Learning training boosts causal attribution tendencies similarly to brief cognitive restructuring, depending on individual differences in learning rate

Agnes Norbury[1*], **Tobias U. Hauser**[2,3,4,5], **Raymond J. Dolan**[2,3] **and Quentin J.M. Huys**[1,3]

1 Applied Computational Psychiatry Lab, Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK **2** Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK **3** Wellcome Centre for Human Neuroimaging, University College London, London, UK **4** Department for Psychiatry and Psychotherapy, Medical School and University Hospital, Eberhard Karls University of Tübingen, Germany **5** German Center for Mental Health (DZPG)

*Corresponding Author: agnes.norbury@cantab.net

Abstract

A core part of cognitive therapy for low mood is learning to identify and challenge negative beliefs. However, it is currently unclear whether improved ability to recognise such beliefs, and the biased interpretations of events which may maintain them, is a mechanism of symptom change during treatment. Here, we investigated the effects of both a learning task (training to identify and select self-enhancing interpretations of events) and a brief cognitive restructuring intervention (how exploring alternative explanations of events may result in improved mood) on a task-based measure of causal attribution tendencies. We found that both learning training and the restructuring intervention decreased tendencies to make unhelpful attributions and increased tendencies to make self-enhancing attributions (e.g., decreased tendency to attribute negative events and increased tendency to attribute positive events to self-related causes). Across two studies, changes in attribution tendencies on the causal attribution task were associated with higher learning rates during learning training – an effect which was specific to learning about different kinds of event attribution. Contrary to expectations, there was no evidence that faster learning was associated specifically to changes in attribution tendencies following cognitive restructuring. Since participants with higher learning rate estimates also provided explicit ratings and free-text descriptions of event causes which were closer to the ground truth, we interpret this as representing a greater benefit of learning training in individuals who were better able to understand the task state space (i.e., to recognise different attribution kinds). It is possible that this kind of training, in conjunction with feedback based on interpretable computational model output, may be a useful form of augmentation or learning-support tool during therapy.

Introduction

A core aspect of cognitive therapy for low mood is learning to identify negative beliefs, and exploring alternative explanations for events which challenge these beliefs ('cognitive restructuring') (Beck et al., 1987; Clark, 2022). However, there is currently little definitive evidence as to whether learning to identify negative beliefs and application of restructuring skills are key drivers of symptom change during psychological therapy for low mood (Lorenzo-Luaces et al., 2015, 2016).

This is a hard problem to solve using data from traditional randomized-controlled trials of psychotherapy treatment programs (e.g., cognitive-behavioural therapy, or CBT), given the multiple types of interventions delivered in each program, and lack of fine-grained resolution required to infer temporal dependencies between changes in beliefs and changes in symptoms (Kazdin, 2009; Lorenzo-Luaces et al., 2015). There is some evidence to suggest that greater self-reported frequency and/or skill in applying cognitive strategies is associated with greater overall symptom reduction following C(B)T (Hundt et al., 2013; Strunk et al., 2014; Hawley et al., 2017; Gumport et al., 2018; Forand et al., 2018; Schmidt et al., 2019). However, the degree of conceptual overlap between self-report measures of cognitive skills and symptoms themselves (the 'jangle' fallacy) means it is often hard to disentangle changes in the former from overall treatment response or residual symptom burden (Hundt et al., 2013; Lorenzo-Luaces, 2023).

Behavioural measures of cognitive processes may be one way to help solve this problem, since they are less close to the target construct of interest - symptom change (Moutoussis et al., 2018; Reiter et al., 2021; Huys et al., 2021). Combining cognitive-behavioural measures with randomized allocation of therapy-like interventions in high-throughput testing is a fast and efficient way to test whether specific components of psychological treatments may causally impact specific cognitive processes, prior to extending testing to resource-intensive clinical settings (Dercon et al., 2023; Norbury et al., 2023).

Here, we use this approach to test whether a behavioural measure of attribution tendencies (how people tend to reason about the causes underlying events) is affected by 1) training in learning to identify different kinds of causal attributions (a learning task intervention) and 2) practice in identifying and challenging unhelpful attributions of events in their own lives (a brief cognitive restructuring intervention). Cognitive therapy is essentially a process of learning (e.g., Moutoussis et al. 2018), and it has been suggested that individuals with greater capacity to engage in learning during treatment may show greater benefits (Bruijniks et al., 2019). We therefore further examined whether individual differences in learning task performance were related to individual differences in response to brief cognitive restructuring.

Across two studies, we found evidence that both learning task training and the cognitive restructuring intervention affected causal attribution tendencies: shifting these away from unhelpful or 'depressogenic' patterns (e.g., lower tendency to attribute negative events to self-related or internal causes) and towards self-enhancing styles (e.g., higher tendency to attribute positive events to internal causes). In both studies, greater shifts in attribution tendencies were associated with higher learning rate estimates on the learning training task. Since we found no association between attribution change and learning rates from a matched control task (which did not concern causal attributions), we interpret this as being due to greater ability to discriminate between different kinds of attributions, or better understanding of the learning task state space. There was no evidence that individuals with faster learning rates showed greater responses to the cognitive restructuring intervention specifically. We discuss these findings with reference to recent proposals to augment psychological treatments with strategies aimed at boosting learning and memory of treatment content, and who this might be most effective for (Harvey et al., 2014; Nord et al., 2023).

Results

Overview of study design and measures

Here, we report results of two studies with similar overall design (Figure 1a). In both studies, participants completed a task-based measure of causal attribution tendencies, before and after two types of intervention: a learning training (or control learning) task, and a brief cognitive restructuring (or control) intervention. This design allowed us to measure changes in tendency to attribute events to causes which are *internal* (related to the self, compared the outside world) and *global* (likely to be active in all situations, rather than this specific one alone), following either type of intervention. Biased interpretation of events along these dimensions is commonly observed in people with depression, and is thought to be critical to the maintenance of low mood according to cognitive theories (see Abramson et al. 1978; Beck et al. 1987, Discussion).

The different study tasks are represented in Figure 1b, and described fully in the Methods. The main dependent measures across studies were attribution tendencies on the causal attribution task. Here, participants were presented with a series of brief descriptions of everyday events, and asked to choose which of four listed causal explanations they think would be the the most likely, if such an event had happened to them. Model-based analysis of task choices yields parameters describing individual tendencies to ascribe positive and negative events to internal (*vs* external) and global (*vs* specific) causes. Of note, these measures have previously been found to exhibit good psychometric properties, and are also correlated with self-reported negative beliefs about the self (as measured by the Dysfunctional Attitudes Scale, Norbury et al. 2023).

The learning training task also presented participants with a series of events and potential causal explanations, but this time using a third-person perspective and explicit feedback. Participants were instructed that that they would be learning about how a hypothetical person in a particular mood might reason about the causes behind events, and they would have to learn to select the 'correct' kinds of reasons for that person via explicit feedback. The control learning task was identical in structure, but required participants to learn to sort different objects according to physical property dimensions, and did not involve understanding different kinds of causal explanations (see Methods). Model-based analysis of data from both tasks yields estimates of learning rates, or how quickly participants are able to update their response option choices based on explicit feedback.

Finally, in between completion of the two causal attribution task versions, and following completion of the learning training or control learning tasks, participants completed a brief therapy-informed intervention. The cognitive restructuring intervention consisted of information about a cognitive model of mood (link between interpretations of events and feelings), interactive exercises identifying helpful and unhelpful attributions of the same events, inviting people to practise generating alternative explanations for recent events in their own lives, and a summary comprehension quiz. The control intervention was closely matched in terms of length, interactivity, and self-relevant exercise content, but based on materials from emotion-focused therapy

Greenberg (2015), that did not reference how different interpretations of events may influence mood. Please note that data from the first study were partially reported previously (Norbury et al., 2023). The learning task data were not included in the previous analysis.

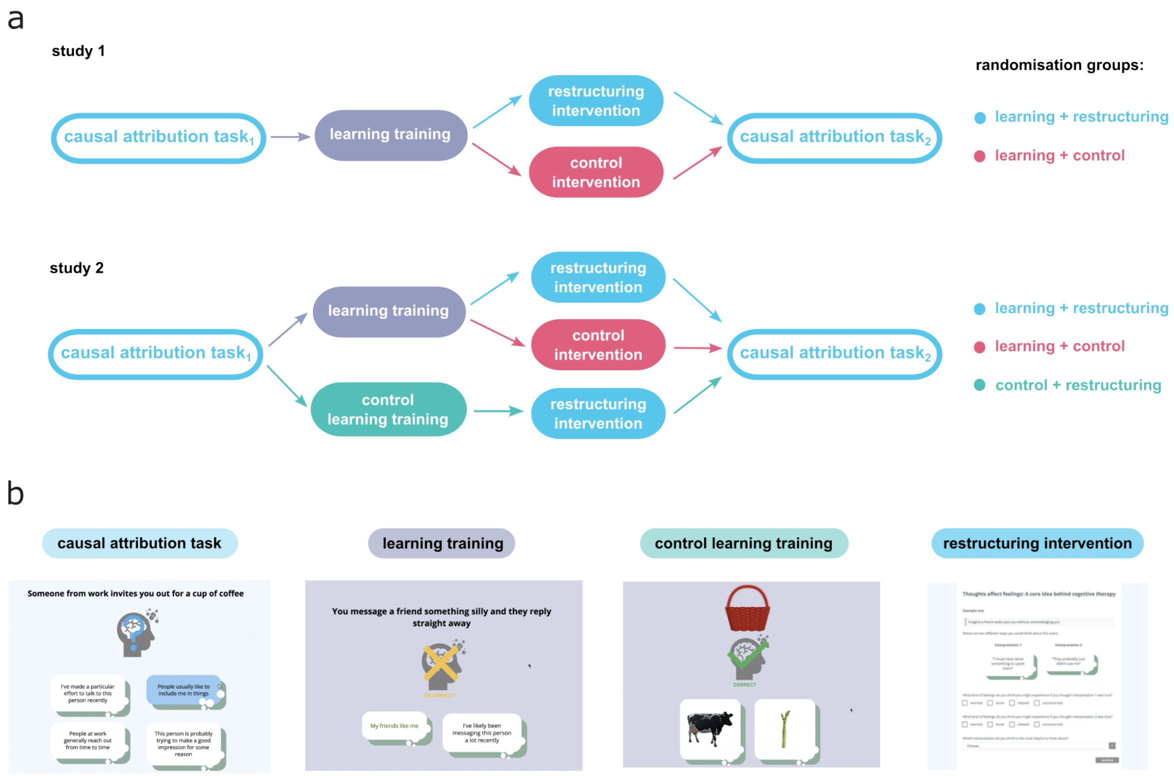
a
study 1
causal attribution task$_1$
learning training
restructuring intervention
control intervention
causal attribution task$_2$
randomisation groups:
● learning + restructuring
● learning + control
study 2
causal attribution task$_1$
learning training
control learning training
restructuring intervention
control intervention
restructuring intervention
causal attribution task$_2$
● learning + restructuring
● learning + control
● control + restructuring
b
causal attribution task
Someone from work invites you out for a cup of coffee
I've made a particular effort to talk to this person recently
People usually like to include me in things
People at work generally reach out from time to time
This person is probably trying to make a good impression for some reason
learning training
You message a friend something silly and they reply straight away
INCORRECT
My friends like me
I've likely been messaging this person a lot recently
control learning training
CORRECT
restructuring intervention
Thoughts affect feelings: A core idea behind cognitive therapy
Example one
Imagine a friend walks past you without acknowledging you
Below are two different ways you could think about this event.
Interpretation 1
Interpretation 2
"I must have done something to upset them"
"They probably just didn't see me"
What kind of feelings do you think you might experience if you thought interpretation 1 was true?
□ worried □ tense □ relaxed □ unconcerned
What kind of feelings do you think you might experience if you thought interpretation 2 was true?
□ worried □ tense □ relaxed □ unconcerned
Which interpretation do you think is the most helpful to think about?
Choose...
continue

Figure 1: **Overview of study designs and measures.** **a** Experimental designs and randomisation conditions for each study. In both studies, a cognitive-behavioural measure of causal attribution tendencies (the causal attribution task), was completed pre- and post- completion of two types of intervention. In study 1, all participants completed the learning training task, and were randomly allocated to complete either brief cognitive restructuring or a control intervention. In study 2, participants were randomly assigned to complete either learning training or a control learning task, followed by either brief cognitive restructuring or a control intervention. All studies took place online, in a single experimental session (around 1 hour in length). **b** Representative screen-shots of different study measures. The causal attribution task asks participants to choose between four different potential explanations of events, if such an event happened to them. The learning training task uses a third-person framing, and requires participants to learn which kinds of explanations are thought to be correct for a hypothetical person in a particular mood state, given explicit feedback. The control learning task is identical in structure, but requires participants to learn about the properties of objects, rather than causal explanations. The brief cognitive restructuring (and control) interventions both took the form of a series of interactive worksheets, which asked participants to learn about a particular therapy model, and apply it to recent events from their own lives.

Participants

Participants for both studies were recruited from an online research participation platform and are described in Table 1. Samples were reasonably diverse in terms of age, gender, and neurodivergence, but were predominantly White, with relatively stable financial and housing status. In both studies, samples showed some evidence of self-selection for mental health research, given 40% reporting of previous treatment for a mental health problem, and mild-to-moderate average endorsement of current low mood and social anxiety symptoms (proportion of participants above cut-off score for clinically-significant depressed mood according to the PHQ9 = 32%, 27%; proportion of participants with significant social anxiety according to the miniSPIN = 48%, 46%; for full distributions of clinical scores by condition in each study see Figure S1).

