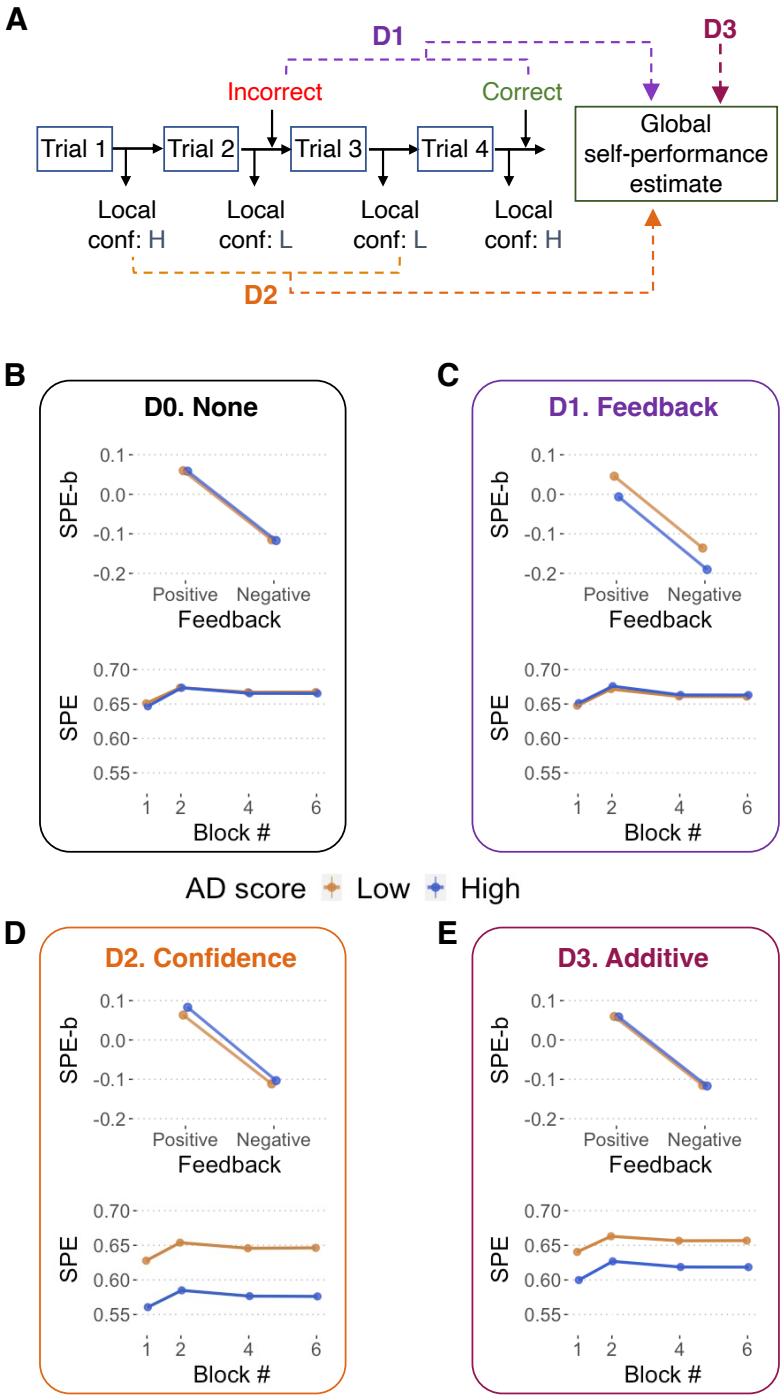


Figure 3. **A)** A schematic of three possible distortions that can explain why underconfidence is maintained in anxiety and depression. D1 – Feedback distortion (purple): global SPE formation is more sensitive to negative than positive feedback. D2 – Confidence distortion (orange): global SPE formation is more sensitive to low than high local confidence. D3 – Additive distortion (maroon): global SPEs are computed from local confidence and feedback without distortion, but with an end-of-block bias. **B-E)** The three accounts make distinct qualitative prediction for how AD symptoms impact baseline-subtracted SPEs on feedback (intervention) blocks (SPE-b, upper sub-panels) and global SPEs across the four non-feedback (baseline plus test) blocks. This is contrasted with a null model without distortions (D0) that does not predict a change in SPE with anxious-depression symptoms.

In a "no distortion" model (D0), high AD individuals show no biases in forming global from either feedback or confidence. Simulations show (Figure 3B) that in this instance high and low AD symptoms should not impact either intervention-block global SPEs (sorted into positive/negative feedback blocks; upper panel) or test-block global SPEs (on block numbers 1, 2, 4, 6; lower panel). According to a feedback distortion account (D1), global SPEs in high AD individuals are more sensitive to negative feedback and less sensitive to positive feedback compared to low AD individuals. Here, global SPEs are expected to be reduced in high AD individuals following feedback (intervention) blocks but not following non-feedback (baseline, test) blocks (upper and lower panels of Figure 3C). Alternatively, according to a confidence distortion account (D2), high AD symptoms lead to global SPEs being more sensitive to low local confidence and less sensitive to high local confidence. Under this model, uncorrected SPEs should be reduced for high AD individuals (Figure 3D; lower panel), but baseline-corrected SPEs on feedback blocks are in fact predicted to slightly increase (upper panel). The latter seemingly counterintuitive effect follows from the fact that on feedback blocks there are fewer trials on which participants rely on their local confidence to form end-of-block global SPEs compared to non-feedback blocks. Thus, there is lesser accumulation of distortion due to AD symptoms from local confidence to global SPEs. Finally, we also considered a non-learning-based account of global SPE distortions (additive bias; D3) where high AD individuals simply decrement their global SPEs at the end of each block proportional to their AD symptom score and do not exhibit asymmetries in their sensitivity to local confidence or feedback. In this account, a reduction in global SPEs is observed in the uncorrected blockwise data (Figure 3E; lower panel), but not baseline-corrected data (upper panel).

We next asked how these alternative models fared in capturing the data in relation to individual variation in AD symptoms within Exps 1 and 2. Figures 4A—4D show the data from Exp 1 and Exp 2 plotted using the same conventions as the model simulations. As expected, overall SPEs were lower in individuals with higher PHQ and AD axis scores in Exp 1 and Exp 2 respectively (Figure 4C and 5D), ruling out null model D0. At the same time, in both datasets, higher AD scores were also associated with slightly higher global SPEs after feedback compared to baseline blocks, thus also discounting the possibility of a feedback-only model

(D1). This slight increase in SPEs as a function of AD symptoms, particularly after positive feedback blocks, moreover suggests (counterintuitively, as explained above) the presence of a distortion in sensitivity to local confidence (low > high) when forming global SPEs (model D2). However, as this effect is small, these data patterns are also consistent with a combination of confidence and feedback distortions, and confidence distortions plus additive biases.

To adjudicate between these possibilities, we next pursued a formal model comparison. In addition to models D0—D3, we also considered composite models that combined feedback with confidence distortions (D4), feedback distortions with additive biases (D5), and confidence distortions with additive biases (D6). Model comparison based on DIC scores (Figure 4E) revealed that a confidence distortion plus additive bias model (D6) provided the best fit to data in both Exp 1 and Exp 2. In general, models that included a confidence distortion either alone or in combination with another distortion (i.e., models D2, D4 and D6) fit the data better than other models. Figure 4F shows the model-estimated posteriors of the regression slope parameters relating anxious-depression symptoms to confidence distortions (β_c) and feedback distortions (β_f) from model D4 in Exp 1 (PHQ) and Exp 2 (AD axis). In both experiments, β_c was significantly negative (Exp 1: 99% HDI = [–.047 –.038]; Exp 2: 99% HDI = [–.09 –.03]) while β_f did not differ from 0 (Exp 1: 99% HDI = [–.047 –.038]; Exp 2: 99% HDI = [–.09 –.03]). Fitting model D6 to both experiments yielded similar values of β_c (Supplementary Figure 6A). In Exp 1, β_c was also negative for GAD (99% HDI = [–.036 –.028]) and mini-SPIN (99% HDI = [–.048 –.034]). Taken together, our modelling results confirm that people with greater anxious-depression symptoms are more sensitive to low vs. high confidence trials when forming global SPEs.

Contrary to our expectation, β_a was significantly positive in both experiments (Exp 1: PHQ, 99% HDI = [2.0e-07 3.9e-05]; Exp 2: AD, 99% HDI = [4.5e-05 7.8e-04]; Supplementary Figure 6B), suggesting that, if anything, AD symptoms are associated with a *positive* additive shift in end-of-block SPEs. This was not only the case for model D5 but also model D3, which modelled additive biases in isolation from distortions in within-block learning. However, the magnitudes of fitted values of β_a were small in both experiments, with simulations using

similar values revealing negligible impact on observed global confidence distortions (Supplementary Figure 6C). Overall, our results from both Exp 1 and Exp 2 indicate individuals

with higher anxious-depression scores have a greater sensitivity to low compared to high local

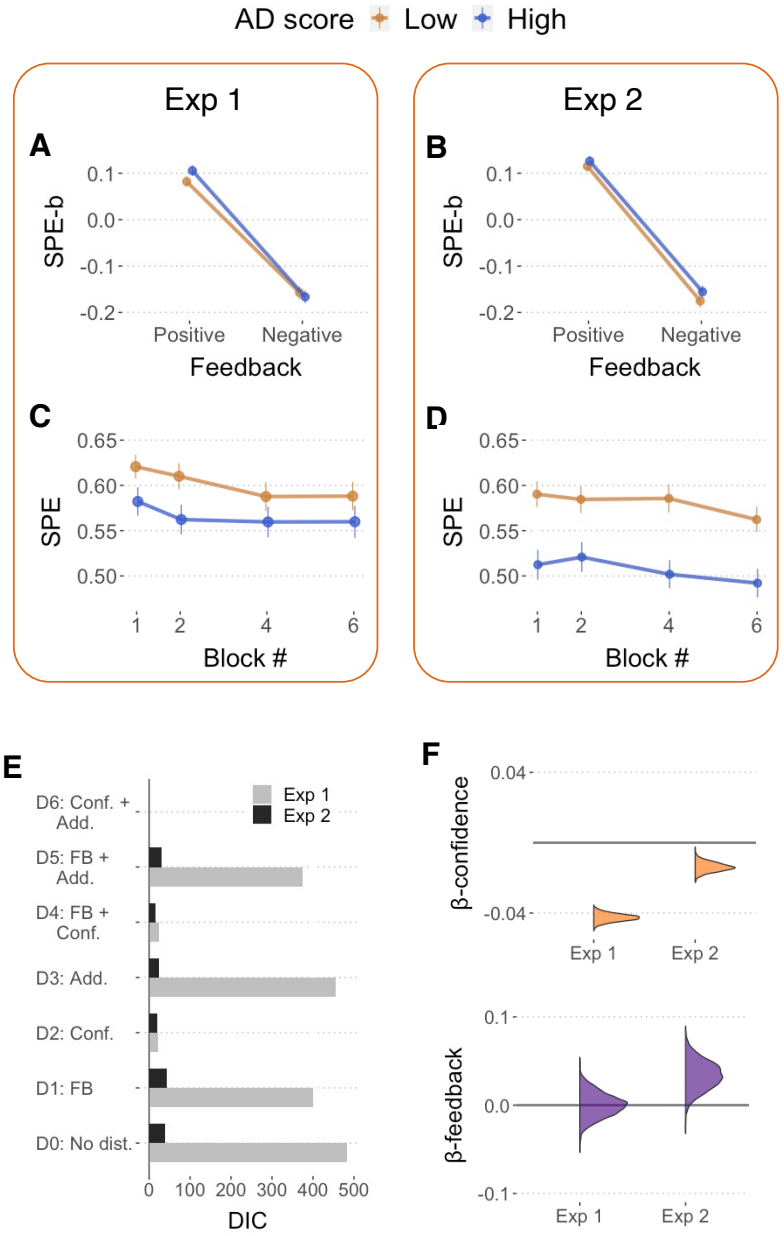


Figure 4. Baseline-subtracted global SPEs (SPE-b) on feedback blocks in **A)** Exp 1, and **B)** Exp 2 for high and low AD individuals. Global SPEs on confidence-only (i.e., non-feedback) blocks in **C)** Exp 1, and **D)** Exp 2. Data from both experiments conformed to qualitative predictions of the confidence distortion model (orange outline). **E)** DIC scores for the seven possible models of how different distortions in global SPE formation relate to AD symptoms score: D0 – No distortion, D1 – Feedback distortion only, D2 – Confidence distortion only, D3 – Additive biases only, D4 – Feedback and confidence distortions, D5 – Feedback distortion and additive bias, D6 – Confidence distortion and additive bias. **F)** Posterior distributions of regression slope parameters relating AD symptoms to confidence and feedback distortions for model D4 in Exp 1 and Exp 2. Regression coefficient values from Exp 2 are divided by 5 (corresponding to differences in the range of PHQ and AD axis scores) to allow direct comparison of coefficients.

confidence when forming global SPEs.

Confidence, feedback, and affective self-beliefs

Finally, we asked whether our feedback intervention, which robustly modulated global SPEs, generalised to a more distal measure of affective self-evaluations (see Methods). We first replicated earlier work showing that individuals with depression and anxiety self-endorse more negative and fewer positive affect words compared to matched control groups[29–32] (Figure 5A; Supplementary Material). We also observed a valence-specific association of self-endorsements with baseline local confidence (Exp 1: $z = -3.61$, $p = .0003$, Supplementary Figure 7D; Exp 2: $z = -6.93$, $p < .0001$, Figure 5B) and baseline global SPEs (Exp 1: $z = -6.70$, $p < .0001$, Supplementary Figure 7E; Exp 2: $z = -2.38$, $p = .018$; Supplementary Figure 8C), with a positive relationship between confidence and self-endorsement to positive words, and a negative relationship between confidence and self-endorsement to negative words.

In Exp 2, we presented one set of words at the beginning of the study and a second set immediately following the first test block (which in turn occurred after the first intervention block). Our key question was whether the type of performance feedback received would impact affective self-evaluation. We observed a significant interaction between *Feedback type* and word valence on the change in self-endorsements (Figure 5C; $t(593) = 2.46$, $p = .014$). When directly comparing self-endorsement profiles following positive vs. negative feedback blocks, there was a significant reduction in self-endorsement for negative words ($t(593) = -2.16$, $p = .03$), and a trend for higher self-endorsement for positive words ($t(594) = 1.03$, $p = .19$). At the same time, participants who received negative performance feedback during

intervention blocks, self-endorsements of negative compared to positive words increased ($t(593) = 4.21$, $p < .0001$). There was no such difference in the change in self-endorsement following positive performance feedback ($t(593) = .65$, $p = .52$). Overall, our results indicate that interventions that change participants global SPEs can also modulate affective self-evaluation.

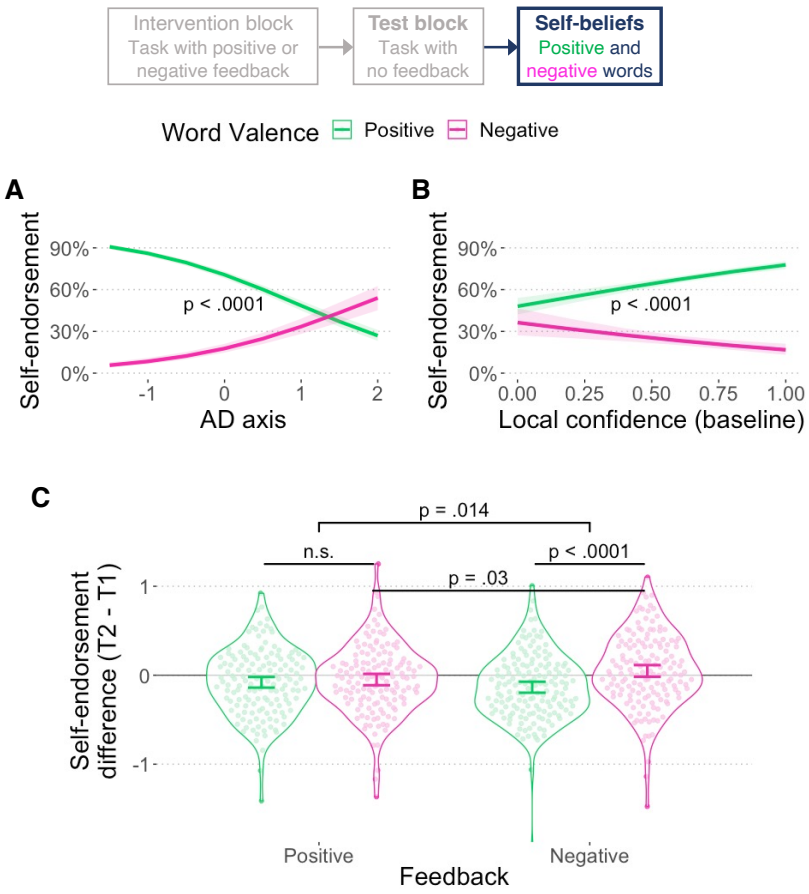


Figure 5. Valence-specific relationship of self-endorsement to positive and negative words at the first timepoint (T1) of the self-referential encoding task in Exp 2 plotted as marginal effects from a logistic regression showing that **A)** the anxious-depressive (AD) transdiagnostic axis is negatively related to positive word endorsement and positively related to negative word endorsement, and **B)** mean confidence during baseline runs is positively related to positive word endorsement and negatively related to negative word endorsement. **C)** Change in word endorsement from T1 to T2 (i.e., post-feedback) separated by word valence and the type of feedback delivered during intervention blocks. While there was no difference in the change self-endorsements between positive and negative words following positive feedback, negative words were endorsed more than positive words following negative feedback. Error bars show 95% bootstrapped confidence intervals.

Discussion

Distortions in self-beliefs represent a central feature of mental ill health[19,33]. Notably, individuals high in anxious-depression (AD) symptoms tend to be chronically underconfident in their abilities both when evaluating their confidence on a particular task instance[2–4] and when providing global self-performance estimates (SPEs)[5]. Here we asked how local feedback and confidence are integrated into global SPEs, and how such integration is distorted by AD symptoms. We show that individuals with higher AD symptoms are unduly sensitive to instances of *low* local confidence when forming their global SPEs. Notably, however, this asymmetry in learning was not observed for explicit (negative vs. positive) feedback – indicating that distortions in global confidence formation are rooted in linkages between local and global metacognition, rather than due to a generalised bias in learning.

Past work has proposed that global SPEs are formed by combining local confidence on individual trials with external feedback about performance[2,8,9,26]. Here we artificially created situations in which participants received predominantly negative feedback, or predominantly positive feedback, despite their performance being equated in both cases. Such situations mimic real-world scenarios in which teachers or supervisors may be stern, mostly giving negative feedback in response to mistakes, or gentle, tending to give positive feedback following successes. In both experiments, we observed strong effects of such feedback asymmetries on global SPEs in both the perception and memory task domains. By demonstrating a robust impact of feedback on global confidence, we extend previous work depicting malleability of local confidence through affective induction[11,12], expected reward[13,34] and feedback[14,15].

We also extend upon a previously proposed computational model[9] of global SPEs by introducing parameters governing an asymmetry in learning from positive vs. negative feedback and high vs. low local confidence. This extension is reminiscent of similar work modelling asymmetries in learning from external reinforcement/information[35–39]. The favoured model provided good fits to our data across both experiments, and in so doing opens

a rich framework within which to understand how AD symptoms result in distortions of global confidence.

Individuals diagnosed with clinical depression and those scoring high on AD symptoms, have been shown to exhibit greater sensitivity to negative compared to positive information, particularly when self-relevant[18–20,30]. Recent work shows that this may involve a learning deficit, where depressed individuals learn more from negative and less from positive information[21–23,40]. We tested if similar distortions existed both with regards to incorporating more negative vs, positive feedback and/or low vs. high local confidence trials into the formation of global confidence. We also considered an alternative account of distortions in global confidence, in which greater AD symptoms lead to additive biases in confidence rather than acting through changes in learning. In both experiments, we found evidence for a distortion in learning from local confidence – individuals with higher AD scores were more sensitive to low than high confidence trials when forming their global confidence estimates. In contrast, no distortion was observed in relation to negative compared to positive feedback – suggesting that asymmetries in learning were specific to metacognitive variables (the link between local and global confidence).

One implication of intact feedback processing in high AD individuals is that asymmetric feedback schedules may be able to "override" metacognitive biases that would otherwise exist in the absence of feedback. However, such an intervention would only be practically meaningful were it to be persistent and generalisable across tasks, and potentially to other more distal metrics of self-evaluation. We indeed found, in both experiments, that an impact of feedback on confidence persisted beyond intervention blocks to also affect confidence on "test" blocks without feedback: during test blocks, local confidence was higher when the previous intervention blocks had more frequent positive compared to negative feedback. Strikingly, this effect of feedback on test block confidence was observed not only when the two tasks were the same, but also when they differed (with the effect being particularly robust when feedback was delivered on a perception task and confidence assayed on a memory task). These results indicate that receiving feedback within a particular domain may have broader,

domain-general effects on people's confidence – consistent with previous work showing that affective induction[11,12] or reward expectations[13] also impact domain-general confidence.

Notably, our model-based approach to estimating the influence of feedback on confidence can be used to simulate an expected change in metacognition under different feedback schedules for different individuals – highlighting a potential for developing personalised interventions to ameliorate systematic biases in over- and under-confidence[2,33,41,42]. As a step in this direction, we investigated if our feedback intervention impacted affective self-evaluations known to be a characteristic of mental health disorders, particularly anxious-depression[29–32]. We found that asymmetries in feedback not only robustly shifted global confidence, but also led to significant changes in self-endorsement of positive vs. negative words – highlighting the potential of metacognitive interventions for changing affective self-beliefs. Future work could seek to calibrate feedback schedules to target adaptive changes in self-beliefs (e.g., in the case of AD, through use of a higher relative frequency of positive feedback, over a longer timescale, and in the context of multiple tasks or social environments).

In summary, we identify a mechanistic source of metacognitive distortions present in individuals with high anxious-depression symptoms. Specifically, high AD individuals exhibit greater sensitivity to instances of low local confidence when forming global self-performance estimates, potentially explaining the maintenance of a chronic underconfidence in this population. The absence of an equivalent asymmetry when processing external feedback points to the potential of calibrated feedback interventions in ameliorating metacognitive biases. Our work provides a principled computational framework within which to understand the causes of distortions in global confidence across the spectra of psychopathology.

Methods

Research questions

We designed our study to answer four inter-related research questions (explicitly preregistered before collection of a replication sample in Exp 2 on 20th January 2023). Here we outline each question and high-level features of the experimental design, before providing details on the methodology and data analysis.

1) Does feedback selectively inform global self-performance estimates (SPEs)?

To address this question, we manipulated the probability with which participants received feedback on correct and incorrect trials. Each participant underwent one positive feedback "intervention" block where they received feedback on approximately 22.5% of correct and 2.5% of incorrect trials and one negative feedback intervention block where these target feedback proportions were reversed (positive-negative order counterbalanced; see Methods; Figure 1B). We hypothesised that participants' SPEs would be higher following positive than negative feedback blocks, despite performance remaining constant. Importantly, we did not give false feedback, and capitalised on variability in correct and incorrect trials within each block to adaptively deliver feedback depending on the intervention block.

2) Do individuals with high self-reported anxious-depression (AD) symptoms exhibit asymmetries when forming SPEs from negative vs. positive feedback and/or low vs. high local confidence?

We first sought to replicate past work showing negative relationships between baseline local and global confidence and AD scores[2,5]. We then tested whether AD moderated the relationship between positive and negative feedback and SPEs. We also tested for asymmetries

in learning rates within a computational model in which SPEs are formed by combining feedback and local confidence[9].

3) Do performance feedback interventions affect (local and global) confidence on subsequent tasks that involve in different cognitive domains?

For this, we used two performance-controlled tasks – a visual perception task (density estimation) and a visual working memory task – while manipulating the type of transfer (to the same or different task) between participant groups.

4) Does a change in global confidence in cognitive tasks elicited through feedback also transfer to affective self-beliefs?

We used a self-referential encoding task (SRET) wherein participants provided self-endorsements to positive and negative adjectives. Higher AD symptoms are characterised by more negative and fewer positive self-endorsements in this task [29–32]. In Exp 1, we asked whether local and global confidence were correlated with biases in self-endorsement. In Exp 2, we tested whether our performance feedback intervention impacted affective self-evaluation by measuring positive / negative self-endorsements before and after task performance.

Note that to assess the impact of feedback on affective self-beliefs, our preregistered analysis aimed to regress the type of feedback intervention (positive/negative) upon the difference between positive and negative self-endorsements (i.e., the double difference between timepoint and word valence). In the main text we deviated from this analysis plan and instead report the interaction of feedback type and word valence upon change in self-endorsements, with the expectation that positive feedback will increases (decreases) positive (negative) self-endorsements, and that negative feedback would show the opposite pattern. We report our preregistered analysis in Supplementary Material.

Participants

We recruited participants through the online platform Prolific (prolific.co), and included individuals between ages 18—55 years who reported being fluent in English. Participants were recruited in two non-overlapping samples – Exp 1 (exploration) and Exp 2 (preregistered

replication; osf.io/7xfqw). After dropouts and exclusions (see Supplementary Materials for details along with sample justification), a total of 230 participants remained for Exp 1 (age mean and SD = 32 ± 9; gender: 127 females, 185 males, 2 nonbinary) and 278 participants for Exp 2 (age mean and SD = 32 ± 9; gender: 176 females, 278 males, 3 nonbinary).

Design

Our key manipulation was the extent to which individuals received (veridical) positive or negative feedback related to their decision at the end of each trial (after they provided a confidence estimate). This manipulation involved experimentally controlling the probability with which feedback was delivered across blocks, with some blocks delivering more positive than negative feedback, and others delivering more negative than positive feedback. Blocks were designated as baseline (blocks 1 and 2; no feedback), intervention (blocks 3 and 5; with feedback), and test (blocks 4 and 6; no feedback). *Feedback type* was manipulated as a within-subjects factor. We also manipulated two between-subject factors: 1) *Intervention task* (perception or memory) and 2) *Transfer type* (whether test blocks followed an intervention block of the same or opposite task). Because *Feedback type* was a within-subjects factor, we controlled for order effects by manipulating *Feedback order* between individuals (i.e., whether positive or negative feedback blocks were encountered first). This resulted in eight groups of participants for Exp 1. Figure 1B shows the order of task blocks (perception and/or memory) for each of the eight groups. Participants assigned to groups transferring to the same task repeated that task in the two baseline blocks while others performed one block of each task whose order was randomised across participants.