		Study 1	**Study 2**
	N	200	164
Age (years)	mean (SD)	37.2 (10.5)	36.9 (10.5)
	range	19-63	20-65
Gender	Woman	110 (55%)	75 (46%)
	Man	86 (43%)	86 (52%)
	Non-binary or other	4 (2%)	3 (2%)
Race/ ethnicity	White	165 (83%)	125 (78%)
	Asian	14 (7%)	13 (8%)
	Black	5 (3%)	12 (7%)
	Mixed	8 (4%)	10 (6%)
	Other	8 (4%)	3 (2%)
Employment status	Employed	147 (74%)	127 (77%)
	Unemployed	19 (10%)	13 (8%)
	Not seeking	33 (17%)	24 (15%)
Financial status	Doing okay	95 (48%)	85 (52%)
	Just about getting by	74 (37%)	61 (37%)
	Struggling	30 (15%)	18 (11%)
Housing status	Homeowner	90 (45%)	87 (53%)
	Tenant	86 (43%)	49 (30%)
	Other	23 (12%)	28 (17%)
Neuro-divergence	Yes	25 (13%)	25 (15%)
	No	167 (84%)	135 (82%)
	Prefer not to say	8 (4%)	4 (2%)
Previous treatment for a mental health problem	Yes	89 (45%)	55 (34%)
	No	103 (52%)	105 (64%)
	Prefer not to say	8 (4%)	4 (2%)
If yes, type of treatment (all that apply)	Talking therapy	62 (31%)	36 (22%)
	Medication	62 (31%)	37 (23%)
	Self-guided	39 (20%)	27 (17%)
	Other	5 (3%)	4 (2%)
PHQ9 total	mean (SD)	7.3 (6.2)	6.3 (5.8)
DAS-SF total	mean (SD)	19.2 (4.6)	18.6 (4.8)
miniSPIN total	mean (SD)	5.8 (3.6)	5.5 (3.4)

Table 1: **Self-reported demographic and clinical data for all study participants.** Self-reported race/ethnicity was based on information provided by Prolific. All other information was recorded via our custom demographic questionnaire (see Methods). Employment status categories were employed (including full-time and part-time employment), unemployed (job seekers and those unemployed owing to ill health), and not seeking employment (stay-at-home parents, students, and retirees). Housing status categories were homeowner (including those with a mortgage), tenant, and other (living with family or friends, homeless, or living in a hostel). Neurodivergence was defined as “a term for when someone processes or learns information in a different way to that which is considered ’typical’: common examples include autism and ADHD”. Categories for previous mental health treatment were talking therapy (including cognitive-behavioural therapies), medication, self-guided (e.g., workbooks or apps), or other. PHQ9 total, Physician’s Health Questionnaire 9-item measure of depressed mood total score (possible range 0-27). miniSPIN total, mini Social Phobia Inventory total score (possible range 0-12). DAS-SF total, Dysfunctional Attitude Scale short-form total score (possible range 9-36).

Separate effects of learning training and brief cognitive restructuring on causal attribution tendencies

We first sought to examine whether there was evidence for separate effects of completing the learning training task and brief cognitive restructuring intervention on attribution tendencies, as measured on the causal attribution task. Specifically, we used a hierarchical Bayesian modelling approach to test whether there was evidence for additional group-level effects of having been randomized to learning training *vs* control learning task conditions, and cognitive restructuring *vs* control intervention conditions (Methods, Equation 6). Analysis models were the same as described previously (see Norbury et al. 2023), with an extension for study 2 data described below.

Since study 1 participants all completed the learning training task, in this data we were only able to examine group-level effects of cognitive restructuring *vs* control intervention conditions. As reported previously, we found that completion of the brief cognitive restructuring intervention resulted in decreased tendency to attribute negative events to internal causes (posterior estimate=-0.48 [90%CI -0.65– -0.31]), and increased tendency to attribute positive events to general or global causes (posterior estimate=0.49 [90%CI 0.19–0.80]) (Figure 2a,b; Table S1). Of interest, the group means for each parameter showed some evidence of shifts between timepoints – with all participants appearing to show slightly higher mean endorsement of internal and global attributions of positive events at the second measurement (Figure 2b). These group-level shifts could represent common effects of completing the cognitive restructuring and control interventions on attribution tendency - although as the control intervention made no reference to how interpretations of events may affect mood, or reappraisal strategies, this might be considered unlikely. An alternative explanation is that these effects are due to completion of the learning training task by all study participants, since this task directly involves learning to recognise different kinds of attributions.

We were able to test this idea directly in study 2 data, since this included randomisation conditions where participants completed the control learning task, as well as cognitive restructuring and control conditions. To formally test whether completion of the learning task resulted in group-level changes in attribution tendencies, we augmented the analysis model for study 2 data such that post-intervention (time 2) attribution tendencies could be influenced by group-level parameters related to learning training condition, as well as restructuring intervention condition (Methods, Equation 7).

Model comparison revealed that the model with additional effects for learning task condition had marginally better predictive accuracy for causal attribution task data than the model with restructuring intervention condition alone (difference in expected log pointwise predictive density for left-out causal attribution task data, $ELPD_{diff}$=-0.4, but of less than 5x than the standard error of the estimate, se=6.8), suggesting that this indeed had an additional impact on changes in attribution tendencies.

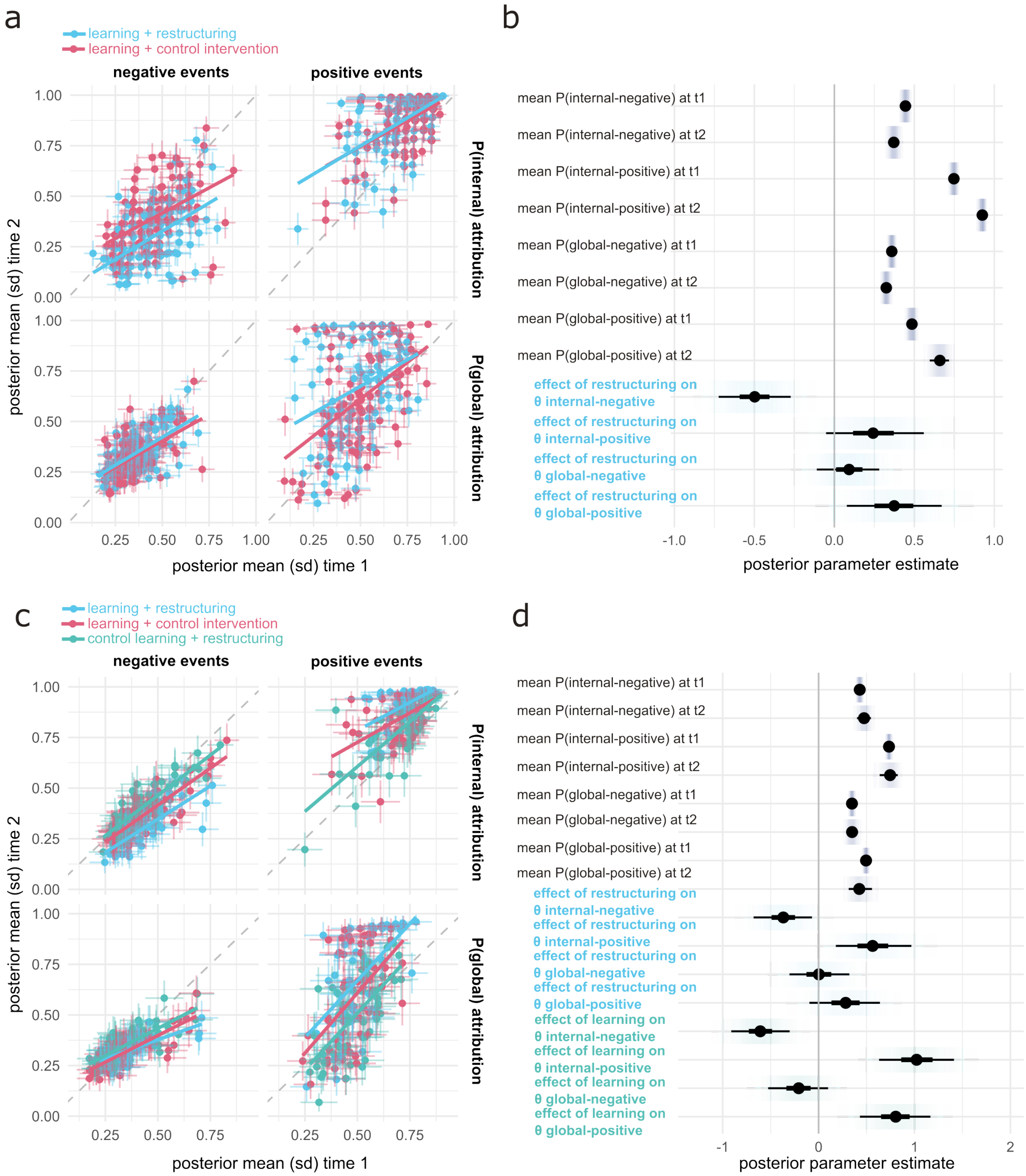

Figure 2: **Independent effects of learning training and brief cognitive restructuring on causal attribution**. **a** Posterior mean (and SD) parameter estimates for the causal attribution task for each participant at time 1 (pre-intervention) and time 2 (post-intervention), by randomisation group, in study 1 participants (N=200). Parameter estimates plotted here represent the probability of endorsing a given kind of attribution for positive and negative events, which are governed by the latent trait parameters (θ). Lines of best fit for mean time 1 *vs* time 2 estimates for individuals in each group are plotted for illustration purposes. **b** Posterior parameter

Figure 2: estimates for group means (over all participants/randomisation conditions) for each parameter at each time point, and the additional effect of the cognitive restructuring intervention at time 2, in study 1 participants. Thick inner lines represent 50%, and thin outer lines represent 90% Credible Intervals, the point estimate is the mean, and shading represents posterior probability density. For visualisation purposes, intervention effects (bold text) have been scaled by the square root of the mean posterior variance estimates for parameter values at time 2, making them roughly equivalent to SMDs. **c** The same plot as (a), for study 2 participants (*N*=164). **d** The same plot as (c), for study 2 participants. Here, group-level effects on time 2 parameter estimated were modelled separately for participants who completed the restructuring *vs* control intervention, and learning *vs* control learning training.

Inspection of changes in individual parameter estimates between time 1 (pre-intervention) and time 2 (post-intervention) shows that participants who completed both the learning training task and cognitive restructuring intervention showed the greatest shifts away from depressogenic (internal, global) attributions of negative events, and towards self-enhancing attributions of positive events (Figure 2c). Examination of posterior parameter estimates for group-level effects revealed that, when accounting for learning task condition, the restructuring intervention both decreased tendency to attribute negative events to internal causes (posterior estimate=-0.30 [90%CI -0.49– -0.11]), and increased tendency to attribute positive events to internal causes (posterior estimate=0.63 [90%CI 0.29–0.97]) (Figure 2d, Table S2). There was also evidence for separate group-level effects of completion of the learning training vs control learning task on attribution tendencies. Specifically, completion of the learning training task resulted in further decreased internal attribution of negative events, as well as increased internal and global attribution of positive events on the causal attribution task (posterior estimates=-0.49 [90%CI -0.68– -0.30], 1.15 [90%CI 0.80–1.49, 0.94 [90%CI 0.60–1.27], Table S2).

It therefore appears that, at the group level, both completion of the restructuring intervention and completion of learning training task affected causal attribution tendencies for everyday events – with both intervention components resulting in decreased tendency to choose unhelpful and increased tendency to choose self-enhancing interpretations.

Learning rates from the learning training task are associated with change in self-enhancing attributions, but not specifically following cognitive restructuring

Since we might reasonably expect that the effects of the learning task intervention depend on individual differences in learning performance, we next explored whether model-based metrics of learning were related to changes in causal attribution tendencies.

Learning rates were estimated from learning training task data using a simple Rescorla-Wagner model (see Methods). Briefly, this algorithm accounts for individual differences in the updating of response option (causal explanation type) values according to feedback using learning rate parameters (α), for individual differences in starting values of different explanation types (bias towards away or away from internal-global explanations of positive and negative events), and for individual differences in how value-driven participants are in their choice behaviour (softmax inverse temperature parameter, β). Full information on model derivation via model comparison, chosen model performance, and simulation-based calibration analysis (including recovery of individual model parameters) can be found in the Methods and Supplementary Results.

Given we observed minimal variation in learning about negative events in our samples (in either raw choice accuracy or posterior model parameter estimates, Figure S2), we focused our analysis on learning estimates for positive events: i.e., speed of learning to select self-enhancing attributions of positive events, rather than speed of learning to avoid unhelpful attributions of negative events. Positive learning rates from the learning task were compared to changes in self-enhancing attributions (internal and global interpretations of positive events) on the causal attribution task.

As a first-pass analysis, we examined whether any relationship was evident between point estimates (posterior parameter means) from separately modelled learning and causal attribution task data. We then carried out a formal test of association (that more appropriately takes into account precision of estimation of different parameters) by analysing learning and causal attribution task data together in a joint hierarchical Bayesian model. This approach allows for the direction estimation of associations between relevant parameters in the form of posterior beta weights (see Methods).

Associations between separately-modelled learning and attribution task data. When comparing point estimates of learning rates to changes in point estimates of parameters governing attribution tendencies, we observed associations between positive learning rates (α_{pos}) and changes in internal and global attributions of positive events (study 1: $R_{\alpha_{\text{pos}},\Delta\text{internal}} = 0.24, p < 0.001$, $R_{\alpha_{\text{pos}},\Delta\text{global}} = 0.10, p = 0.15$, study 2: $R_{\alpha_{\text{pos}},\Delta\text{internal}} = 0.24, p < 0.001$, $R_{\alpha_{\text{pos}},\Delta\text{global}} = 0.20, p < 0.001$; all correlations weighted by the posterior precision of α_{pos} estimates; Figure 3a,d). These relationships were not evident for learning rates derived from the control learning task ($Rs = 0.14, 0.10$; Figure 3d).