Exp 2 had a similar design to Exp 1, except we no longer manipulated *Transfer type*. Instead, test blocks always consisted of the opposite task to the one used during the intervention block, where we sought to replicate findings in Exp 1 of cross-domain transfer in the effects of feedback on confidence. We assessed whether the effects of feedback on confidence modulated affective self-beliefs. To this end, participants self-endorsed two sets of 20 affective words (details below), once, before beginning the confidence tasks, and again after the first test block.

The order of the two sets of words was counterbalanced across participants. Thus, Exp 2 also had eight groups of participants (Supplementary Figure 1).

Tasks

Perception and memory tasks

The tasks were embedded in a gamified environment. Participants were informed they were helping people of Fruitville whose livelihood depends on harvesting and packaging fruits. The perception task required participants to decide which of two types of berries – raspberries or blackberries – were more numerous, to aid the farmers in deciding which berry to harvest. The memory task required participants to decide which fruit was present in a box of fruits that was recently opened in front of them, to help the fruit packers correctly label the contents of that box. Participants were also asked to report their confidence in their choices and told that their reported confidence would help people of Fruitville adjust how much to rely on their advice. To provide a natural rationale for the intermittent feedback schedule, participants were instructed that occasionally Fruitville would hire an "auditor" who would evaluate participants' choices and give them feedback as to whether their choice was correct or incorrect. The stimulus background was a cartoon-like nature scene and participants were instructed, in both tasks, to attend to a green bush in the centre within which all stimuli would appear.

Figures 1B and 1C illustrate a one trial sequence of each of the perception and memory tasks respectively. In Exp 1, each perception or memory task block consisted of 40 trials. In Exp 2, perception and memory task blocks also consisted of 40 trials except for the two test blocks, which consisted of 20 trials (we reduced the trial number in Exp 2 after observing any cross-task transfer of feedback to confidence in Exp 1 was largely restricted to the first half of each test block).

The perception task comprised a density estimation task, as used in previous psychophysical studies of perceptual metacognition[2,43]. On each trial, after initially attending to the blank central bush for a duration of 600—750 ms, 121 berry stimuli appeared at random non-

overlapping locations within the area of the central bush. Some of the berries were red (raspberries) while others were dark purple (blackberries). The berries appeared on the screen for 1000 ms within which time they were randomly replotted within the bush every 250 ms to make the stimulus dynamic and engaging and preclude explicit counting strategies. One of the two types of berries was always greater in number. 200 ms after stimulus offset, participants were shown one of each type of berry and asked to choose by pressing the left or the right arrow key, which they thought were presented in greater numbers. Once a choice was made, the chosen berry increased in size and a vertical slider bar appeared in the centre of the screen. Participants were then instructed to "Click/Drag the slider to select confidence in your choice" with their mouse. After the appearance of the slider, and before selection of a confidence value, participants had the option to change their berry choice by pressing the opposite arrow key. Once the slider was clicked, 5 confidence markers (none, low, medium, high, full) appeared to the left of the slider along with a continuous percentage between 0—100% on the right indicating their selected confidence. After reporting their confidence, participants were asked to end the trial by pressing one of the arrow keys. On selected trials of the intervention blocks, participants then saw a message notifying them that, "The auditor of Fruitville is here to evaluate your response. Check how you performed." Participants then pressed a button and were informed if they were Correct or Incorrect. The correct and incorrect feedback text was accompanied by a smiley or slightly frowning face respectively, together with text randomly chosen from one of three (positive or negative) feedback messages (see Supplementary Material). As a measure of global SPEs, at the end of the block, participants were shown a slider and asked to indicate how many trials they believed they answered correctly on that block.

The memory task was a visual working memory task. Each trial began with participants attending to the central bush for a variable duration of 500–600 ms, followed by the presentation of several different fruits (between 1–12 fruits) within the area of the bush for 1500 ms. After stimulus offset, there was a delay of 1000 ms following which participants were then shown two fruit options, one of which was present in the previous stimulus set and one that was not. Participants were asked to indicate the fruit that was in the previous set. They

then reported their local confidence, received feedback and provided global SPEs as described above for the perception task.

The difference between the number of raspberries and blackberries for the perception task, and the number of fruits displayed for the memory task (the set size), were each adjusted from trial to trial using a 1-up-2-down staircase. We calculated the probability of giving positive and negative feedback on intervention blocks based on the expected ~71% accuracy rate associated with such a staircase. Out of the 40 trials, we aimed to provide feedback on 9 correct trials and 1 incorrect trial on positive feedback blocks and 9 incorrect trials and 1 correct trial on negative feedback blocks. Feedback for each trial was thus delivered with the probabilities of .3169 (= 9/(40*.71); correct trial) and .0862 (= 1/(40*(1 – .71)); incorrect trial) on positive feedback blocks and .0352 (correct trial), and .7759 (incorrect trial) on negative feedback blocks.

Self-referential encoding task

Participants also completed a self-referential encoding task (SRET). Exp 1 involved a single instance of the SRET presented before the perception and memory task blocks. Here, participants were asked to self-endorse 42 words (20 pleasant or positive affect adjectives, 20 unpleasant or negative affect adjectives and 2 catch words; words from[44]; Supplementary Figure 9). Words were shown on the screen (in random order) for 1 second, and participants were asked, "Does this word describe you?" They then selected one of two options, "Yes" or "No." After responding to all the words, participants were asked further questions about their Prolific ID, gender, and how many online studies they had completed in the last 24 hours and last month. This was followed by a surprise memory test where they were asked to recall as many words as possible from the self-referential encoding task phase (data not analysed here).

In Exp 2, participants were administered the SRET at two time points – before and after the performance feedback intervention. Positive and negative affect words from Exp 1 were split into two sets of 10 positive, 10 negative and 1 catch words, with one set presented at each time point. Words were split based on how well they predicted individual global self-performance

estimates on perception and memory tasks during the baseline trials of Exp 1, such that both set of words had roughly similar relationships with global confidence (Supplementary Figure 9). With this design we aimed to control for any non-specific differences in the sensitivity of the two word sets to changes in global confidence. For the SRET in Exp 2, we sought a continuous answer to the question "How much does this word describe you?" Participants responded using a slider with 5 equidistant markers, "Not at all," "A little", "Somewhat", "A good amount" and "Very much." Each word was presented for 1.5 sec and responses were self-paced.

Procedure

The entire study was programmed using the Phaser 3 game framework for JavaScript to provide a gamified look and feel and hosted on the online platform Pavlovia. Participants from Prolific were given the study link (run.pavlovia.org/sucharit/fruitville/) where they were first required to read and agree to a Study Information and Consent Form (both approved by the UCL Research Ethics Committee; approval number 21029/001) before proceeding. Each participant then proceeded through multiple self-paced phases of the study.

The first phase comprised the self-referential encoding task. This included instructions about how to perform the task along with four practice words, the 42 test words (21 in Exp 2), and the surprise memory test (only in Exp 1). The next phase comprised the Fruitville game. This commenced with extensive step-by-step instructions on how to perform one or both tasks (depending on the participant's group) along with a description of the game scenario and 5 training trials. Participants were given the option to repeat the instructions and training trials as many times as they wished. Instructions for each task were followed by 50 practice trials of that particular task, which also served the purpose of initiating the staircase. From the practice trials, we used the mean value of the last 5 reversals of staircase level as the starting level of task difficulty used for the main part of the study, which in turn involved participants performing six task blocks. Figure 1B and Supplementary Figure 1 shows sequences of the

blocks for different groups in Exp 1 and Exp 2 respectively. In Exp 2, the second block of SRET occurred after block 3 of the perception/memory task.

In the final phase, participants completed several mental health questionnaires. For Exp 1 this included the PHQ-9 for depression[45], GAD-7 for general anxiety[46] and mini-SPIN for social phobia[47]. For Exp 2, this included a set of 71 questions[28] designed to allow efficient determination of scores along three transdiagnostic mental health axes[27] – anxious-depression (AD), compulsivity and intrusive thought (CIT), and social withdrawal (SW).

Across the entire study, there were four 'catch' questions designed to detect participants who may not have been paying sufficient attention. Two catch questions were included during the self-referential encoding task, and two were included within the mental health questionnaires.

Statistical analyses

Model-free data analysis and figure generation was performed in R[48]. One set of analyses tested effects of performance feedback interventions on subsequent global SPEs and local confidence. For these analyses, we baseline-corrected global SPEs by subtracting baseline block SPEs, and local confidence by subtracting baseline block average local confidence. Consequently, we use the label SPE-b in what follows to indicate baseline-corrected global confidence. As these analyses were performed at a within-subject level, we used linear mixed regression models (LMM) implemented by the *lmer* function in the *lmerTest* package (version 3.1-3). All LMMs included a random intercept for individual participants. When modelling global SPEs, we included random intercepts for group and run number as well as main effects of mean accuracy, mean confidence, mean reaction times within the relevant block (for each task, reaction times > 3 times interquartile range away from the median were excluded and then z-scored). When modelling local confidence, we also included random intercepts for trial number. Random intercepts for group, run number and trial number were dropped if they did not explain sufficient variance in the data, as indicated by singular model fits. For the LMMs, *p* values were calculated by determining degrees of freedom using the Satterthwaite method. Any 2- or 3-way interactions also included main effects of all terms in those interactions. Any

2- or 3-way interactions that were not significant at an alpha = .05 were excluded from models and models were re-evaluated without those interactions (by first reducing 3-way and then 2-way interactions). All tests were two-tailed.

For between-participant analyses for relating local/global confidence on baseline blocks, mental health scores and SRET self-endorsements, we used the *lm* function in the *stats* package (version 4.1.0). For mediation analyses, we used the *mediation* package (version 4.5.0) in R. Marginal effects obtained from regression models were plotted using the *plot_model* function of the *sjPlot* package (version 2.8.11).

Computational models

We modelled the dynamics of global SPEs by building on a model developed by Rouault et al[9]. This model maintains Beta distributions over expected success for each task as a proxy for global self-performance estimates. A Beta distribution is characterised by *a* and *b* parameters, with higher values of *a* tending to "pull" the distribution towards 1 (higher SPE) and higher values of *b* pulling it towards 0 (lower SPE). Higher values of both *a* and *b* result in higher precision, and more certainty around a particular SPE. In Rouault et al's[9] model, for trials with explicit feedback, Beta parameters are updated trial-by-trial as follows:

$$a_{t+1} = \begin{cases} a_t + 1 & if\ correct \\ a_t & if\ incorrect \end{cases}$$

$$b_{t+1} = \begin{cases} b_t & if\ correct \\ b_t + 1 & if\ incorrect \end{cases}$$

When feedback is provided on all trials, this algorithm naturally leads to mean of the Beta distribution (the mean SPE) converging on one's true underlying probability correct, with increasing precision as more data (trials) are acquired. In contrast, when participants do not get feedback, the best information they have about their performance is their local confidence estimate. Rouault et al. proposed local confidence can be used as a proxy of the probability of

a correct response on a given trial. Thus, in the absence of feedback, the a and b parameters are updated by reported confidence (normalised to 0—1):

$$a_{t+1} = a_t + conf_t$$
$$b_{t+1} = b_t + (1 - conf_t)$$

In the present study, we were interested in a potential asymmetry in forming SPEs from positive and negative feedback, and high and low confidence, and whether such asymmetries relate to anxious-depression (AD) symptoms. Thus, we modified the above model by introducing parameters that controlled the asymmetry in updating global SPEs from 1) positive and negative feedback (ΔLR_f), and 2) high and low confidence (ΔLR_c), as follows.

In the presence of feedback:

$$a_{t+1} = \begin{cases} a_t + (1 + \Delta LR_f) & if\ correct \\ a_t & if\ incorrect \end{cases}$$
$$b_{t+1} = \begin{cases} b_t & if\ correct \\ b_t + (1 - \Delta LR_f) & if\ incorrect \end{cases}$$

In the absence of feedback:

$$a_{t+1} = a_t + conf_t * (1 + \Delta LR_c)$$
$$b_{t+1} = b_t + (1 - conf_t) * (1 - \Delta LR_c)$$

In these equations, the ΔLR parameters control a boost in learning rate on correct/high confidence trials relative to a boost in learning rate on incorrect/low confidence trials. Thus, if participants learn equally from both trial types, ΔLR should be zero. Moreover, the asymmetry parameters may be similar or different for the two tasks (perception and memory). We compared different versions of the model in terms of goodness of fit (using Deviation Information Criteria; DIC): 1) ΔLR = 0 – global SPE is formed equally from positive and negative feedback, and high and low local confidence, 2) ΔLR ≠ 0, but is the same for two tasks, and is applied equally to both feedback and local confidence (1 parameter), 3) ΔLR ≠ 0, and is the same for two tasks but differs for feedback and local confidence (2 parameters), 3) ΔLR ≠ 0, and is different for the two tasks but the same for feedback and local confidence

(2 parameters), and 4): ΔLR ≠ 0, and differs both across tasks and between feedback and local confidence (4 parameters).