There was no strong evidence that the strength of these correlations differed significantly between participants who received the cognitive restructuring compared to control interventions (for change in internal-positive attribution tendencies, study 1: $Rs = 0.27, 0.21$, study 2: $Rs = 0.12, 0.22$; for change in global-positive attribution tendencies, study 1: $Rs = 0.20, 0.04$, study 2: $Rs = 0.25, 0.18$, all $p > 0.9$, Fisher's R-to-Z tests).

Joint hierarchical Bayesian modelling of learning and attribution task data. In order to formally assess associations between different task parameters, we constructed joint models of learning and attribution task data. Specifically, the model of the causal attribution task data was extended via inclusion of beta weight parameters that reconstructed group-level effects on time 2 parameter estimates as a linear combination of an intercept term plus a beta weight*individual learning rate parameter estimates. The prior values for all beta weights were centred around zero – such that a significant contribution of learning rate parameters to time 2 (post-intervention) parameter estimates should be reflected in a posterior beta weight estimate with a Credible Interval excluding zero (see Methods).

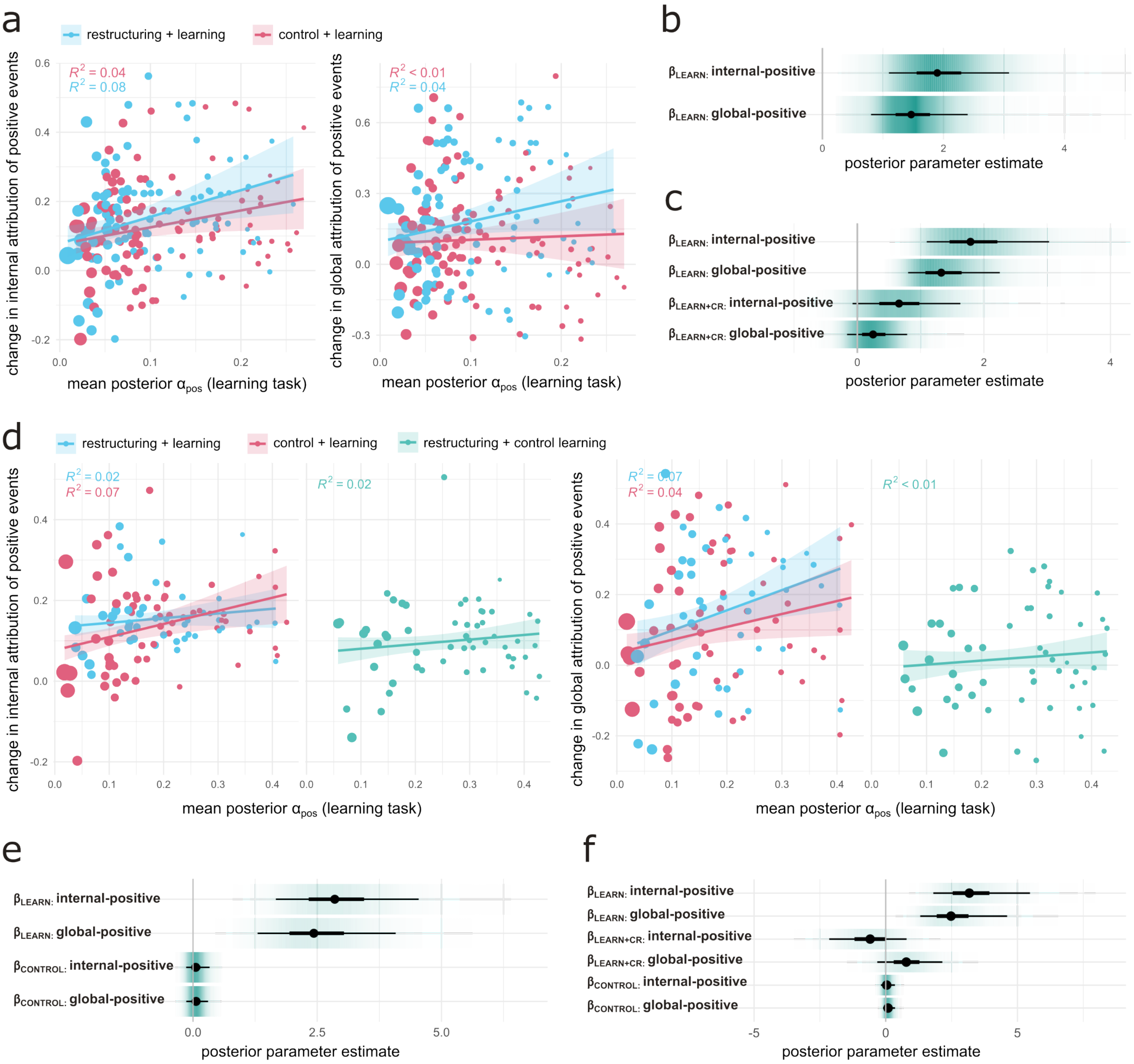

Figure 3: **Changes in self-enhancing attributions were positively associated with learning rate estimates from the learning training task, but this effect was not greater in participants who completed cognitive restructuring.** **a** Correlations between posterior mean estimates for positive learning rate from the the learning training task (α_{pos}) and changes in mean values of parameters governing tendency to select internal and global attributions of positive events in study 1 participants. Point weights represent precision of estimation of α_{pos} (1/posterior SD). **b** Posterior estimates of group-level effects from joint models of learning and causal attribution task data. β_{LEARN}, posterior estimates for weight of α_{pos} estimates on change in internal and global attributions of positive events. For visualization purposes, β estimates have been scaled by the ratio of SDs of the predictor (α_{pos}) to outcome (mean posterior parameter variance estimates), making them roughly equivalent to standardized regression coefficients. Black lines represent 50,90% posterior Credible Intervals, and shading represents posterior probability density. **c** The same plot as b, for a joint model with additional β weights for participants who completed brief cognitive restructuring in addition to learning training ($\beta_{LEARN+CR}$).

Figure 3: **d** The same plot as a, for study 2 participants. **e** The same plot as b, for study 2 participants. $\beta_{CONTROL}$, posterior estimates for weight of control learning task learning rate estimates on change in attribution tendencies. **f**, The same plot as e, for a joint model with additional weights for participants who completed completed brief cognitive restructuring in addition to learning training.

Two types of joint model were constructed. The first set of models tested if learning rates were associated with changes in attribution tendencies regardless of restructuring intervention condition (β_{LEARN}, Equation 10). For study 2 data, the model contained separate weights for each learning task condition, in order to test if associations were specific to learning about causal attributions (β_{LEARN}, $\beta_{CONTROL}$, Equation 11). The second set of models tested whether these associations might differ in strength depending on whether or not participants completed the cognitive restructuring intervention: i.e., if faster learning about attributions *plus* practice in identifying and challenging these in participants' own lives might result in the greatest improvement. This was done via inclusion of an additional beta weight parameter contributing to time 2 parameter estimates, only for participants who completed brief cognitive restructuring ($\beta_{LEARN+CR}$, see Methods, Equation 12, Equation 13).

Results of the first joint models provided strong evidence of positive relationships between α_{pos} estimates and changes in internal and global attributions of positive events, across intervention conditions, in study 1 participants (β_{LEARN} internal-positive=0.53 [90%CI 0.34–0.75], β_{LEARN} global-positive=0.44 [90%CI 0.28–0.62], Figure 3b, Table S3). These effects were replicated in study 2 data (β_{LEARN} internal-positive=0.29 [90%CI 0.19–0.40], β_{LEARN} global-positive=0.26 [90%CI 0.16–0.37]) - but were not evident for learning rates estimated from the control learning task ($\beta_{CONTROL}$ internal-positive=0.01 [90%CI -0.01–0.03], $\beta_{CONTROL}$ global-positive=0.01 [90%CI -0.01–0.02], Figure 3e, Table S4). This suggests that associations between speed of learning and subsequent change in self-enhancing attribution tendencies were specific to learning training in the domain of causal attributions.

Results of the second joint models provided marginal evidence for an additional influence of α_{pos} estimates on change in internal-positive attributions in participants who completed the restructuring intervention in study 1 ($\beta_{LEARN+CR}$ internal-positive=0.22 [90%CI 0.02-0.42]), but this effect was not replicated in study 2 ($\beta_{LEARN+CR}$ internal-positive=-0.07 [90%CI -0.19–0.05]. In neither study was there evidence for an additional influence of α_{pos} estimates on change in global-positive attributions in restructuring group participants (study 1: $\beta_{LEARN+CR}$ global-positive=0.09 [90%CI -0.03–0.22],Figure 3c, Table S5, study 2: $\beta_{LEARN+CR}$ global-positive=0.09 [90%CI -0.01–0.19], Figure 3f, Table S6). There was therefore no strong evidence in favour of a selective interaction between faster learning on the learning training task and response to the cognitive restructuring intervention, specifically.

Importantly, when the likelihood of the attribution task data was compared between the original analysis model and joint models, both joint models had superior predictive accuracy in left-out data (Table S7). This suggests that overall estimates of learning rates from the learning task were providing relevant information for inferring post-intervention causal attribution task parameter values.

Learning rates from learning task data reflect understanding of the task state-space

In order to understand why this might be the case, we explored relationships between learning rates estimates and other learning task data. Specifically, after each learning task scenario, participants were asked to provide explicit ratings of the kinds of causes that were thought to be 'correct', along internal-external and global-specific dimensions, and also provided free-text descriptions of each cause. This data provides insight into how well participants understood the ground truth of each scenario, as reflected in specific types of attributions they had to learn to select.

Full analysis of learning task data (choice accuracy, response times, explicit-cause ratings and free-text cause descriptions) in available in the Supplementary Results, alongside further details of model-based analysis of learning task data. In summary, at the group level, participants were able to learn to perform the task (improved in choice accuracy over trials and scenarios), although in both studies participants were slower to learn to select self-enhancing attributions for positive events than to avoid unhelpful attributions of negative events (Figure S2). Overall, participants were also able to describe the kinds of causes that were correct for each scenario using explicit ratings scales (Figure S3). Natural-language processing (NLP) classification of free-text descriptions also distinguished between scenarios, and participants with more accurate explicit cause ratings provided free-text descriptions with higher probabilities for ground truth cause labels (at least along the internal-external dimension; Figure S4) – suggesting these measures to some extent tap common understanding of task structure.

Relationships between positive learning rates and explicit cause ratings. Posterior mean estimates of learning rates for positive events (α_{pos}) were positively associated with the explicit ratings of 'correct' causes for each task scenario. In other words, participants who learned faster to select internal-global attributions of positive events during the task were also able to better identify that correct causes were internal (self-related) and global (general), using explicit rating scales (study 1: $Rs = 0.2 - 0.35, p < 0.005$, study 2: $Rs = 0.20 - 0.33, p <= 0.033$, Figure S5). These relationships persisted in linear mixed-effects models controlling for scenario number and mean posterior inverse temperature (β) parameter values, weighted by posterior precision of α_{pos} estimates (internal-external cause ratings: study 1: $F_{1,238} = 17.2$, study 2: $F1, 114.5 = 9.5, p < 0.005$; global-specific cause ratings: study 1: $F_{1,235} = 13.2; p < 0.001$, study 2: $F1, 112.7 = 4.1, p < 0.05$). For the control learning task, there was little variance in explicit ratings data (see Figure S3c), so these relationships were not examined.

Relationships between positive learning rates and free-text cause description label probabilities. In study 1, posterior mean estimates of α_{pos} were significantly correlated with classifier label probabilities for positive events in each scenario, along the internal-external dimension ([events were caused by] "myself", $Rs = 0.24 - 0.28, p < 0.001$; [events were caused by] "other people", $Rs = 0.2 - 0.32, p < 0.001$; Figure S6). These effects persisted in linear mixed-effect models controlling for scenario number and posterior mean inverse temperature (β) parameter values, weighted by posterior precision of α_{pos} estimates ($F_{1,239} = 12.1, p < 0.001; F_{1,267} = 5.6, p < 0.02$). In study 2, this association was only marginally evident ([events were caused by] "myself", $Rs = 0.17 - 0.22, p < 0.07$; [events were caused by] "other people", $Rs = 0.14 - 0.25, p < 0.10$), and did not survive in the controlled model ($F_{1,111} = 2.13, p = 0.15$, $F_{1,140} = 2.59, p = 0.11$). No relationships were evident between learning rates and classifier label probabilities for the free-text descriptions of positive events in the global-specific dimension in either sample, likely as this dimension was represented much more noisily in classifier output (see Supplementary Results).

Relationships between positive learning rates and self-reported demographic and clinical data. Across studies, there was no evidence that mean posterior α_{pos} estimates varied according to participant age, gender identity, neurodivergence, or previous experience of talking therapy (all $Rs < 0.1$, Figure S7, Figure S8). Interestingly, whilst in study 1, there was no evidence of a relationship between learning rates and current depression symptom severity (PHQ9 total score, $R < 0.1$), in study 2 participants with higher current depression symptom severity had lower positive learning rates ($R = 0.26, p = 0.005$). Control task learning rates did not vary by gender, neurodivergence, or current mental health symptoms, but were negatively associated with age ($R = 0.50, p < 0.001$, Figure S7).

Together, this indicates that participants who learned more quickly to select self-enhancing attributions in a reinforced setting were also able to better describe the types of causes reinforced as correct during the task. Participants with higher learning rate estimates (α_{pos}) may therefore have had greater understanding of the ground truth dimensions along which response options (potential causal explanations) varied, allowing them to more quickly choose the 'correct' responses for a given scenario. This ability did not appear to be systematically influenced by sociodemographic factors or related to current experience of mental health symptoms – although inconsistent findings across studies may mean this could be usefully explored in future work.

Discussion

Negative beliefs about the self and biases towards interpreting negative events as being due to enduring, overly-general and self-related factors are core drivers of low mood according to cognitive theories of depression (Beck et al., 1987; Abramson et al., 1978; Clark and Beck, 2010). There is widespread evidence from cross-sectional studies that depression is related to increased endorsement of negative self-beliefs and decreased tendency to exhibit self-enhancing attributions (e.g., internal attribution of positive events, and external attribution of positive events; Mezulis et al. 2004; Chahar Mahali et al. 2020; Bartucz et al. 2022), with some evidence this kind of attributional style may be predictive of future depressed mood (Pearson et al., 2015). Cognitive restructuring, a form of psychology therapy which involves learning to identify and challenge overly negative or unhelpful interpretations of events, is an effective treatment for low mood (Clark, 2013) – however, we currently lack conclusive evidence as to whether intervening on negative self-beliefs or attributional tendencies is necessary or sufficient for treatment success (Lorenzo-Luaces et al., 2015, 2016) – and if there are individuals for whom this may be more or less effective.

Meta-analyses of RCT data suggest that treatment effects on depression symptom severity are positively related to changes in self-reported negative beliefs - however this is also the case for non-cognitive treatments and pharmacotherapies, which do not directly target such beliefs (Garratt et al., 2007; Cristea et al., 2015; Kazantzis et al., 2018). Where more fine-grained temporal data is available, some studies have identified that changes in negative cognition precede improvements in depression symptoms (Lorenzo-Luaces et al., 2015, 2016; Schmidt et al., 2019) – however this has not been consistently observed (e.g., Lemmens et al. 2017), and in larger samples relationships have been found to be small and bidirectional (Persons et al., 2023). Similarly, there is some evidence that self-reported frequency and ability to use cognitive skills during the course of treatment is related to success of psychological interventions for depression (Hundt et al., 2013; Strunk et al., 2014; Hawley et al., 2017; Forand et al., 2018), including in online settings (Gumport et al., 2018). Although in reality relationships between negative self-beliefs, interpretational biases and current depression symptoms are likely to be complex and bidirectional, our ability to detect and disentangle such influences in existing data may hampered by 'contamination' of self-report measures by similarity to low mood symptoms themselves (Hundt et al., 2013; Reiter et al., 2021; Lorenzo-Luaces, 2023).

Further, theories of cognitive restructuring suggest it is a process based on *learning* (Moutoussis et al., 2018). It has been proposed that individual differences in learning and memory of therapy content may be a key moderator of symptom change during treatment (Harvey et al., 2014; Bruijniks et al., 2019). Inspired by recent demonstrations that clinically-relevant inference processes can be reliably measured using computerised learning tasks (Dorfman et al., 2019; Hopkins et al., 2021), here we sought to explore whether ability to recognise and learn about different attributions during a learning task was related to the subsequent changes in causal attribution tendencies, in the absence or presence of a brief cognitive restructuring intervention.

Contrary to our expectations, we found little evidence that individual differences in learning on the learning task were specifically related to change in attribution tendencies following the restructuring intervention. Rather, we found robust evidence to support the idea that completion of the learning task had additive effects to completion of either intervention condition, in particular in boosting shifts towards self-enhancing (internal and global) attributions of positive events. Across studies, the magnitude of these effects were related to how quickly participants were able to update their choices according to reinforcement of an (implicit) internal-global response dimension on the learning task. Participants with faster learning rate estimates also showed greater ability to explicitly label correct responses along these ground truth dimensions, suggesting better overall understanding of the task state-space. Together, this suggests that individuals for whom understanding these of dimensions is more intuitive may be most likely to respond to this kind of training.

Several previous studies which have attempted to shift appraisals of everyday events using online task-based training. For example, a single session of app-based reappraisal training (ambiguous story vignettes with partial positive or negative completion prompts) was found to result in maladaptive response biases to ambiguous imagined scenarios in individuals given negative training, and adaptive biases in individuals given positive training, in non-clinically anxious or depressed participants (Woud et al., 2013). Similarly, three weeks of online training (user-reported current life stresses with peer-based feedback on reinterpretation) was found to increase self-reported reappraisal skill use – with participants who reported low levels of reappraisal use at baseline benefiting more in terms of improvement in depression symptoms (Morris et al., 2015). A recent network-based meta analysis concluded that, compared to other task-based approaches such as attentional bias modification, cognitive bias training (where participants are typically presented with ambiguous everyday scenarios and trained to resolve them to favour of neutral or positive interpretations) significantly reduced symptoms of anxiety compared to sham (active control) training, and reduced symptoms of both and anxiety and depression compared to waitlist control (Fodor et al. 2020).

Novel aspects of the learning task described here is the use of third-person perspective, alongside explicit reinforcement. It is possible that this is an effective strategy in helping participants learn to recognise different kinds of causal attribution tendencies, since distancing techniques are often employed during cognitive restructuring (for example, “what would you think about the reasons behind this event would be it happened to a friend?”) (Wisco and Nolen-Hoeksema, 2010). One advantage of tasks that are able to measure attribution biases along multiple dimensions – in conjunction with interpretable computational models – is that this information could potentially be fed back to users over time (for example, “compared to your peers, we have noticed that you tend to attribute negative events more to overly-general causes”). Future studies could explore the impact of this kind of informed training on learning speed, self-relevant attribution, and symptom change, as a form of acute psychological treatment augmentation (Nord et al., 2023). Finally, these findings raise the question of further considering the importance of training targeted at increasing self-enhancing attributions, as well as decreasing depressogenic attributions, during cognitive therapy. It has previously been highlighted that learning to adopt self-enhancing attributions may be psychological protective (as it is reliably observed in healthy individuals), and represents a potential resilience mechanism against relapse of low mood in the face of stressful life events (e.g., Alloy et al. 2011).

The major limitation of the studies presented here is that participation was not restricted to individuals currently experiencing clinically-significant levels of psychological symptoms, and that the brief restructuring intervention used here does not represent a real-world (proven to be effective) psychological treatment component. It will be important to test in future work whether findings extend to these settings. However, measuring the impact of isolated therapy components on their proposed cognitive mechanisms in experimental settings has been proposed to be a useful first step in understanding how and when psychotherapeutic techniques result in meaningful clinical improvement (Bruijniks et al., 2018; Huibers et al., 2021). There are also some methodological limitations of our results. Although inference procedures for the for the learning task model were well-calibrated and parameter recovery was adequate, we do not provide empirical data that directly speaks to the test-retest reliability of this measures – something which can be harder to establish for learning compared to choice-based tasks due to meta-learning or practice effects and may require modifications to task structure (see e.g., Zorowitz et al. 2023a). This may limit our ability to infer reliable individual differences in learning between participants. It is also possible that our single-session online experimental design, whilst supporting fast and high-throughput measurement in a relevant sample of individuals for internet-based cognitive treatments, may result in increased demand effects (e.g., participants updating their responses on the second choice task in line with perceived purpose of the study). Future work should establish whether effects observed here are evident over a longer time-scale, and if they generalise to less close cognitive measures – including interpretations of the causes of events in users’ own lives.

A fundamental aim of this kind of research is to help address barriers to the uptake and use of existing psychological interventions – in particular in the case of remotely-delivered treatments where the potential for impact is large, but where initializing engagement and high attrition rates are acute problems (Graham et al., 2019). One factor that has been identified by users of digital mental health products is a “need..to experience a sense of ‘self’ in the treatment” (Knowles et al., 2014). It is possible that using cognitive tasks with interpretable model-based output, and, critically, feeding this information back to users may help address this need. For example, information about relative performance on learning or choice-based attribution measures could be used to suggest to users that they may wish to engage with cognitive restructuring or reappraisal training treatment components earlier in their treatment course (a ‘capitalisation’ approach), or, alternatively, may wish to learn more about these kinds of cognitive biases prior to treatment practice (a ‘compensation’ approach) (Cheavens et al., 2012). The utility of these approaches needs to be established in empirical studies, ideally incorporating participation from relevant stakeholders. Promisingly, e-mental health applications offer the potential to test these questions directly and at scale in an agile way, which may help substantially reduce the time between development and implementation of new treatment strategies (Seiferth et al., 2023).

Methods

Data and code availability statement

Code for implementing all tasks and analyses described here, alongside anonymized study data is available at the the study github repository.

Ethical approval

All participants gave written informed consent and all studies were approved by the UCL Research Ethics Committee (project ID 21029/001).

Participants

For all studies, participants were recruited from an online research participation platform (Prolific), and were required to be based in the UK, 18-65 years old, and fluent in English.

Power analysis

Power analysis for study 1 was based on pilot data concerning the effects of brief cognitive restructuring on proportionate choice of internal-negative attributions (see Norbury et al. 2023 for full details), which determined that we could replicate an effect half the pilot data effect size ($d = 0.48$) in *N*=48 participants with 95% power (repeated-measures ANOVA between-within interaction with 2 groups, 2 measures per group, assuming 0.6 correlation across repeated-measures and alpha=0.05, G*Power 3.1; Faul et al. 2007). Given the relative ease of online data collection, subsequent studies were super-powered to *N*=100 per sample. The data analysed here are the combined initial discovery and replication samples from Norbury et al. 2023, yielding a final *N* of 200.

Power analysis for study 2 was based on the observed simple correlation between mean posterior positive learning rates from the learning training task and change in internal attributions of positive events in study 1 data (all participants $R = 0.27$). Analysis using G*Power 3.1 revealed that *N* of 111 would allow us to replicate an association of this size with 90% power (point biserial model, one-tailed, alpha=0.05). Given that only 2/3 of participants in the proposed study design would complete the learning training task, the target *N* was set to 165 (approximately 55 participants per study arm).

Study design

The design of each study is described in Figure 1a. Upon recruitment to each study, participants were assigned to each randomisation arm using a random number generation-based procedure (ratios 1:1 for study 1, and 1:1:1 for study 2). All studies took place in a remote (online) setting over a single session, which lasted approximately 1 hour.

Measures

Code for implementing each task is available here. All tasks was coded in javascript using the jsPsych library, version 7.2.1 (de Leeuw, 2015).

Causal attribution task

The causal attribution task was as described in Norbury et al. 2023. For full details of task development, design optimization, and measurement properties please see this paper. Of note, output parameters from the associated analysis model have excellent identifiability and test-retest reliability (posterior mean R=0.82-0.90), and have previously been found to be associated with self-reported negative self-beliefs and current depression symptom severity (e.g., Dysfunctional Attitudes Scale and PHQ2 total scores and internal-negative attribution tendency, R=0.35, R=0.26).

Briefly, participants were instructed to imagine themselves in various everyday situations. For each situation, they were asked to picture the situation described as clearly as they could ("as if the events were happening to them right now"), and then choose which of several possible explanations listed below they thought most likely, if it had actually happened to them.

Participants were presented with 32 event scenarios (16 positive and 16 negative events, randomly interleaved), divided into two blocks. Event scenarios were drawn from interpersonal (e.g., ”Someone you are close to tells you that they admire you”), professional/academic (”You and your friends do a general knowledge quiz and you get the lowest score”), and general life-functioning domains (”You fix something around the house that you have been meaning to get done for a while”). For each event, participants were asked to choose between four response options that varied orthogonally in terms of internal-external and global-specific explanation types, derived from examples provided in Abramson et al. 1978. For example, for the event ”You find out that someone you consider a friend has talked about you negatively behind your back”, possible explanations were ”Deep down, my friends don't really like me” (internal-global), ”I probably did something recently to annoy them” (internal-specific), ”Everyone has bad things said about them sometimes” (external-global), and ”My friend was probably just in a bad mood and letting off steam” (external-specific).

Learning training task

The learning training task was developed as a measure of how easily participants are able to learn to select different kinds of causal attributions, in a reinforced setting. In order to differentiate from the causal attribution task (which aims to measure how participants tend to think about the causes behind events, if they actually happened to them, the learning training task used a third-person framing.

Specifically, participants were informed that ”some researchers believe that the the kinds of explanations we think are most likely for events can vary, depending on our moods". They were told that that they would be learning about how a hypothetical person in a particular mood might reason about the causes behind events, across three different scenarios, which represent different kinds of mood that person may be in. For each scenario or mood, it was their job to learn to select the correct kinds of explanations for that person in that mood, which they had to learn via trial and error. Given that the visual format of of the task was somewhat similar to the causal attribution task, participants were provided with explicit instructions prior to each task stressing the differences between tasks, and required to pass a multiple-choice post-instructions quiz before proceeding to each task.

Participants completed three blocks of 20 trials, which they were told represented three different mood state scenarios. Each trial consisted of a description of an everyday event, with event descriptions and different potential causal explanations drawn from the battery of items tested during the development of the causal attribution task (but not included in the final causal attribution task; see Norbury et al. 2023). Across each scenario, events were balanced in terms of positive and negative valence, and whether they concerned interpersonal interactions. Transition between each scenario was signalled by a message stating they were about to encounter a new scenario (where the kinds of reasons thought to be correct may be different to the previous scenario), and a change in screen background colour.