For model fitting, the initial values of *a* and *b* were determined by allowing free parameters for the mean (μ_0) and variance (v_0) of the Beta distributions for the two tasks. Beta parameters *a* and *b* are defined by mean and variance as follows:

$$a_0 = \frac{\mu_0}{v_0} * (\mu_0 - {\mu_0}^2 - v_0)$$

$$b_0 = \frac{(1 - \mu_0)}{v_0} * (\mu_0 - {\mu_0}^2 - v_0)$$

The participant's reported global SPE at the end of a given block B is used to fit Beta parameters on the final trial *n* of each block as follows:

$$SPE_B \sim Beta(a_n^B, b_n^B)$$

The best fitting model from those listed above served as the "no distortion" model used to investigate the influences of AD symptoms on learning (D0; see Supplementary Material), which we then extended to test for the presence of several potential cognitive distortions impacting global confidence formation in high AD individuals. When forming global confidence, we considered that high AD individuals could D1) have higher sensitivity to negative vs. positive feedback trials, D2) have higher sensitivity to low vs. high local confidence, or D3) simply be biased to report lower SPEs at the end of the block without differing in learning rates. We adjudicated between these potential mechanisms by regressing individual mental health scores (MH_i) upon learning rate asymmetry parameters ΔLR_f (for D1) and ΔLR_c (D2), and global SPE values (D3), and estimated the corresponding regression slopes β_f, β_c and β_a respectively, as follows:

$$\Delta LR_f = \Delta LR_{f0} + \beta_f * MH_i$$

$$\Delta LR_c = \Delta LR_{c0} + \beta_c * MH_i$$

$$SPE_B = \mu_n^B + \beta_a * MH_i$$

Here, μ_n denotes the Beta mean computed after updating the last trial:

$$\mu_n^B = \frac{a_n^B}{a_n^B + b_n^B}$$

The regression models deviate from those noted in the preregistration document in two ways: 1) At the time of preregistration we did not consider the non-learning bias model, D3, as an alternative to our key hypotheses regarding feedback and confidence learning distortions. 2) In our preregistered methods, we only considered the regression slope (β) terms, and erroneously omitted the intercepts terms for Models D1 and D2 (i.e., ΔLR_{f0} and ΔLR_{c0}). Simulations showed that the absence of an intercept term can introduce artificially significant effects that load on the slope term. Indeed, after including the intercept terms, we no longer observed $\beta_f < 0$ for Exp 1 as stated in preregistered Hypothesis 2. However, consistent with the same preregistered hypothesis, we still observed $\beta_c < 0$.

As anxious-depression symptoms are known to have a significant relationship with local confidence[2], we sought to avoid this relationship confounding interpretation of biases in global SPE formation by *z*-scoring local confidence within each participant prior to analysis.

Models were fit using MCMC sampling (3 chains of 2000 samples with 1000 burn-in samples) implemented by the JAGS toolbox (version 3.4.1[49]) in MATLAB (Mathworks Inc.; 2022b). All free parameters were inferred at the group level. The three regression slope parameters – β_f, β_c and β_a – were fully recoverable in simulation (Figure 3A and Supplementary Figure 5). All models were run using uninformative priors.

Acknowledgments

SK was supported by a grant from Koa Health. SMF is a CIFAR Fellow in the Brain, Mind and Consciousness Program, and funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust. QJMH acknowledges support by the UCLH NIHR BRC and grant funding from the Wellcome Trust (221826/Z/20/Z) and Carigest S.A.. The Max Planck-UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by University College London and the Max Planck Society. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any author-accepted manuscript version arising from this submission.

Data availability

All data is publicly available at https://github.com/sucharitk/confidence-distortion-AD/

Code availability

All code for stimulus presentation and data analysis is publicly available at https://github.com/sucharitk/confidence-distortion-AD/

Competing interests

This research was funded by Koa Health. The funder did not participate in conceptualization, design, data collection, analysis, or preparation of the manuscript. However, the funder was informed of the concept and design prior to data collection. They were also provided an earlier draft of this manuscript before it was submitted for publication. QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience.

References

1. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**, 191 (1977).
2. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biol. Psychiatry* **84**, 443–451 (2018).
3. Seow, T. X. F. & Gillan, C. M. Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Sci. Rep.* **10**, 2883 (2020).
4. Benwell, C. S. Y., Mohr, G., Wallberg, J., Kouadio, A. & Ince, R. A. A. Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *Npj Ment. Health Res.* **1**, 1–17 (2022).
5. Hoven, M., Luigjes, J., Denys, D., Rouault, M. & van Holst, R. J. How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nat. Ment. Health* **1**, 337–345 (2023).
6. Seow, T. X. F., Rouault, M., Gillan, C. M. & Fleming, S. M. How local and global metacognition shape mental health. *Biol. Psychiatry* (2021) doi:10.1016/j.biopsych.2021.05.013.
7. Fox, C. A. *et al.* Metacognition in anxious-depression is state-dependent: an observational treatment study. Preprint at https://doi.org/10.31234/osf.io/uk7hr (2023).
8. Wittmann, M. K. *et al.* Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. *Neuron* **91**, 482–493 (2016).

9. Rouault, M., Dayan, P. & Fleming, S. M. Forming global estimates of self-performance from local confidence. *Nat. Commun.* **10**, 1141 (2019).
10. Rouault, M., Will, G.-J., Fleming, S. M. & Dolan, R. J. Low self-esteem and the formation of global self-performance estimates in emerging adulthood. *Transl. Psychiatry* **12**, 1–10 (2022).
11. Allen, M. *et al.* Unexpected arousal modulates the influence of sensory noise on confidence. *Elife* **5**, e18103 (2016).
12. Koellinger, P. & Treffers, T. Joy Leads to Overconfidence, and a Simple Countermeasure. *PLOS ONE* **10**, e0143263 (2015).
13. Lebreton, M. *et al.* Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Sci. Adv.* **4**, eaaq0668 (2018).
14. Marcke, H. V., Denmat, P. L., Verguts, T. & Desender, K. Manipulating prior beliefs causally induces under- and overconfidence. 2022.03.01.482511 Preprint at https://doi.org/10.1101/2022.03.01.482511 (2022).
15. Siedlecka, M. The immediate effects of accuracy feedback on metacognitive processing. (2022).
16. Daniel, R. & Pollmann, S. Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* **59**, 3457–3467 (2012).
17. Guggenmos, M., Wilbertz, G., Hebart, M. N. & Sterzer, P. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* **5**, e13388 (2016).
18. Alloy, L. B. & Ahrens, A. H. Depression and pessimism for the future: Biased use of statistically relevant information in predictions for self versus others. *J. Pers. Soc. Psychol.* **52**, 366–378 (1987).

19. Beck, A. T. *Depression: Clinical, experimental, and theoretical aspects*. (University of Pennsylvania Press, 1967).
20. Everaert, J., Podina, I. R. & Koster, E. H. W. A comprehensive meta-analysis of interpretation biases in depression. *Clin. Psychol. Rev.* **58**, 33–48 (2017).
21. 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).
22. 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).
23. Garrett, N. *et al.* Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Front. Hum. Neurosci.* **8**, (2014).
24. Katyal, S. & Fleming, S. Construct validity in metacognition research: balancing the tightrope between rigor of measurement and breadth of construct. Preprint at https://doi.org/10.31234/osf.io/etjqh (2023).
25. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016).
26. Lee, A. L. F., de Gardelle, V. & Mamassian, P. Global visual confidence. *Psychon. Bull. Rev.* **28**, 1233–1242 (2021).
27. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).

28. Hopkins, A. K., Gillan, C., Roiser, J., Wise, T. & Sidarus, N. Optimising the measurement of anxious-depressive, compulsivity and intrusive thought and social withdrawal transdiagnostic symptom dimensions. Preprint at https://doi.org/10.31234/osf.io/q83sh (2022).
29. Dainer-Best, J., Lee, H. Y., Shumake, J. D., Yeager, D. S. & Beevers, C. G. Determining optimal parameters of the Self Referent Encoding Task: A large-scale examination of self-referent cognition and depression. *Psychol. Assess.* **30**, 1527–1540 (2018).
30. Derry, P. A. & Kuiper, N. A. Schematic processing and self-reference in clinical depression. *J. Abnorm. Psychol.* **90**, 286 (1981).
31. Kuiper, N. A. & Derry, P. A. Depressed and nondepressed content self-reference in mild depressives. *J. Pers.* **50**, 67–80 (1982).
32. Shestyuk, A. Y. & Deldin, P. J. Automatic and Strategic Representation of the Self in Major Depression: Trait and State Abnormalities. *Am. J. Psychiatry* **167**, 536–544 (2010).
33. Wells, A. *Metacognitive therapy for anxiety and depression*. (Guilford press, 2011).
34. Salem-Garcia, N., Palminteri, S. & Lebreton, M. Linking confidence biases to reinforcement-learning processes. *Psychol. Rev.* No Pagination Specified-No Pagination Specified (2023) doi:10.1037/rev0000424.
35. Ciranka, S. *et al.* Asymmetric reinforcement learning facilitates human inference of transitive relations. *Nat. Hum. Behav.* **6**, 555–564 (2022).
36. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science* **306**, 1940–1943 (2004).

37. Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S. & Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 1–9 (2017).
38. Rosenbaum, G. M., Grassie, H. L. & Hartley, C. A. Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. *eLife* **11**, e64620 (2022).
39. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* **14**, 1475–1479 (2011).
40. Rouhani, N. & Niv, Y. Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology (Berl.)* **236**, 2425–2435 (2019).
41. Hoven, M. *et al.* Abnormalities of confidence in psychiatry: an overview and future perspectives. *Transl. Psychiatry* **9**, 1–18 (2019).
42. Moritz, S. *et al.* Sowing the seeds of doubt: a narrative review on metacognitive training in schizophrenia. *Clin. Psychol. Rev.* **34**, 358–366 (2014).
43. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).
44. Katyal, S., Hajcak, G., Flora, T., Bartlett, A. & Goldin, P. Event-related potential and behavioural differences in affective self-referential processing in long-term meditators versus controls. *Cogn. Affect. Behav. Neurosci.* **20**, 326–339 (2020).
45. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9. *J. Gen. Intern. Med.* **16**, 606–613 (2001).

46. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Arch. Intern. Med.* **166**, 1092–1097 (2006).
47. Wiltink, J. *et al.* Mini - social phobia inventory (mini-SPIN): psychometric properties and population based norms of the German version. *BMC Psychiatry* **17**, 377 (2017).
48. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4. *ArXiv14065823 Stat* (2014).
49. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. in *Proceedings of the 3rd international workshop on distributed statistical computing* vol. 124 1–10 (Vienna, Austria., 2003).
50. Kumle, L., Võ, M. L.-H. & Draschkow, D. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behav. Res. Methods* **53**, 2528–2543 (2021).

Supplementary Material

Methods

Participants

Recruitment and group allocation

In Exp 1 our goal was to have at least 25 participants per group. To achieve this, we first randomly allocated eight participants to each group to ensure the task ran successfully from beginning to end for all eight groups. We then allocated 192 participants into eight equal groups using a pre-randomised list. After exclusions, groups that lacked the minimum of 25 participants were then randomised by recruiting participants in two iterations till we reached our criterion. In all, 390 participants started Exp 1 and 314 completed it.

For our replication study, Exp 2, we determined a target sample size by calculating the number of participants that would be needed to achieve 90% power for the interaction between proportion of negative feedback trials and depression scores (PHQ-9) regressed upon intervention-block self-performance estimates. The power calculation was performed in R using the *mixedpower* package (version 0.1.0) that allows power calculations for mixed effects models [50]. Accounting for a 25% exclusion rate based on Exp 1, we estimated a sample size of 460, which we preregistered (osf.io/7xfqw). When recruiting participants, we again randomised the first eight participants to ensure the task ran for all eight groups. Out of the remaining participants, we allocated group numbers for the first 448 participants using a pre-randomised list comprising eight groups equally. As some participants would start the task, receive a group allocation but subsequently drop-out, any remaining participants were then allocated to one of the eight groups randomly till we

reached our predetermined sample size. A total of 591 participants started Exp 2 and 460 completed it.

Exclusion

Participants were excluded from all analyses if they missed one of the four "catch" questions (two during the self-referential encoding task and two during the mental health questionnaires).

For analyses involving local confidence and global self-performance estimates, we also excluded participants, 1) whose performance was outside the interval (.60 .85] on any one of the 6 task blocks, 2) who did not exhibit sufficient variability in trial-by-trial confidence ratings defined as having <.05 SD across trials for each task (on a continuous confidence scale of 0—1), and 3) who did not have stable behavioural staircases in the perception/memory tasks as assessed visually.

Final sample for confidence analyses

Table S1. Number of participants in the two experiments for each of the eight groups

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Exp 1	30	26	27	30	29	27	29	32
Exp 2	36	40	32	36	32	30	32	40

The three performance-/confidence-based exclusion criteria used in the confidence analysis above were not relevant (and overly stringent) for analyses involving the self-referential encoding task (SRET). Instead, for the SRET, we used a less stringent criterion of excluding participants whose mean accuracy was below .6 across all blocks on average (and were thus presumably not paying sufficient attention to the experiment). Additionally, for this task we found that some subjects had extremely long RTs ranging from 10 seconds to several minutes. Such participants were presumably also not doing the task sincerely and would especially add noise to the pre-post feedback intervention changes in self-beliefs

we expected to test in Exp 2. We thus removed participants for whom any word RT was greater than 5 IQRs from the median (~4.9 sec). The results were not substantially impacted by the specific criterion – similar results were obtained, for example, if we used stricter (e.g., >3 IQRs or ~3.2 sec) or more relaxed (e.g., >7 IQRs or ~6.4 sec) exclusion criteria. Final sample sizes for SRET were N = 300 for Exp 1 and N = 335 for Exp 2.