Since we were primarily interested in how quickly participants were able to learn to select self-enhancing causal explanations of positive events and avoid unhelpful explanations of negative events (the goal of cognitive restructuring), the 'correct' (reinforced) attributions for events in each scenario were always internal-global explanations for positive events, and non internal-global for negative events. Response options (potential causal explanations) are unique on each trial, and opposite contingencies are required for positive and negative events, making the task relatively hard. On the basis of pilot testing, it was determined that two response options per trial and a deterministic reinforcement structure (i.e., 100% reinforcement of correct choices) was required to make the task solveable for participants. Specifically, response options (left-right randomized on every trial) were internal-global *vs* internal-specific explanations for scenario 1, internal-global *vs* external-global explanations for scenario 2, and internal-global *vs* external-specific explanations for scenario 3 (see Figure S2a,c). The former explanations were always correct for positive events, and the latter explanations always correct for negative events. On each trial, after participants chose an option, their choice was highlighted, visual feedback given as to whether that choice was correct or not, and the correct response option highlighted in green text.

Given that solving the task requires understanding that response options on each trial can vary according to internal-external and global-specific dimensions, we sought to orient all participants to these aspects of the task state-space at the start of the task. Specifically, before starting the task, participants were asked to think about something negative and positive that happened to them over the last few weeks, and think about the main reason they thought that event happened. They were then asked to rate that reason on slider scales ranging from [caused] "completely by myself"..."completely by other people or circumstances" (internal-external dimension) and [caused] "by things that affect all areas of my life"..."by things related to the specific circumstances" (global-specific dimension) (Figure S3a,c). After each scenario, participants were asked to complete the same ratings for the explanations that were thought to be correct in each scenario (again, separately for positive and negative events). Finally, after each scenario, participants were also asked to provide a brief free-text description of the kinds of causes that were correct in that scenario (”Please describe your general impression of the correct reasons for negative/positive events during the previous scenario”) (Figure S3b,d).

In order to maintain sustained attention on the task in a remote setting, a maximum response time of 15s was applied to each trial. If this was exceeded, participants saw a time-out message, and the trial was repeated. Participants were informed that submissions with either a high percentage of timed-out trials (>10%) or very short average choice times (<1s) may be rejected, since completing the task required sustained attention and sufficient time to read the information on each trial. In order to motivate performance, participants were also paid a small bonus depending on the number of correct responses over the course of the task.

Control learning task

The control learning task was exactly matched in trial type and reinforcement structure to the causal learning task. Participants were informed that ”some researchers believe how quickly we learn about things can differ, according to our mood”. They would see a series of different coloured and shaped baskets, below which would be two different objects that could potentially belong to them. For each scenario, it was their job to learn which kinds of objects belonged in each basket, by trial and error.

Response options (objects) were again trial unique, with opposite reinforcement contingencies depending on trial ’valence’ (here, basket shape/colour). Response options varied along the dimensions human-made - natural and smaller - bigger than a shoebox. Response option stimuli were images drawn from a previously-published database of object images for psychological experiments, which has specifically validated all images along these specific dimensions using the Object Memorability Image Normed Database Software (O-MINDS) v0.1.5 (Duncan Lab, 2022). O-MINDS generates low-variance stimulus sets with images that are approximately matched for human-rated memorability, nameability, and emotionality. Specifically, response options (left-right randomized on every trial) were natural-smaller (than a shoebox) *vs* natural-bigger objects for scenario 1, natural-smaller *vs* humanmade-smaller objects for scenario 2, and natural-smaller *vs* objects humanmade-bigger for scenario 3. The former types of objects were always correct for red baskets, and the latter types of objects always correct for blue baskets.

Participants were asked to provide explicit slider ratings (along the relevant response dimensions) for example objects at the start of the task, to orient them to the task state space. Again, participants provided explicit ratings and free-text descriptions of objects that belonged in each type of basket at the end of each scenario. All other aspects of task design were identical to the learning training task.

Brief cognitive restructuring and control interventions

The brief cognitive restructuring and control interventions were in the form of a series of interactive worksheets, requiring participants to select answers from multiple potential options during worked examples, and provide input based on recent positive and negative experiences from their own lives.

The cognitive restructuring intervention was based on cognitive therapy materials (Beck et al., 1987), and consisted of information about a cognitive model of mood (link between interpretations of events and feelings), interactive exercises identifying helpful and unhelpful attributions of the same events, inviting people to practise generating alternative explanations for recent events in their own lives, and a summary comprehension quiz. The control intervention was based on materials from emotion-focused therapy (Greenberg, 2015), and was closely matched in terms of length, interactivity, and self-relevant exercise content – although, importantly, it did not contain reference to cognitive interpretations influencing feelings or include reappraisal activities (e.g., reflection on whether a particular emotional reaction is helpful or not). The full content of each intervention is available here.

Self-reported demographic and clinical information

At the end of each study, participants completed a set of brief self-report measures to provide information about their recent experience of mental health symptoms, and other sociodemographic information. Symptoms of low mood were measured using The 9-item Patient Health Questionnaire (PHQ9) (Kroenke et al., 2001). A brief measure of social anxiety symptoms, the 3-item Social Phobia Inventory (miniSPIN) (Connor et al., 2001), was also included given our previous observations that social anxiety is relatively elevated in online research participation samples. The Dysfunctional Attitudes Scale, short form, (DAS), a measure of negative self-beliefs observed in some depressed people (Beevers et al., 2007), was included as it has previously been shown to be sensitive to cognitive treatment of low mood (Cristea et al., 2015).

The demographic measure included questions about participant gender identity, age, neurodivergence (defined as "a term for when someone processes or learns information in a different way to that which is considered 'typical': common examples include autism and ADHD"), previous treatment for a mental health problem, disability across World Health Organization Disability Assessment 2.0 domains of functioning (World Health Organization, 2012), and financial, housing, and employment status (as per Buckman et al. 2022). All self-report batteries included two infrequency items (in which some responses are logically invalid or highly improbable), in order to detect potential inattentive responding Zorowitz et al. 2023b.

Analysis

All analyses were carried out in R version 4.1.2 (The R Foundation for Statistical Computing, 2021), using RStudio version 2022.02.0 (RStudio, PBC, 2022).

Initial statistical analysis of learning task data

Preliminary statistical analysis of learning task data was via mixed-effects linear regression models, as implemented in `lme4` (Bates et al., 2015). Specifically, choice accuracy (whether the chosen response option was correct or not) and choice reaction times (RTs) were modelled as:

$$accuracy \sim trialWithinBlock * eventValence * scenarioNo + (1|subID) \tag{1}$$

$$RT \sim trialWithinBlock * eventValence * scenarioNo + (1|subID) \tag{2}$$

Explicit ratings scale data and classification label probabilities for free text data (see below) were modelled as:

$$value \sim eventValence * scenarioNo + (1|subID) \tag{3}$$

Classification of learning task free-text data

In order to measure how well participants were able to describe the ground-truth causes in each scenario in their own words, free-text responses were passed to a zero-shot NLP classification pipeline. Specifically we used Facebook's BART-MNLI-LARGE transformer model (Hugging Face, 2023), with the non mutually-exclusive candidate labels [”myself”, ”other people”, ”in general”, ”specific situations”]. Output probabilities for each candidate label were then passed to further analysis as above.

Hierarchical Bayesian modelling

General methods. Model evaluation and fit procedures were carried out according to Bayesian workflow recommendations (Gelman et al., 2020), with results of Bayesian analyses reported in accordance with recent guidelines (Kruschke, 2021). Model parameters were estimated using Markov-Chain Monte Carlo (MCMC) sampling as implemented in Stan 2.21.0 (Carpenter et al., 2017), using RStan 2.21.3 (Stan Development Team, 2021). MCMC chains were initiated with random starting values, and posterior distributions were formed using 4 chains of 2000 iterations, with 1000 discarded warm-up samples (i.e., 4000 kept iterations per model). Convergence of sampling chains was assessed via inspection of trace plots and Gelman-Rubin ($\hat{R}$) statistics for each parameter (Gelman and Rubin, 1992). All models used generic weakly-informative priors (see Supplementary Methods for details).

Different models of the same data were compared using a cross-validation procedure suitable for hierarchical Bayesian models, which guards against over-fitting by comparing predictive accuracy in left-out samples. Specfcally, models were compared in terms of expected log pointwise predictive density ($ELPD_{diff}$) using the R package `loo` (Vehtari et al., 2017).

For experimental effects of interest, parameters were assessed using 90% credible intervals (CIs), with a 90% CI excluding zero interpreted as representing evidence for a meaningful contribution to posterior parameter estimates (McElreath, 2016). Distributions of posterior parameter estimates and CIs were visualized using the R package `tidybayes` (Kay, 2022).

Hierarchical Bayesian analysis of causal attribution task data

Modelling of causal attribution task data was as previously described in Norbury et al. 2023, using an analysis model for which task design was previously optimized. Specifically, participants' choices on each trial were coded along two dimensions, according to whether an internal (*vs* external) or global (*vs* specific) response option was chosen ($y\_internal$ and y_global, respectively), with the resulting data analysed within a single hierarchical model with 4 free participant-level parameters:

$$\begin{aligned} y\_internal_{p,t,v} &\sim Bern(\theta_{internal,p,t,v}) \\ y\_global_{p,t,v} &\sim Bern(\theta_{global,p,t,v}) \end{aligned} \tag{4}$$

where $\theta_{internal,p,t,v}$ and $\theta_{global,p,t,v}$ represent the latent traits governing a participant (*p*)'s tendency to make an internal or global attribution at that time point (*t*), separately for positively and negatively valenced (*v*) event scenarios.

As previously, data from the two task time-points (pre- and post-intervention) were fit using a single hierarchical model, with separate group means for each parameter at each time-point, and individual parameter estimates at each time-point assumed to be drawn from a multivariate normal distribution, given a uniform prior over $[-1, 1]$ on correlation of individual parameter values across time-points. Also as previously (given evidence of correlations between individuals' tendencies to make global and internal attributions for positive and negative events), we assumed that individual tendencies to make internal and global attributions for each type of event within a given time-point were drawn from a multivariate normal distribution:

$$\begin{aligned}
\begin{bmatrix} \theta_{internal,1,neg} \\ \theta_{global,1,neg} \\ \theta_{internal,2,neg} \\ \theta_{global,2,neg} \end{bmatrix} &\sim MVNormal\left(\begin{bmatrix} \theta_{internal,\mu,1,neg} \\ \theta_{global,\mu,1,neg} \\ \theta_{internal,\mu,2,neg} \\ \theta_{global,\mu,2,neg} \end{bmatrix}, \sigma_{\theta,neg} \right) \\
\begin{bmatrix} \theta_{internal,1,pos} \\ \theta_{global,1,pos} \\ \theta_{internal,2,pos} \\ \theta_{global,2,pos} \end{bmatrix} &\sim MVNormal\left(\begin{bmatrix} \theta_{internal,\mu,1,pos} \\ \theta_{global,\mu,1,pos} \\ \theta_{internal,\mu,2,pos} \\ \theta_{global,\mu,2,pos} \end{bmatrix}, \sigma_{\theta,pos} \right)
\end{aligned} \tag{5}$$

where $\theta_{internal,\mu,t,v}$ and $\theta_{internal,\mu,t,v}$ are the group-level means for each parameter and time-point (modelled separately for positive, *pos*, and negative, *neg*, events), and σ is the covariance between individual-level parameters across attribution types and time points. For full descriptions of parameter constrains and model priors see Supplementary Methods.

Participant-level parameter estimates were constructed using non-centered reparameterization in order to separate the hierarchical parameters and lower-level parameters in the prior (Papaspiliopoulos et al., 2007). For each parameter (e.g., ϕ) and time point *t*, participant-level estimates ($\phi_{p,t}$) were constructed from a group mean ($\phi_{\mu,t}$) and an individual offset ($\tilde{\phi}_{p,t}$). The between-subjects effects of intervention group were then modelled as:

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2}, & \text{if control intervetion + learning task} \end{cases}
\end{aligned} \tag{6}$$

where ϕ_{CR} is a group-level parameter describing potential effects of allocation to the CR intervention on parameter estimates at time 2. For all models, the priors for effects of intervention conditions on parameter estimates were centred on 0 (e.g., $\phi_{CR} \sim N(0,1)$).

For study 2, this model was augmented to include potential group-level effects of allocation the learning training task condition (ϕ_{LEARN}):

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \phi_{LEARN}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{LEARN}, & \text{if control intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR}, & \text{if CR intervention + control learning task} \end{cases}
\end{aligned} \tag{7}$$

Hierarchical Bayesian analysis of learning task data

For model-based analysis of learning task data, choices were collapsed to binary selection of internal-global and non internal-global responses, separately for positive and negative events, in order to allow for repeat assessment of learning across the three task scenarios. Choice data were then modelled using a series of simple reinforcement-learning models based on the

Rescorla-Wagner algorithm. In this framework, values of each response option (internal-global and non-internal global explanations) in each state (for a positively or negatively valence event) are updated on each trial using a surprise term, which is simply the difference between trial feedback (correct or incorrect) and the previously estimated value for that option in that state, multiplied by a learning rate.

Model comparison. In order to determine the best model of task performance, several candidate models of study 1 learning task data were compared in terms of predictive accuracy in left-out data (see above). Specifically, a base model, with a single learning rate parameter, and where choice values were reset at the start of the each scenario (in line with task instructions that different kinds of explanations may be correct in each scenario), was compared to a set of related models, where learning rates and initial starting values were allowed to vary between valence conditions and between first and subsequent scenarios, motivated by features of the pilot and study 1 datasets (for full details see Table S8). All compared models used a softmax observation function to link values to observed choices, with a single free parameter governing inverse temperature (degree of value-drivenness) of this function (see below).

Three models with separate learning rates for positive and negative events, as well as individual-level free parameters governing starting values for internal-global attributions of positive and negative events, performed similarly well (difference in expected log pointwise predictive density less than 5x than the standard error of the estimate; Vehtari et al. 2017; Table S8). Of these, the model with superior parameter recovery according to simulation-based calibration analysis was taken forward for further analysis (see below). Re-running analyses with the alternate ('winning') model produced a very similar pattern of results to those reported below, with all reported main effects surviving.