Tasks

Perception and memory tasks

On the perception and memory tasks, feedback on correct and incorrect trials was accompanied by a randomly chosen message from one of three messages each.
Messages for Correct trials:

1) “Great going!”
2) "The residents of Fruitville thank you for your help!”
3) "You are getting good at this!”

Messages for Incorrect trials:

1) "You chose the wrong option!",
2) "Should have chosen the other one!",
3) "The residents of Fruitville chose the wrong fruit based on your suggestion!"

For the perception and memory tasks, we excluded trials where RTs were more than 3 IQR (inter-quartile range) away from the median RT evaluated separately for the two tasks.

Self-referential encoding tasks

For the self-referential encoding task, we found some subjects had extremely long RTs ranging from 10 seconds to several minutes. Such participants were presumably not doing the task sincerely and would especially add noise to the pre-post feedback intervention changes in self-beliefs we expected to test in Exp 2. We thus removed participants for whom any word RT was greater than 5 IQRs from the median. The results were not

substantially impacted by the specific criterion – similar results were obtained, for example, if we used a stricter exclusion criterion of >3 IQRs from the median.

Results

Domain-general transfer of feedback to confidence

We tested if the impact of feedback on confidence transferred across distinct task domains. In Exp 1 (Supplementary Figures 4A and 4B), we observed a significant 3-way interaction between *Feedback type* (on the preceding intervention block), *Task* and *Transfer type* (within domain, across domains) upon test-block local confidence ($t(16772) = 3.28$, $p = .001$). Test-block local confidence was significantly higher following positive compared to negative feedback when both intervention and test tasks were the same, i.e., within-domain transfer (perception-to-perception transfer: positive – negative = 0.021 ± .003, $t(16772) = 6.30$, $p < .0001$, Supplementary Figure 4A top-left; memory-to-memory transfer: positive – negative = 0.010 ± .003, $t(16775) = 2.86$, $p = .004$, bottom-right) confirming an enduring effect of feedback on subsequent confidence within the same domain. Importantly, we also observed cross-domain transfer of feedback to local confidence. Local confidence on *memory* test blocks was significantly higher following positive compared to negative feedback *perception* blocks (positive – negative = 0.009 ±.003; $t(16779) = 2.57$, $p = .010$; Supplementary Figure 4A bottom-left). However, this was not the case for local confidence on perception test blocks following memory intervention blocks (positive – negative = –0.002 +.003; $t(16774) = –.59$, $p = .56$; Supplementary Figure 4A top-right). Generally, within-domain transfer of feedback to test-block confidence was stronger ($t(16774) = 6.46$, $p < .0001$) than cross-domain transfer ($t(16778) = 1.45$, $p = .15$). Moreover, we found within-domain (but not cross-domain transfer) to be mediated by global SPEs, which suggests that global SPEs may (at least to some degree) act as summary metacognitive beliefs about one's performance carried forward to future instances of tasks (Supplementary Material).

Exp 2 only included measures of cross-domain transfer, with intervention blocks always being followed by test blocks of a different type. Here, we observed a significant main effect of *Feedback type* (positive – negative = –.012 ± .004; $t(10304) = –2.87$, $p = .004$; Supplementary Figures 4C and 4D) on subsequent confidence, in the absence of any interaction with *Task* ($t(10318) = –1.07$, $p = .28$), indicating a cross-domain effect of feedback

on local confidence was present both when the intervention block was perception, and the test block memory, and when the intervention block was memory and the test block perception. Taken together, the two experiments demonstrate that feedback interventions impacted subsequent local confidence in a domain-general manner.

Mediation of feedback transfer to local confidence by SPE

A hierarchical model of metacognition[6] suggests that summary metacognitive beliefs are formed from local metacognitive evaluations, with more global estimates then used as priors for metacognitive evaluations across other domains or timescales. We tested if global SPEs may act as low-dimensional summary statistics mediating the effect of feedback interventions on local confidence in subsequent test blocks. In Exp 1, a mediation analysis (Supplementary Figures 10A and 10C) confirmed that intervention-block SPE fully mediated a positive relationship between positive feedback and test-block confidence (mediated effect = 0.03, 95% CI = [.004 .06], p = .016) and negative relationship between negative feedback and test-block confidence (mediated effect = –0.06, CI = [–.12 –.01], p = .017).

Unlike in Exp 1, in Exp 2 we did not observe any significant mediation of positive and negative feedback by intervention-block SPE in predicting test confidence (both > .9), which we suspect was due to the relatively weak overall effects of confidence transfer in Exp 2 compared to Exp 1 (due to Exp 2 only including cross-domain transfer conditions). Post-hoc analyses confirmed this intuition. We found that when re-analysing Exp 1, a mediation effect was statistically reliable only when transferring to the same task (positive feedback: mediated effect = 0.04, 95% CI = [.003 .10], p = .025; negative feedback: mediated effect = –0.11, 95% CI = [–.21 –.02], p = .011) with no such mediation effect observed when transferring to the opposite task (positive feedback: mediated effect = 0.01, 95% CI = [–.02 .05], p = .38; negative feedback: mediated effect = –0.03, 95% CI = [–.09 .03], p = .35).

Overall, these analyses are consistent with global confidence acting as a mediator of changes in local confidence, although this may be limited to within-domain transfer.

Longevity of impact of feedback on SPEs

In Exp 1, we observed that feedback effects on local confidence dissipated towards the second half of the test block (beyond ~20 trials). However, it is possible that a lower-dimensional summary of performance in the test block – an end-of-block SPE – may inherit some of the feedback effect, leading to slower-timescale dynamics in global confidence. As with local confidence on test blocks, this shift in global confidence could be domain-specific or domain-general.

For Exp 1, we first performed a factorial analysis of test-block SPEs, which showed a significant main effect of *Feedback type* (positive – negative = .030 ± .009, $t(222) = 3.45$, $p = .0007$), in the absence of 2- or 3-way interactions with *Task* and *Transfer type* (all $p > .2$), indicating that intervention-block feedback effects continue to exert effects on more distant test-block SPEs in a domain-general fashion (Supplementary Figure 11A). We also performed a mediation analysis to test if intervention-block SPE mediated the transfer of feedback to test-block SPE. Here again, there was a significant mediation of the effect of positive (mediated effect = 0.06, CI = [.03 .11], $p < .0001$) and negative (mediated effect = –0.15, 95% CI = [–.23 –.08], $p < .0001$) feedback on test-block SPE by intervention-block SPE (Supplementary Figures 10B and 10D).

For Exp 2, we did not find a significant main effect of *Feedback type* (positive – negative = .011 ± .009, $t(275) = 1.28$, $p = .20$), or an interaction with *Task* ($t(277) = –1.52$, $p = .13$) on test-block SPEs, although the sign of the main effect was in the expected direction (positive > negative; Supplementary Figure 11B). Exploratory analysis however did reveal that when only considering perception-to-memory transfer, there was a significant effect of positive > negative feedback on test SPE ($t(280) = 1.99$, $p = .048$). This ordering of transfer effect magnitude (perception-to-memory > memory-to-perception) was also observed for test-block SPEs in Exp 1 (Supplementary Figure 11A). Finally, there was also no significant

mediation of positive and negative feedback upon test-block SPEs by intervention-block SPEs (both $p > .8$).

These results suggest the possibility of feedback effects potentially affecting slower-timescale global confidence estimates beyond their immediate influence on local confidence, with such an effect being stronger within compared to across domains.

Choice of the "no distortion" model

Before fitting the different models of distortions in global SPEs with individual AD symptoms, we compared different accounts of group-level learning asymmetries to obtain a best fitting "no distortion" model. As in the model comparison in the main text, the learning asymmetries could be similar or different for feedback and confidence, and similar or different for the two tasks (perception/memory; see Methods). A comparison of DIC values revealed that in both Exp 1 and Exp 2 (Supplementary Figure 12), the best fitting model was one with separate, task-specific asymmetries in learning from feedback and confidence. This model provided better fits than other model variants. This model also provided good qualitative fits to the global SPE data across 6 blocks and 8 eight groups in both Exp 1 and Exp 2 (Supplementary Figure 13). Participants in both experiments consistently underestimated their true performance (~71% correct), a bias which was captured by the model. Interestingly, providing biased positive feedback propelled participants' SPEs towards values that matched their performance level.

Local and global confidence in relation to the CIT axis

Recent work has shown that while the CIT axis is positively associated to local confidence, it is negatively related to global confidence [5]. Such a contradictory finding could be explained by individuals scoring high on the CIT axis tending to overweigh low vs. high local confidence when forming global confidence. In other words, we might expect the β_c parameter to be more negative in individuals with higher CIT scores. When fitting our model to the Exp 2 data including all three transdiagnostic axes (instead of just the AD axis for our main hypothesis), this is indeed what we found – that β_c (99% HDI = [–.20 –.09])

was significantly negative for the CIT axis over and above the effect observed along the AD axis, which continued to be significantly negative (99% HDI = [–.10 –.01]; Supplementary Figure 14A). Model-free analyses again showed slightly elevated SPE values during feedback relative to baseline blocks with a reduction in SPEs during non-feedback blocks in individuals with high CIT axis scores (Supplementary Figures 14B—C). For completeness, we also report model-fit regression slopes for the SW axis, finding that β_c was significantly greater than 0 (99% HDI = [.07 .17]). Taken together, therefore, these findings indicate that greater scores on all 3 transdiagnostic axes are associated with greater sensitivity to lower rather than higher local confidence. Future studies specifically targeting different symptom axes are needed to understand potential differences in global confidence formation across dimensions[5].

Baseline correlations of affective self-endorsements with mental health and confidence

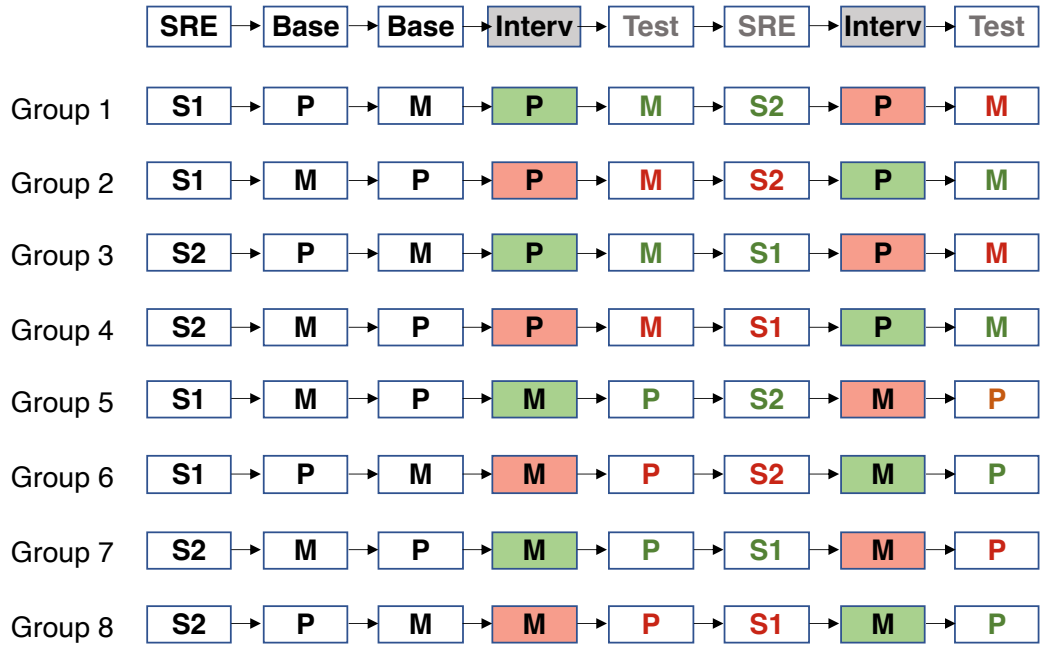
Replicating previous work, we observed a highly significant interaction between word valence and depression and anxiety scores upon self-endorsements in Exp 1 (PHQ, z = 34.19, p < .0001; GAD, z = 31.96, p < .0001; SPIN, z = 31.45, p < .0001), characterised by significant increases in self-endorsed negative words and significant decreases in self-endorsed positive words with AD symptoms (Supplementary Figure 7). In Exp 2, we found a similar interaction of word valence with the AD transdiagnostic axis (z = 20.14, p < .0001; Figure 5A), which was again strongly related both to an increase in negative self-endorsements (z = 12.91, p < .0001) and decrease in positive self-endorsements (z = –15.74, p < .0001). Similar interactions were also observed for the other two transdiagnostic axes (CIT: z = 6.04, p < .0001; SW: z = 4.51, p < .0001), with the CIT interaction specifically driven by a positive relationship with negative self-endorsements (Supplementary Figures 8A–B).

Preregistered SRET analysis

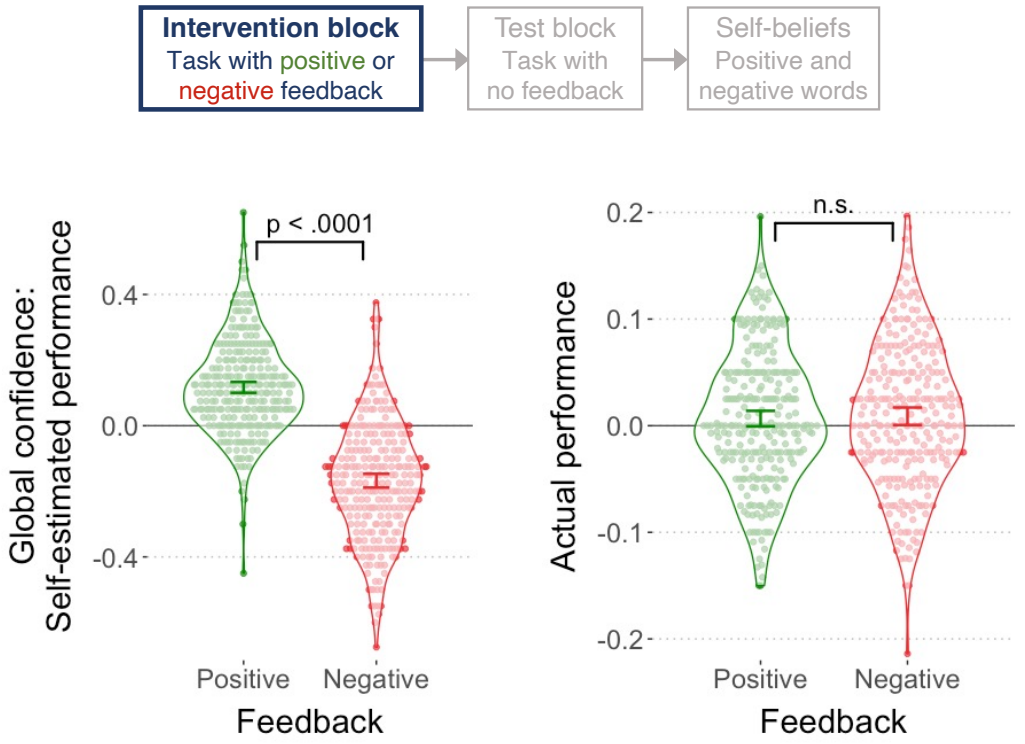
The analysis reported in Results slightly deviated from our preregistered analysis when assessing changes in self-beliefs with *Feedback type* (positive / negative) in Exp 2 (see

Methods). As per our preregistered analysis, we also regressed *Feedback type* and its interaction with *Task* upon the double difference of word valence (positive – negative self-endorsements) and timepoint (T2 – T1). We observed a significant main effect of *Feedback type* upon the double difference of self-endorsements ($t(262) = -2.05$, $p = .042$) in the expected direction in the absence of an interaction with *Task* ($t(260) = .76$, $p = .45$). Thus, as hypothesised, the difference between positive vs. negative self-endorsements between pre- and post-feedback blocks was greater for positive compared to negative feedback.

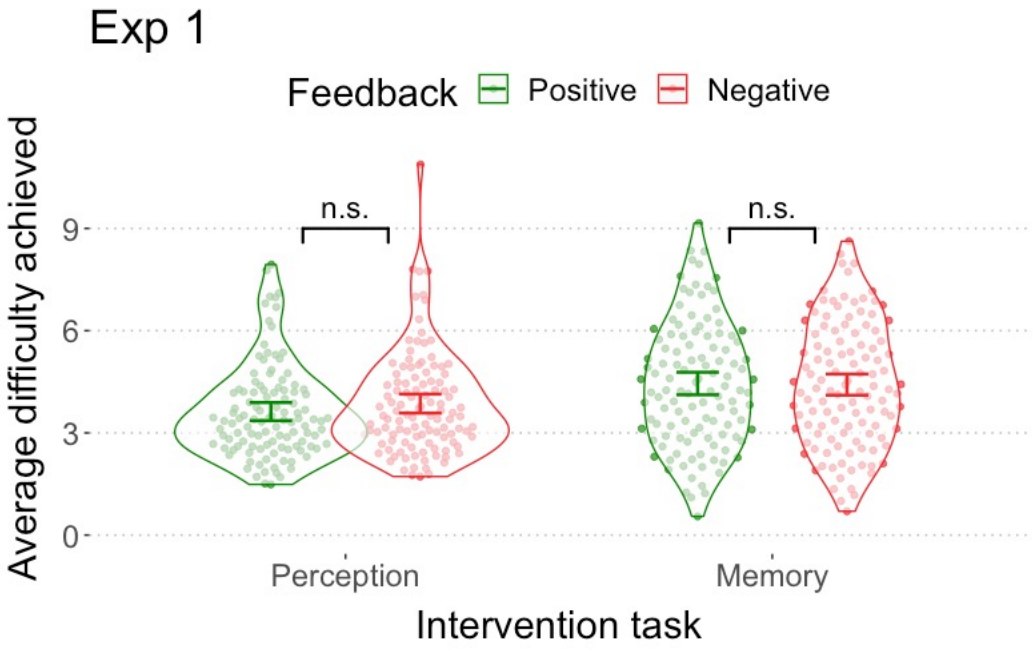
Supplementary Figures

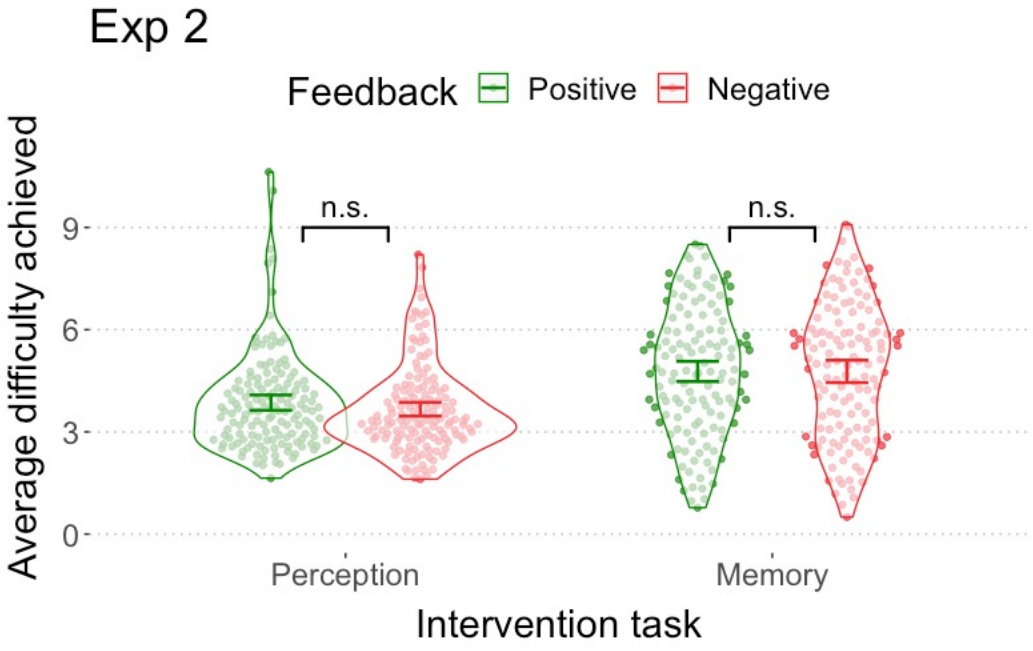


Supplementary Figure 1. Eight possible groups of participants in Exp 2 with the order in which they performed the three tasks, perception (P), memory (M) and self-referential encoding (SRE; with one of two sets of words, S1 or S2) task. The study started with one block of the SRE task. This was be followed by two baseline blocks of the perception and memory task (order randomised across participants). Next, on the intervention block (Interv) they were provided either more frequent positive (green) or negative (red) feedback. Next, they underwent a test block of the opposite task from the intervention block. Next, they were administered the SRE task with the other set of words from the baseline SRE block. Finally, they performed another set of intervention and test blocks where feedback on the intervention block was opposite of the feedback (more negative or more positive) given on the first intervention block.

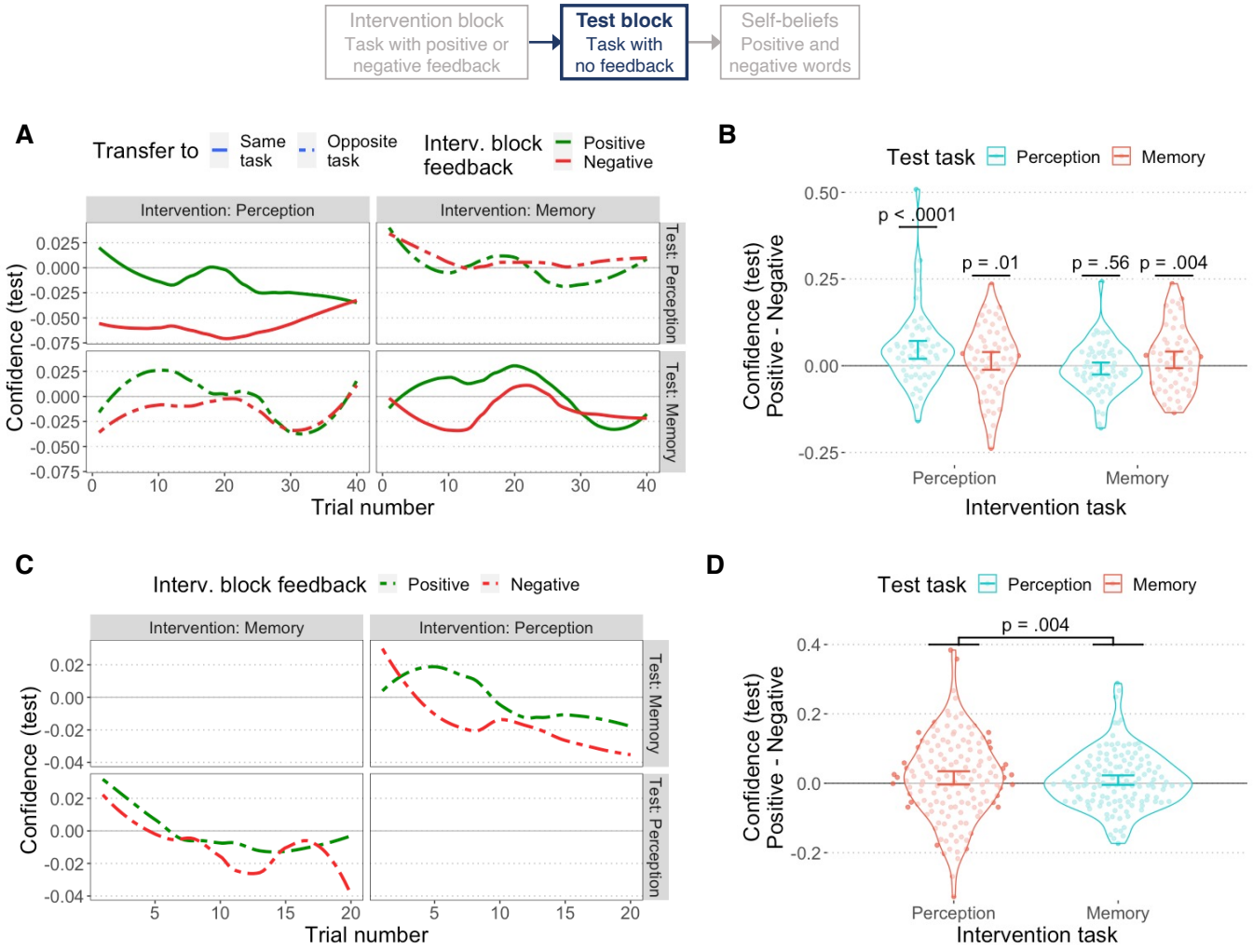


Supplementary Figure 2. The effect of positive (green) and negative (red) feedback on SPE of the intervention block for the two tasks in Exp 2 (baseline block SPE values subtracted for each individual). Each dot is individual participant data. Error bars are 95% bootstrapped confidence intervals.