For all subsequent analyses, learning task data were analysed as:

$$Q_{v,c,t} = Q_{v,c,t} + \alpha_{v,p} * (outcome_{p,t} - Q_{v,c,t}) \tag{8}$$

where $Q_{v,c,t}$ is the value of each choice (*c*) for each event valence (*v*) on trial *t*, $alpha_{v,p}$ is the learning rate parameter for each participant (*p*) for each event valence (i.e., α_{pos}, α_{neg}), and the outcome for that trial is either correct (1) or incorrect (0).

Starting values of (initial bias towards or away from) internal-global explanations for each event valence were set to separate free parameters for the start of the first scenario (individual initial starting bias) and the second and third scenario (representing degree of expectation reset for each participant at the start of subsequent scenarios). Q values were assumed to map onto observed choice data (y) using a softmax likelihood function with inverse temperature parameter β:

$$y_{p,t} \sim categorical\_logit(\beta_p * [Q_{v,:,t}]); \tag{9}$$

As both learning training and control learning tasks had identical trial type and reinforcement structure, and in order to facilitate joint analysis, the same model identified above was applied to both learning and control learning task data in study 2. Since linear-mixed effects analysis indicated some differences in the form of learning between tasks (in both overall speed of learning and starting biases; see Figure S2, Supplementary Results), different group-level mean and variance parameters were specified between tasks types (governing all individual-level parameters, except the inverse temperature parameter β). Formal model comparison confirmed that a model with separate group means for different task versions had better predictive accuracy than a model with single group means ($ELPD_{diff}$=-124.8, se=14.6).

Simulation-based calibration analysis. Simulation-based calibration (SBC) analysis was used to validate inference procedures for the learning task models (Talts et al., 2020). Briefly, this involves generating draws from the prior predictive distribution of the generative model (creating N simulated datasets), then fitting the model to each simulated dataset and obtaining D independent draws from the model posterior. For each parameter of interest, the rank of the simulated value within the posterior draws is then calculated. If the data generation and inference procedure works as expected, then the resulting ranks should be uniformly distributed across $[0, D]$ (Modrák et al., 2022). Here, we generated N=150 datasets based on independent draws from the prior distributions of each parameter, which were specified generously based on the empirical posterior estimates of parameter distributions observed in pilot data. We then took D=2000 posterior draws (after discarding 1000 warm-up samples), across two sampling chains. Graphical summaries of SBC results were generated using the R package SBC (Kim et al., 2023), and are available for the chosen learning model in Figure S9.

Model performance. Two model-agnostic 'goodness-of-fit' measures are reported. Posterior predictive accuracy was calculated as the match between replicated choice data generated stochastically from posterior parameter estimates and task trial arrays, and the observed data from each participant. Pseudo-r^2 statistics reflect the amount of variance explained by the model relative to a model of pure chance (Daw, 2011).

Associating separately-modelled causal attribution and learning task data parameters

As simple first-pass check of association between parameter estimates from the causal attribution and learning training task, we examined correlations between point estimates (posterior means) of each parameter, weighted by the posterior precision of the predictor variable (i.e., 1/posterior SD α_{pos}. This is not an optimal way to test for associations between different estimates, since it neglects information about the individual precision of both parameter estimates.

Joint modelling of causal attribution and learning task data

In order to formally test for associations between parameters, we constructed a series of joint models of causal attribution and learning task data. Joint modelling allows maximum use of participant-level data, whilst more appropriately retaining information about the uncertainty or precision of each kind of measurement (Turner et al., 2017; Haines, 2021).

For the first joint models, the causal attribution task analysis model was extended such that individual estimates for positive learning rates from the learning task (α_{pos}) were allowed to influence relevant post-intervention (time 2) causal attribution task parameter estimates via the inclusion of β weight parameters (β_{LEARN}; see Haines et al. 2020; Hopkins et al. 2021 for previous examples of this approach). These β weights can interpreted similarly as in a standard regression model, with the group-level intervention effects (e.g., ϕ_{CR}) now representing the intercept (see below).

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \beta_{LEARN} * \alpha_{pos}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \beta_{LEARN} * \alpha_{pos} & \text{if control intervention + learning task} \end{cases} \tag{10}$$

For study 2 data, the first joint model included separate β weights for participants who completed the learning training *vs* control learning tasks (β_{LEARN}, $\beta_{CONTROL}$):

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \phi_{LEARN} + \beta_{LEARN} * \alpha_{pos}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{LEARN} + \beta_{LEARN} * \alpha_{pos}, & \text{if control intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \beta_{CONTROL} * \alpha_{pos} & \text{if CR intervention + control learning task} \end{cases} \tag{11}$$

The second joint models added additional β weights for participants randomized to complete the cognitive restructuring intervention ($\beta_{LEARN+CR}$), in order to test for the presence of larger influences of learning rates on time-2 attribution changes in participants who received both learning training and brief cognitive restructuring.

For study 1:

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \\ \beta_{LEARN} * \alpha_{pos} + \beta_{LEARN+CR} * \alpha_{pos}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \beta_{LEARN} * \alpha_{pos} & \text{if control intervention + learning task} \end{cases} \tag{12}$$

For study 2:

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \phi_{LEARN} + \\ \beta_{LEARN} * \alpha_{pos} + \beta_{LEARN+CR} * \alpha_{pos}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{LEARN} + \beta_{LEARN} * \alpha_{pos}, & \text{if control intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{CR} + \beta_{CONTROL} * \alpha_{pos} & \text{if CR intervention + control learning task} \end{cases} \tag{13}$$

For all joint models, the priors for β effects were centred on zero (e.g., $\beta_{LEARN} \sim N(0,1)$). Posterior estimates for β weights with a 90% credible interval that excluded zero were taken as evidence that learning rate estimates were meaningfully informative with respect to post-intervention changes in causal attribution tendencies.

Acknowledgements

This research was funded by a research grant from Koa Health to QJMH, SF and RD and a Wellcome Trust grant to QJMH (221826/Z/20/Z). We acknowledge support by the UCLH NIHR BRC. TUH is supported by a Sir Henry Dale Fellowship (211155/Z/18/Z; 211155/Z/18/B; 224051/Z/21) from Wellcome and The Royal Society. QJMH acknowledges grant funding from the Wellcome Trust (221826/Z/20/Z). 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.

AN declares no competing interests. QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience. TUH has obtained fees and options for consultancies for Limbic Ltd.

References

Abramson, L. Y., Seligman, M. E., and Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 87:49–74.

Alloy, L. B., Wagner, C. A., Black, S. K., Gerstein, R. K., and Abramson, L. Y. (2011). The breakdown of self-enhancing and self-protecting cognitive biases in depression. In *Handbook of Self-enhancement and Self-protection*, pages 358–379. Guilford Press.

Bartucz, M. B., David, D. O., and Matu, S. A. (2022). Cognitive vulnerabilities and Depression: A Culture-Moderated Meta-Analysis. *Cognitive Therapy and Research*, 46(3):502–516.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Beck, A. T., Rush, A. J., Shaw, B. F., and Emery, G. (1987). *Cognitive Therapy of Depression*. Guildford Press.

Beevers, C. G., Strong, D. R., Meyer, B., Pilkonis, P. A., and Miller, I. W. (2007). Efficiently assessing negative cognition in depression: An item response theory analysis of the Dysfunctional Attitude Scale. *Psychological Assessment*, 19:199–209.

Bruijniks, S. J. E., DeRubeis, R. J., Hollon, S. D., and Huibers, M. J. H. (2019). The Potential Role of Learning Capacity in Cognitive Behavior Therapy for Depression: A Systematic Review of the Evidence and Future Directions for Improving Therapeutic Learning. *Clinical Psychological Science*, 7(4):668–692.

Bruijniks, S. J. E., Sijbrandij, M., Schlinkert, C., and Huibers, M. J. H. (2018). Isolating therapeutic procedures to investigate mechanisms of change in cognitive behavioral therapy for depression. *Journal of Experimental Psychopathology*, 9(4):2043808718800893.

Buckman, J. E. J., Saunders, R., Stott, J., Cohen, Z. D., Arundell, L.-L., Eley, T. C., Hollon, S. D., Kendrick, T., Ambler, G., Watkins, E., Gilbody, S., Kessler, D., Wiles, N., Richards, D., Brabyn, S., Littlewood, E., DeRubeis, R. J., Lewis, G., and Pilling, S. (2022). Socioeconomic Indicators of Treatment Prognosis for Adults With Depression: A Systematic Review and Individual Patient Data Meta-analysis. *JAMA Psychiatry*, 79(5):406–416.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76:1–32.

Chahar Mahali, S., Beshai, S., Feeney, J. R., and Mishra, S. (2020). Associations of negative cognitions, emotional regulation, and depression symptoms across four continents: International support for the cognitive model of depression. *BMC Psychiatry*, 20(1):18.

Cheavens, J. S., Strunk, D. R., Lazarus, S. A., and Goldstein, L. A. (2012). The compensation and capitalization models: A test of two approaches to individualizing the treatment of depression. *Behaviour Research and Therapy*, 50(11):699–706.

Clark, D. A. (2013). Cognitive Restructuring. In *The Wiley Handbook of Cognitive Behavioral Therapy*, pages 1–22. John Wiley & Sons, Ltd.

Clark, D. A. (2022). Cognitive Reappraisal. *Cognitive and Behavioral Practice*, 29(3):564–566.

Clark, D. A. and Beck, A. T. (2010). Cognitive theory and therapy of anxiety and depression: Convergence with neurobiological findings. *Trends in Cognitive Sciences*, 14(9):418–424.

Connor, K. M., Kobak, K. A., Churchill, L. E., Katzelnick, D., and Davidson, J. R. (2001). Mini-SPIN: A brief screening assessment for generalized social anxiety disorder. *Depression and Anxiety*, 14(2):137–140.

Cristea, I. A., Huibers, M. J. H., David, D., Hollon, S. D., Andersson, G., and Cuijpers, P. (2015). The effects of cognitive behavior therapy for adult depression on dysfunctional thinking: A meta-analysis. *Clinical Psychology Review*, 42:62–71.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Oxford University Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47:1–12.

Dercon, Q., Mehrhof, S. Z., Sandhu, T. R., Hitchcock, C., Lawson, R. P., Pizzagalli, D. A., Dalgleish, T., and Nord, C. L. (2023). A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychological Medicine*, pages 1–11.

Dorfman, H. M., Bhui, R., Hughes, B. L., and Gershman, S. J. (2019). Causal Inference About Good and Bad Outcomes. *Psychological Science*, 30(4):516–525.

Duncan Lab (2022). O-MINDS.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39:175–191.

Fodor, L. A., Georgescu, R., Cuijpers, P., Szamoskozi, , David, D., Furukawa, T. A., and Cristea, I. A. (2020). Efficacy of cognitive bias modification interventions in anxiety and depressive disorders: a systematic review and network meta-analysis. *The Lancet Psychiatry*, 7(6):506–514. Publisher: Elsevier.

Forand, N. R., Barnett, J. G., Strunk, D. R., Hindiyeh, M. U., Feinberg, J. E., and Keefe, J. R. (2018). Efficacy of Guided iCBT for Depression and Mediation of Change by Cognitive Skill Acquisition. *Behavior Therapy*, 49(2):295–307.

Garratt, G., Ingram, R. E., Rand, K. L., and Sawalani, G. (2007). Cognitive Processes in Cognitive Therapy: Evaluation of the Mechanisms of Change in the Treatment of Depression. *Clinical Psychology: Science and Practice*, 14(3):224–239.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian Workflow.

Graham, A. K., Lattie, E. G., and Mohr, D. C. (2019). Experimental Therapeutics for Digital Mental Health. *JAMA Psychiatry*, 76(12):1223–1224.

Greenberg, L. S. (2015). *Emotion-focused therapy: Coaching clients to work through their feelings, 2nd ed.* Emotion-focused therapy: Coaching clients to work through their feelings, 2nd ed. American Psychological Association, Washington, DC, US.

Gumport, N. B., Dong, L., Lee, J. Y., and Harvey, A. G. (2018). Patient learning of treatment contents in cognitive therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, 58:51–59.

Haines, N. (2021). *Integrating Trait and Neurocognitive Mechanisms of Externalizing Psychopathology: A Joint Modeling Framework for Measuring Impulsive Behavior*. PhD thesis, The Ohio State University.

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., Myung, J. I., Turner, B. M., and Ahn, W.-Y. (2020). Anxiety Modulates Preference for Immediate Rewards Among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis. *Clinical Psychological Science*, 8(6):1017–1036.

Harvey, A. G., Lee, J., Williams, J., Hollon, S. D., Walker, M. P., Thompson, M. A., and Smith, R. (2014). Improving Outcome of Psychosocial Treatments by Enhancing Memory and Learning. *Perspectives on Psychological Science*, 9(2):161–179.

Hawley, L. L., Padesky, C. A., Hollon, S. D., Mancuso, E., Laposa, J. M., Brozina, K., and Segal, Z. V. (2017). Cognitive-Behavioral Therapy for Depression Using Mind Over Mood: CBT Skill Use and Differential Symptom Alleviation. *Behavior Therapy*, 48(1):29–44.

Hopkins, A. K., Dolan, R., Button, K. S., and Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry*, 5(1):21–37.

Hugging Face (2023). facebook/bart-large-mnli · Hugging Face.

Huibers, M. J. H., Lorenzo-Luaces, L., Cuijpers, P., and Kazantzis, N. (2021). On the Road to Personalized Psychotherapy: A Research Agenda Based on Cognitive Behavior Therapy for Depression. *Frontiers in Psychiatry*, 11.