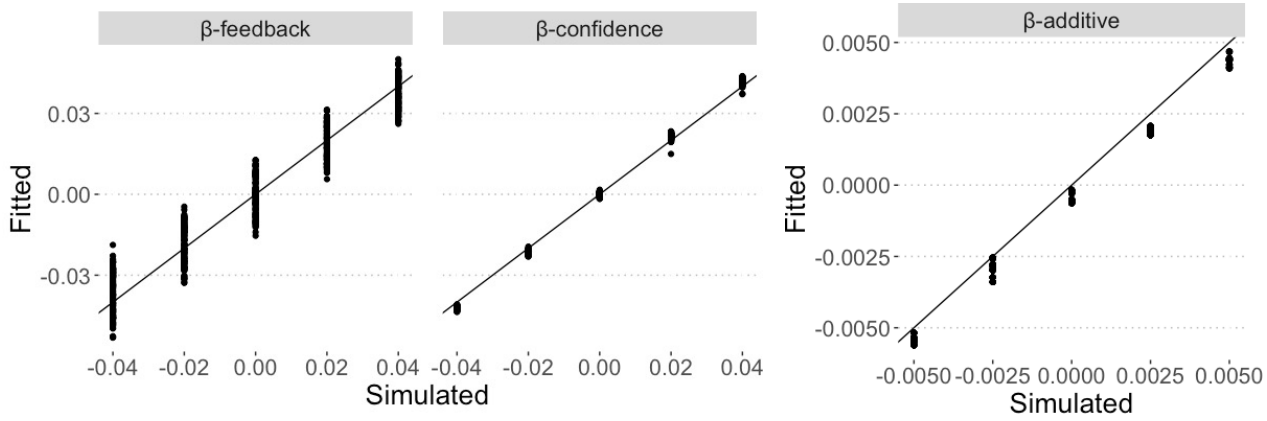




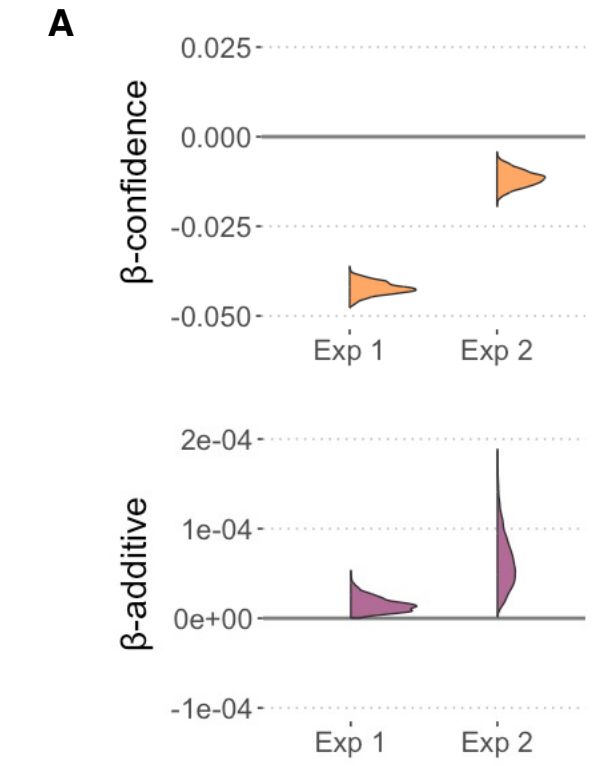
Supplementary Figure 3. The effect of positive (green) and negative (red) feedback on average difficulty achieved (mean staircase level) of the intervention block for the two tasks in Exp 1 (top) and Exp 2 (bottom) plotted separately for the two tasks Each dot is individual participant data. Error bars are bootstrapped 95% confidence intervals.



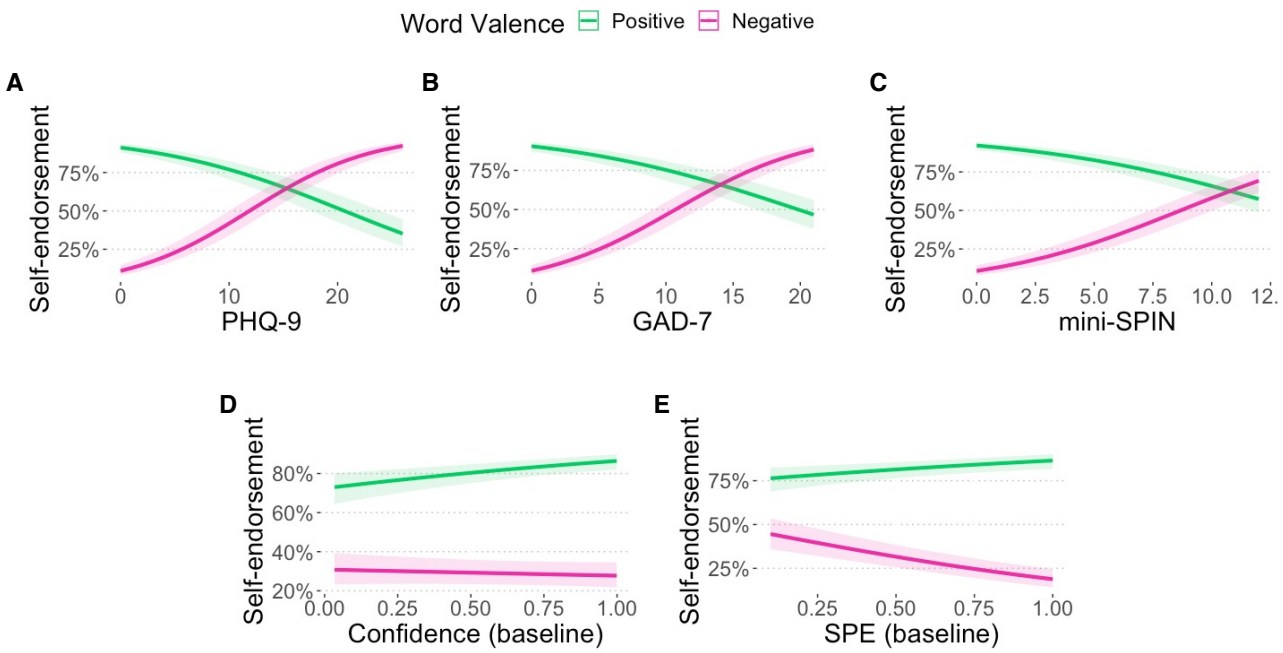
Supplementary Figure 4. Transfer of feedback during intervention to local confidence during test blocks for **A, B)** Exp 1 and **C, D)** Exp 2. **A, C)** Mean local confidence on test blocks plotted as a function of trial number (smoothed using a local regression (loess) span of 0.6) for positive (green) and negative (red) feedback blocks plotted separately for the two intervention tasks in panels along the columns, and test tasks along the rows. **B, D)** Difference in mean confidence of test blocks following positive and negative intervention blocks for the two intervention tasks (x-axis) and two test tasks (colours).



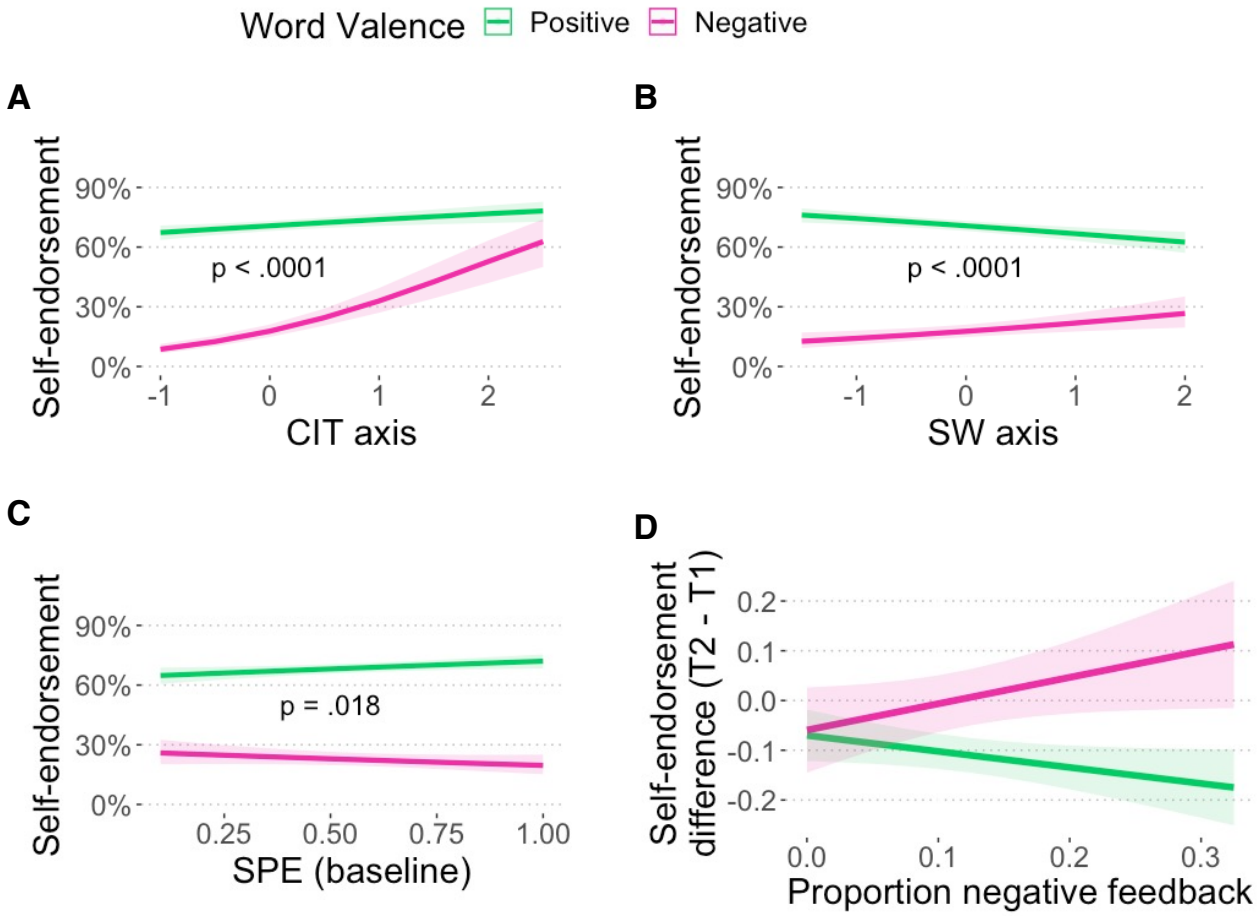
Supplementary Figure 5. **A)** Supplementary Figure 5. Parameter recovery of regression slope parameters linking symptoms to feedback distortion (left) and confidence distortion (middle) and additive bias (right). Simulated values on x-axis and model recovered values on y-axis.



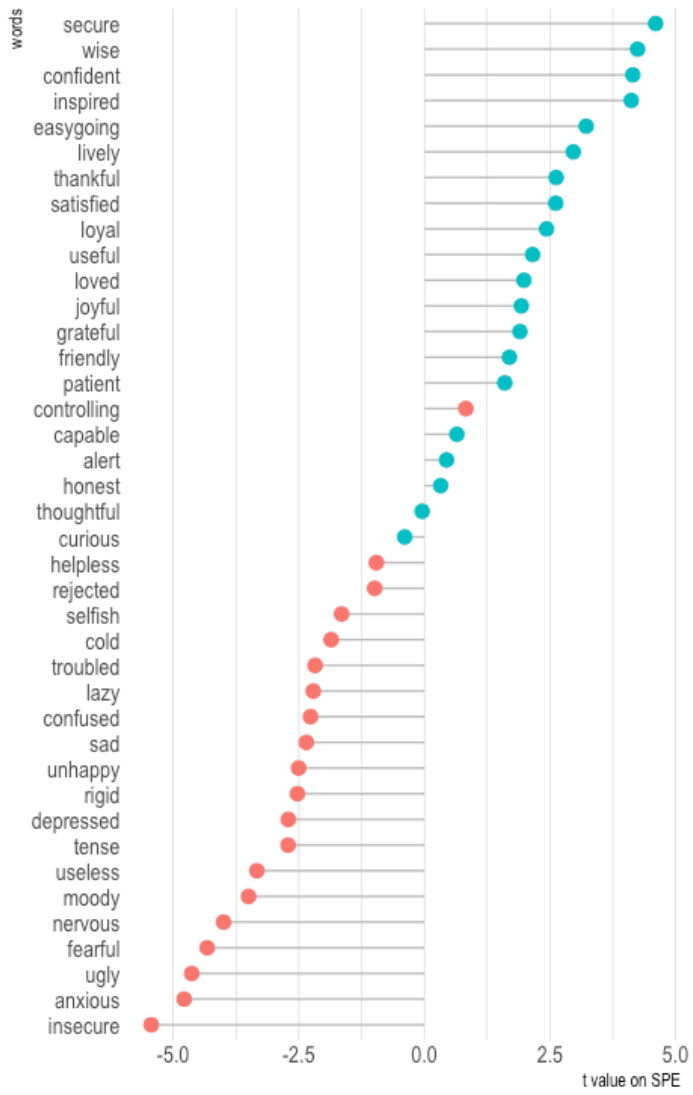
Supplementary Figure 6. Figure 6. Posterior distributions of regression slope parameters for model D6 relating AD symptoms to **A)** confidence distortion and **B)** additive bias in Exp 1 and Exp 2. Regression coefficient values from Exp 2 are divided by 5 (corresponding to differences in the range of PHQ and AD axis scores) to allow direct comparison of coefficients. **C)** A simulation of the difference in global SPEs with low and high AD scores where the additive bias is of the order found in Exp 1 (similar results for Exp 2) showing little impact of AD scores on global SPEs with just additive bias.