Hundt, N. E., Mignogna, J., Underhill, C., and Cully, J. A. (2013). The Relationship Between Use of CBT Skills and Depression Treatment Outcome: A Theoretical and Methodological Review of the Literature. *Behavior Therapy*, 44(1):12–26.

Huys, Q. J. M., Browning, M., Paulus, M. P., and Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, 46(1):3–19.

Kay, M. (2022). ggdist: Visualizations of distributions and uncertainty.

Kazantzis, N., Luong, H. K., Usatoff, A. S., Impala, T., Yew, R. Y., and Hofmann, S. G. (2018). The Processes of Cognitive Behavioral Therapy: A Review of Meta-Analyses. *Cognitive Therapy and Research*, 42(4):349–357.

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research*, 19(4-5):418–428.

Kim, S., Moon, H., Modrák, M., and Säilynoja, T. (2023). SBC: Simulation Based Calibration for rstan/cmdstanr models.

Knowles, S. E., Toms, G., Sanders, C., Bee, P., Lovell, K., Rennick-Egglestone, S., Coyle, D., Kennedy, C. M., Littlewood, E., Kessler, D., Gilbody, S., and Bower, P. (2014). Qualitative Meta-Synthesis of User Experience of Computerised Therapy for Depression and Anxiety. *PLOS ONE*, 9(1):e84323.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613.

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5(10):1282–1291.

Lemmens, L. H. J. M., Galindo-Garre, F., Arntz, A., Peeters, F., Hollon, S. D., DeRubeis, R. J., and Huibers, M. J. H. (2017). Exploring mechanisms of change in cognitive therapy and interpersonal psychotherapy for adult depression. *Behaviour Research and Therapy*, 94:81–92.

Lorenzo-Luaces, L. (2023). Identifying active ingredients in cognitive-behavioral therapies: What if we didn't? *Behaviour Research and Therapy*, 168:104365.

Lorenzo-Luaces, L., German, R. E., and DeRubeis, R. J. (2015). It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, 41:3–15.

Lorenzo-Luaces, L., Keefe, J. R., and DeRubeis, R. J. (2016). Cognitive-Behavioral Therapy:

Nature and Relation to Non-Cognitive Behavioral Therapy. *Behavior Therapy*, 47(6):785–803.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, New York.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., and Hankin, B. L. (2004). Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias. *Psychological Bulletin*, 130:711–747.

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2022). Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity.

Morris, R. R., Schueller, S. M., and Picard, R. W. (2015). Efficacy of a Web-Based, Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression: Randomized Controlled Trial. *Journal of Medical Internet Research*, 17(3):e4167.

Moutoussis, M., Shahar, N., Hauser, T. U., and Dolan, R. J. (2018). Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Computational Psychiatry*, 2(0):50–73.

Norbury, A., Hauser, T. U., Fleming, S., Dolan, R. J., and Huys, Q. (2023). Different components of cognitive-behavioural therapy affect specific cognitive mechanisms.

Nord, C. L., Longley, B., Dercon, Q., Phillips, V., Funk, J., Gormley, S., Knight, R., Smith, A. J., and Dalgleish, T. (2023). A transdiagnostic meta-analysis of acute augmentations to psychological therapy. *Nature Mental Health*, 1(6):389–401.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 22(1):59–73.

Pearson, R. M., Heron, J., Button, K., Bentall, R. P., Fernyhough, C., Mahedy, L., Bowes, L., and Lewis, G. (2015). Cognitive styles and future depressed mood in early adulthood: The importance of global attributions. *Journal of Affective Disorders*, 171:60–67.

Persons, J. B., Marker, C. D., and Bailey, E. N. (2023). Changes in affective and cognitive distortion symptoms of depression are reciprocally related during cognitive behavior therapy. *Behaviour Research and Therapy*, page 104338.

Reiter, A. M., Atiya, N. A., Berwian, I. M., and Huys, Q. J. (2021). Neuro-cognitive processes as mediators of psychological treatment effects. *Current Opinion in Behavioral Sciences*, 38:103–109.

Schmidt, I. D., Pfeifer, B. J., and Strunk, D. R. (2019). Putting the “cognitive” back in cognitive therapy: Sustained cognitive change as a mediator of in-session insights and depressive symptom improvement. *Journal of Consulting and Clinical Psychology*, 87(5):446–456.

Seiferth, C., Vogel, L., Aas, B., Brandhorst, I., Carlbring, P., Conzelmann, A., Esfandiari, N., Finkbeiner, M., Hollmann, K., Lautenbacher, H., Meinzinger, E., Newbold, A., Opitz, A., Renner, T. J., Sander, L. B., Santangelo, P. S., Schoedel, R., Schuller, B., Stachl, C., Terhorst, Y., Torous, J., Wac, K., Werner-Seidler, A., Wolf, S., and Löchner, J. (2023). How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. *Nature Mental Health*, 1(8):542–554.

Strunk, D. R., Hollars, S. N., Adler, A. D., Goldstein, L. A., and Braun, J. D. (2014). Assessing Patients’ Cognitive Therapy Skills: Initial Evaluation of the Competencies of Cognitive Therapy Scale. *Cognitive Therapy and Research*, 38(5):559–569.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat].

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., and Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychol-*

ogy, 76:65–79.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Wisco, B. E. and Nolen-Hoeksema, S. (2010). Interpretation bias and depressive symptoms: The role of self-relevance. *Behaviour Research and Therapy*, 48(11):1113–1122.

World Health Organization (2012). WHO Disability Assessment Schedule (WHODAS 2.0).

Woud, M. L., Postma, P., Holmes, E. A., and Mackintosh, B. (2013). Reducing analogue trauma symptoms by computerized reappraisal training – Considering a cognitive prophylaxis? *Journal of Behavior Therapy and Experimental Psychiatry*, 44(3):312–315.

Zorowitz, S., Karni, G., Paredes, N., Daw, N., and Niv, Y. (2023a). Improving the Reliability of the Pavlovian Go/No-Go Task.

Zorowitz, S., Solis, J., Niv, Y., and Bennett, D. (2023b). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour,* 7:1667–1681.

Supplementary Material

Supplementary Methods

Hierarchical Bayesian modelling of causal attribution task data

Priors for group-level parameter means were specified using standard normal distributions, $\phi_{\mu,s} \sim N(0,1)$. Priors for group-level parameter standard deviations were specified as $\phi_{\sigma,s} \sim cauchy(0,1)$. Priors for individual participant deviations from group-level parameter estimates ($\theta_{internal,p,t,neg}$, $\theta_{internal,p,t,pos}$, $\theta_{global,p,t,neg}$, $\theta_{global,p,t,pos}$) were also specified using standard normal distributions ($\tilde{\phi}_{p,t} \sim N(0,1)$). The prior over the correlation matrix relating parameter estimates across time-points was set to be uniform over $[-1,1]$ using an $LKJ(1)$ prior.

The priors for group-level effects of interventions on parameter estimates at time 2 (ϕ_{CR} and ϕ_{LEARN}), and group-level β weights governing influence of learning rates on effects of interest (β_{LEARN}, $\beta_{CONTROL}$, $\beta_{LEARN+CR}$), were also specified as standard normal distributions (i.e., centred on zero).

Individual parameter estimates for latent traits governing tendency to attribute positive and negative events to internal and global causes were unconstrained but passed to the Bernoulli observation function (eq. 4) using an inverse logit transform, scaling probability of endorsement to the range $[0,1]$ (see e.g., Figure 2).

Hierarchical Bayesian modelling of learning task data

Priors for group-level parameter means were specified using standard normal distributions, $\phi_{\mu,s} \sim N(0,1)$. Priors for group-level parameter standard deviations were specified as $\phi_{\sigma,s} \sim cauchy(0,1)$. Priors for individual participant deviations from group-level parameter estimates (α_{neg}, α_{pos}, β, $q0\_1\_neg$, $q0\_2\_pos$, $q0\_23\_neg$, $q0\_23\_pos$) were also specified using standard normal distributions ($\tilde{\phi}_{p,t} \sim N(0,1)$).

Individual parameter estimates for learning rates (α_{neg}, α_{pos}) were constrained to be in range $[0,1]$, and inverse temperature parameters (β) were constrained to be positive and in the range $[0,20]$.

Supplementary Results

Initial statistical analysis of learning task data

Response accuracy. Choice data for the learning task is shown in Figure S2a,c. Analysis of choice accuracy via mixed-effects linear models showed that, within each scenario, participants were able to learn to select the correct attribution type (main effect of trial number within block on response accuracy, study 1: $F_{1,11793} = 81.7, p < 0.001$, study 2: $F_{1,6365} = 60.3, p < 0.001$), and that this effect was greater for later task scenarios (main effect of scenario number, scenario*trial number interaction, study 1: $F_{1,11793} = 128.8, 8.1, p < 0.005$, study 2: $F_{1,6365} = 83.3, p < 0.001$), suggesting some learning carried over between scenarios. As can be seen in Figure S2, there was also a significant influence of event valence on choice accuracy – with lower overall accuracy and slower learning over the task for positive events (main effect of event valence, valence*trial number interaction, valence*trial*scenario number interaction, study 1: $F_{1,11793} = 245.0, 38.5, 19.6, p < 0.001$, study 2: $F_{1,6365} = 167.6, 22.0, 8.7, p < 0.005$). This suggests that participants found it harder to learn to select self-enhancing (internal-global) attributions of positive events compared to unhelpful (non internal-global) attributions of negative events.

Choice reaction times. This valence asymmetry was also reflected in choice RTs (Figure S2b,d). Overall, participants were slower to choose responses for positive events (main effect of event valence on choice reaction time, study 1: $F_{1,11793} = 8.4, p < 0.005$, study 2: $F_{1,6365} = 4.1, p < 0.05$), although this was mainly evident in the first scenario (valence*trial*scenario number interaction, study 1: $F_{1,11793} = 33.4, p < 0.001$, study 2: $F_{1,6365} = 15.0, p < 0.001$). Choice times indicated maintenance of considered responding over the course of the task (mean RT>4s).

Explicit post-scenario ratings data. Across response dimensions and scenarios, participants were able to recognise that the characteristics of 'correct' causes differed between positive and negative events (main effect of event valence on ratings study 1: $F_{1,2189} = 1091.7, p < 0.001$, study 2: $F_{1,1284} = 638.1, p < 0.001$; Figure S3c), with this knowledge improving over the task (valence*scenario number interaction study 1: $F_{2,2189} = 6.8, p < 0.005$, study 2: $F_{1,1284} = 6.85, p < 0.005$). Both of these effects were of smaller magnitude for the global-specific compared to the internal-external response scale ratings (scale*valence interaction, study 1: $F_{1,2189} = 16.8, p < 0.001$, study 2: $F_{1,1284} = 19.3, p < 0.001$) – suggesting that participants found this response dimension harder to parse.

Free text post-scenario descriptions. A zero-shot natural language classifier (BART-LARGE-MNLI) was also able to distinguish ground truth cause types from participants' free text descriptions of each scenario (Figure S3d). Specifically, there were significant differences in output label probabilities in the expected direction for the internal-external ("myself", "other people") dimension (event valence*label interactions on output scores, study 1: $F_{1,2189} = 434.7, p < 0.001$, study 2: $F_{1,1177} = 301.8, p < 0.001$), with differences in label probabilities increasing over the task (valence*label*scenario number interaction, study 1: $F_{2,2189} = 37.7, p < 0.005$, study 2: $F_{2,1177} = 15.7, p < 0.001$). For global-specific ("in general", "specific situations"), there were significant differences in output label probabilities in the expected direction only in study 1 ($F_{1,2189} = 33.3, p < 0.001$; study 2: $F_{1,1284} = 3.3, p = 0.07$), although in both cases differences in label probabilities increased over the task (valence*label*scenario number interaction, study 1: $F_{2,2189} = 6.3, p < 0.005$, study 2: $F_{2,1284} = 6.5, p < 0.005$).

Relationship between explicit ratings and free text post-scenario descriptions. Explicit ratings and free text classification label probabilities for the internal-external dimension were also weakly correlated with each other (study 1: $Rs = 0.17 - 0.36, p <= 0.01$, study 2: $Rs = 0.26 - 0.48, p < 0.001$; Figure S4), suggesting that these measures were capturing at least partially shared information. Specifically, participants with more accurate post-scenario internal-external explicit cause ratings provided free text descriptions that were more easily classifiable with ground truth cause type labels for this response dimension. For the noisier global-specific dimension, associations were not significant (study 1 and 2: $Rs < 0.14, p > 0.05$).

Differences between causal learning training and control tasks. When choice accuracy data for the causal attribution and control learning tasks were combined in the same model, there was evidence for lower overall accuracy for the control learning task (main effect of task type on response accuracy, $F_{1,5381} = 91.3, p < 0.001$) – likely as performance was not aided by the presence of group-level initial biases towards correct response options (as was the case for the causal task, Figure S2c). Control task participants did not also show a valence asymmetry in response accuracy (for the control task, ‘valence’ represents sorting basket colour/type rather than positive or negative events; task type*valence interaction, $F_{1,9662} = 284.7, p < 0.001$), and did not show slower learning over the task for ‘positive’ events (task type*valence*scenario number, task type*valence*trial number, and task type*valence*trial*scenario number interactions, $F_{1,9662} = 148.1, 76.0, 33.7, p < 0.001$) – suggesting this effect was specific to a reticence to select self-enhancing attributions on the causal learning task.

When choice time data for both tasks were analysed together, choice times were significantly faster for the control learning task (main effect of task type, $F_{1,564} = 269.1, p < 0.001$) - likely reflecting faster processing speed for images compared to text-based stimuli (Figure S2d). Control task participants were also not slower to choose response options for ‘positive’ (red basket) stimuli (task type*valence interaction, $F_{1,9662} = 4.1, p < 0.05$).