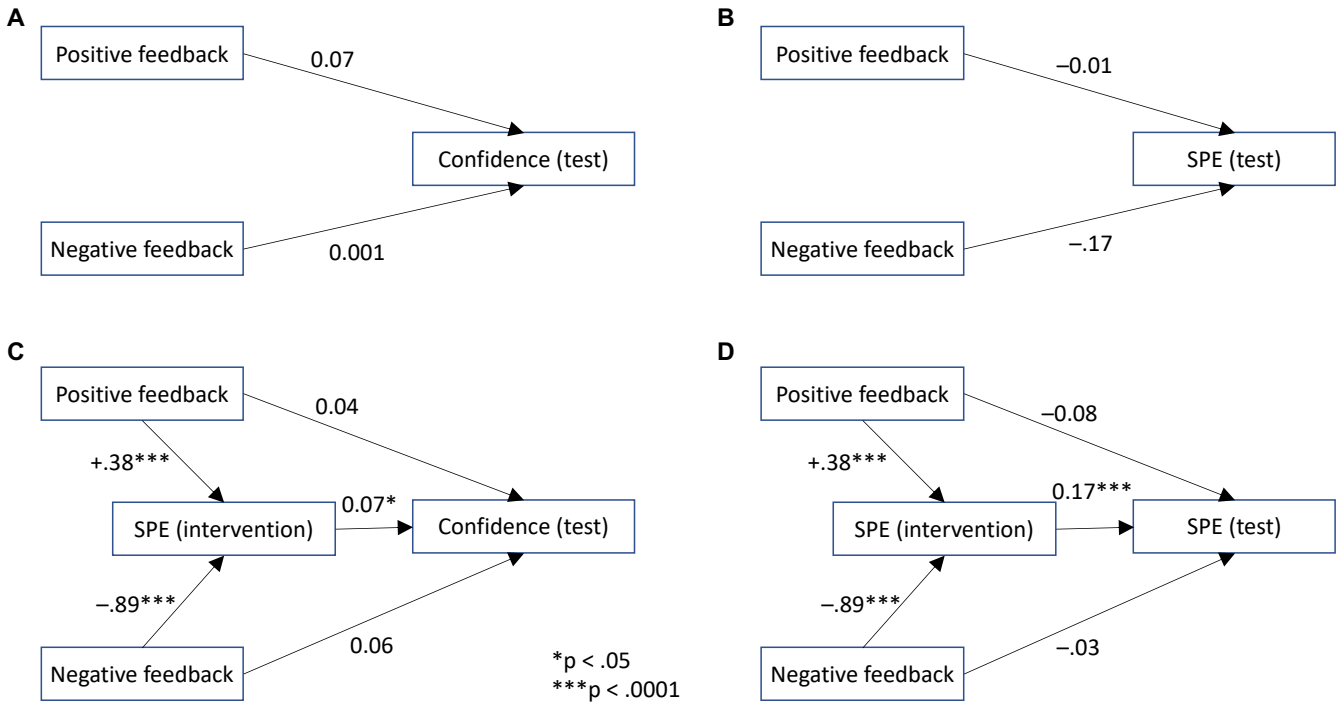
Supplementary Figure 7. Valence-specific relationship of self-endorsement to positive and negative word during the self-referential encoding task in Exp 1 plotted as marginal effects from a logistic regression showing that **A)** PHQ-9, **B)** GAD-7, and **C)** mini-SPIN scores are negatively related to positive word endorsement and positively related negative word endorsement. Similar plots in Exp 1 for **D)** global SPEs, and **E)** mean local confidence during baseline showing that they are associated positively with positive and negatively with negative word endorsements.



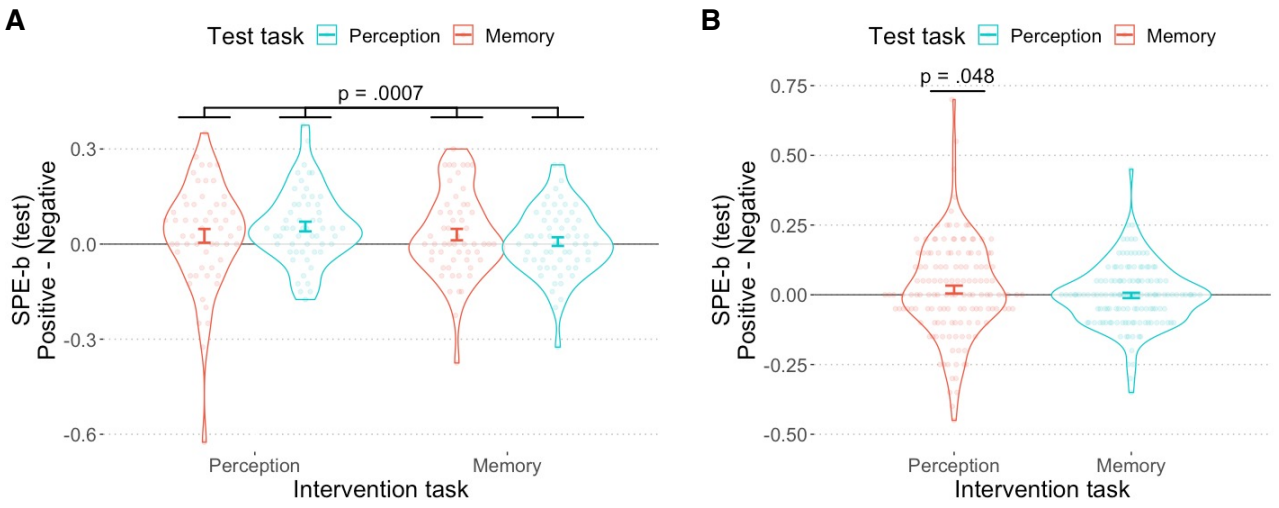
Supplementary Figure 8. Valence-specific relationship of self-endorsements for positive and negative word during the self-referential encoding task in Exp 2 plotted as marginal effects from a logistic regression showing that **A)** Compulsivity and Intrusive Thought transdiagnostic axis is positively related negative word endorsement, and **B)** Social Withdrawal transdiagnostic axis is both negatively related to positive word endorsement and positively related negative word endorsement. Similar plots in Exp 2 for **C)** global SPEs during baseline showing that they are associated positively with positive and negatively with negative word endorsements. Marginal effects of the interaction of proportion of negative feedback and word valence on the self-endorsement difference between SRET timepoint 2 – timepoint 1. The difference in self-endorsement increases for negative and decreases for positive words with the increase in the proportion of trials on which negative feedback was delivered on the previous task block.



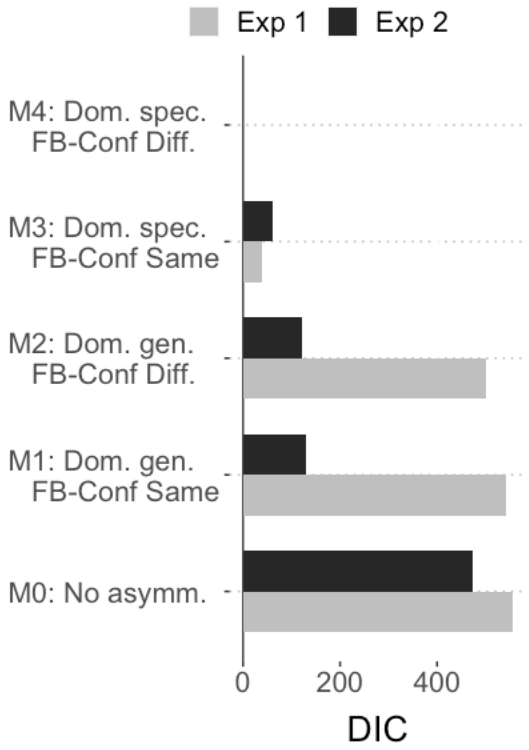
Supplementary Figure 9. t values of for the regression of each positive (blue) and negative (red) word endorsement upon SPE during baseline blocks of Exp 1



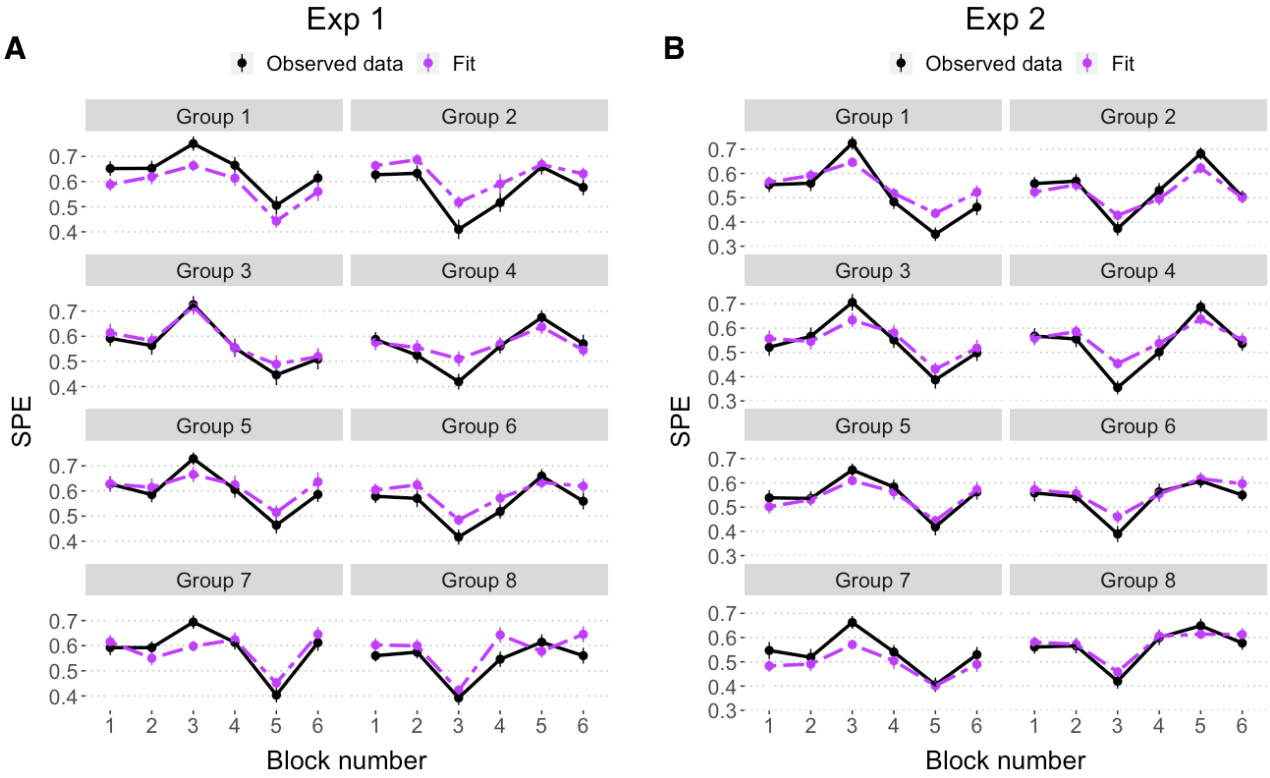
Supplementary Figure 10. A depiction of mediation analyses of positive and negative feedback's effect on test-block local confidence and test-block SPE by intervention-block SPE in Exp 1. Direct effect of feedback on test-block **A)** local confidence and **B)** SPE. Indirect effect of feedback on test-block **C)** local confidence and **D)** SPE, that both were fully mediated by SPE of the intervention block.



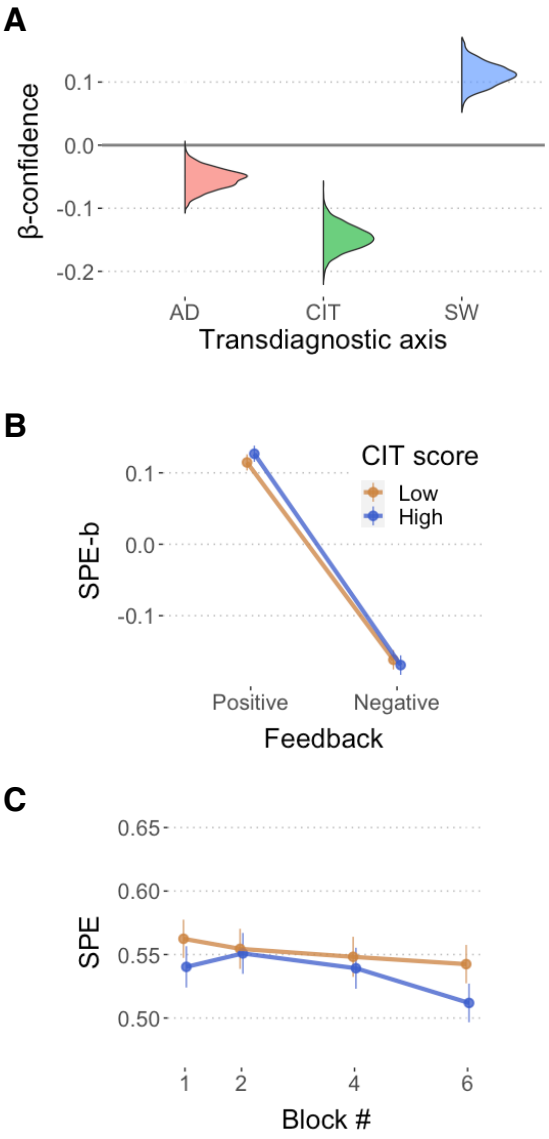
Supplementary Figure 11. Difference in SPEs between test blocks following positive and negative intervention blocks plotted separately for the two intervention tasks and two test tasks in **A)** Exp 1, and **B)** Exp 2.



Supplementary Figure 12. DIC values for five models relating local confidence and performance feedback fit to the SPE data. In both Exp 1 and Exp 2, Model 4, which contained separate asymmetry parameters for feedback and confidence that were domain specific (i.e., separate for memory and perception) provided the best fit.



Supplementary Figure 13. SPE values (black) across the six task blocks averaged for participants belonging to groups that received positive feedback on block 3 and negative feedback on block 5 (i.e., Group 1, 3, 5, 7) shown in the top panel and those that received negative feedback on block 3 and positive feedback on 5 (Groups 2, 4, 6, 8) shown in the bottom panel. Dashed purple lines are model fits. Error bars show SEMs.



Supplementary Figure 14. **A)** Posterior distributions of the regression slope parameter for distortion in using local confidence to form global SPEs in individuals with high AD symptoms along the three transdiagnostic axes. **B)** Baseline-subtracted global SPEs on feedback blocks, and **C)** Global SPEs on confidence-only (i.e., non-feedback) blocks, in Exp 2 for high and low CIT scores.