Ratings values for the control task were substantially less variable and more extreme (Figure S3c), suggesting that the response dimensions for this task were more explicit and easily parsed by participants (task type*valence interaction in both tasks model, $F_{1,1782} = 33.8, p < 0.001$).

If the free-text responses from the control learning task were classified using the same candidate labels as for the causal learning task (which should not be relevant), output label probabilities were significantly lower across response dimensions (main effect of task type, $F_{1,162} = 139.6, 46.7, p < 0.001$), and were not sensitive to trial ‘valence’ (task type*valence*label interaction, $F_{1,1782} = 102.1, 6.44, p < 0.02$) - suggesting that the classifier results were somewhat specific to the task for which candidate labels represented the ground truth, rather than, for example, picking out general language features not related to task content.

Model-based analysis of learning task data

Model performance. The mean posterior predictive accuracy of the model (agreement between real choices and simulated choice data generated from posterior parameter estimates) in study 1 was 0.88 (SD 0.08), and in study 2 0.88 (SD 0.07). Pseudo-r^2 (ratio of variance explained compared to a random model) was 0.59 in study 1 and 0.57 in study 2.

For study 2 data, model performance was similar when separately considering the likelihood of choice data of participants from either task type (causal learning task, mean posterior predictive accuracy=0.89, pseudo-r^2=0.59; control learning task, mean posterior predictive accuracy=0.86, pseudo-r^2=0.52).

Supplementary Figures

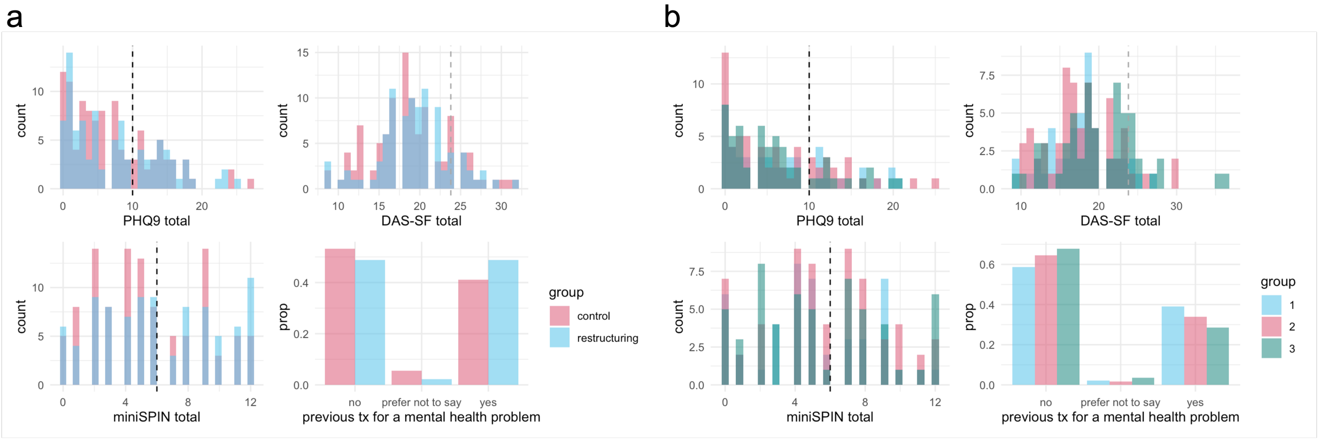

Figure S1: **Distribution of self-reported clinical scores for both studies.** **a** Study 1 participants. The restructuring group represents participants randomized to the cognitive restructuring intervention, with the control group representing participants randomized the control (emotion-focused) intervention. Both groups completed the causal attribution learning task prior to completing the intervention. **b** Study 2 participants. Group 1 represents participants randomized to complete the learning task + cognitive restructuring intervention. Group 2 represents the learning task + control intervention condition. Group 3 represents the control learning task + cognitive restructuring intervention condition. PHQ9 total, Physician's Health Questionnaire 9-item measure of depressed mood total score. miniSPIN total, mini Social Phobia Inventory total score. DAS-SF total, Dysfunctional Attitude Scale short-form total score. Black dotted lines represent previously-published cut-off scores for clinically-significant levels of symptoms. For the DAS-SF, where no such cut-off score is available, grey dotted lines represent mean scores in previously-published samples of depressed inpatients. Participants were also asked if they had ever previously received treatment (tx) for a mental health problem (see Table 1).

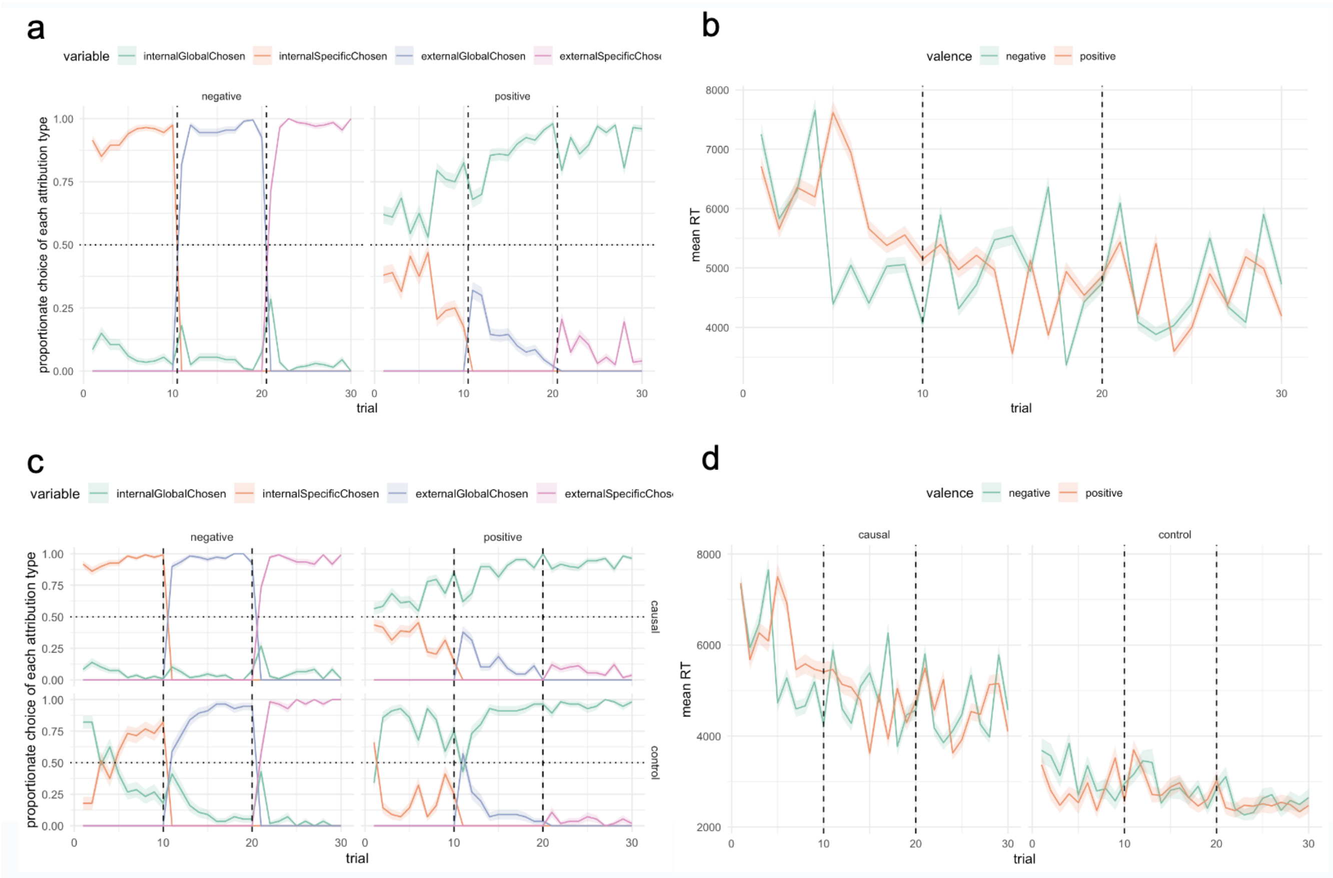

Figure S2: **Choice accuracy and response time data for the learning training tasks.** **a** Study 1 learning task choice accuracy data. Participants were instructed that they would learn about three different scenarios, each of which represented a different kind of mood a person could be in. For each scenario, they had to learn (by trial and error) which kind of explanations for events were thought to be correct for a person in that particular mood. In truth, the correct (reinforced) attributions were always self-enhancing explanations (i.e., internal-global attributions for positive events, and non internal-global attributions for negative events).**b** Choice reaction times during the task, by event valence (in ms). **c** Study 2 learning task choice accuracy data. Here, the top panels represent the same task as in a (the 'causal' learning training task), and bottom panels represent data from the control learning task. In the control learning task, rather than selecting between different causes of events (trial-unique responses that varied according to internal-external and global-specific response directions), participants were asked to choose between images of trial-unique objects that varied according to human made-natural and smaller-larger than a shoebox response dimensions. Trial type and reinforcement structure was identical to the learning training task, with opposite response options reinforced as correct for different coloured/shaped object 'baskets' (analogous to event valence in the causal attribution learning task). **d** Choice reaction times for each learning task, in study 2 participants. Line graphs in all panels represent the mean and standard error of participants' data.

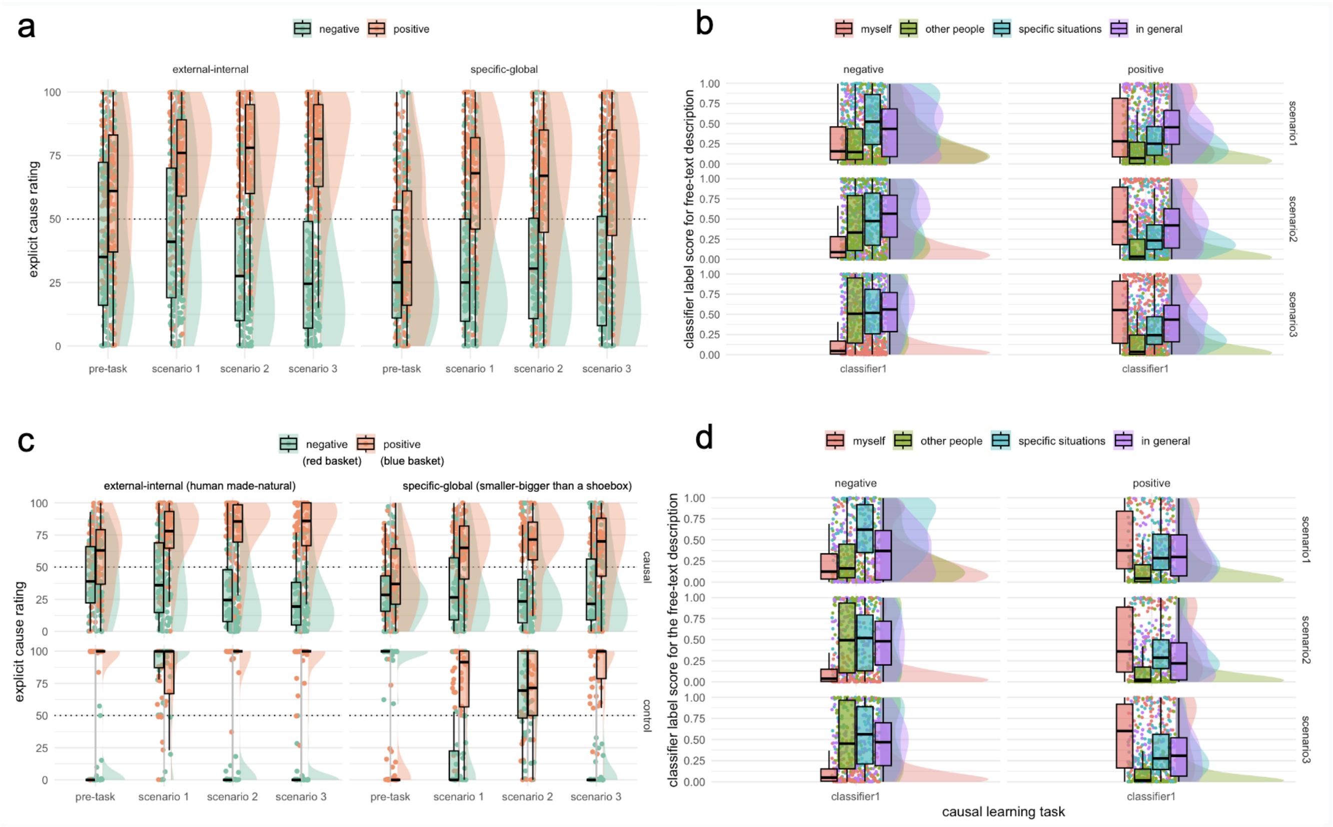

Figure S3: **Explicit ratings and free-text description data from the learning training tasks**. **a** Within-task explicit cause ratings data for study 1. After each scenario, participants were asked to rate the kinds of causes of events that were thought to be correct, along two separate dimensions (external-internal and specific-global). Prior to starting the task, participants were asked to think about a recent positive and negative event from their own lives, and asked to rate the causes of these events along these two dimensions, in order to help familiarise them with the response option state space. **b** Within-task free-text cause description data for study 1. After each scenario, participants were also asked to provide a free-text description of the kinds of causes that were thought to be correct, separately for positive and negative events. This data was passed to a natural language processing algorithm (BART-LARGE-MNLI), which output classification probabilities for the candidate labels [events were caused by] “myself”, “other people” “specific situations”, and “in general” (labels were non mutually-exclusive). In all panels, raincloud plots show individual participant data, summarised by boxplots (median and interquartile range).

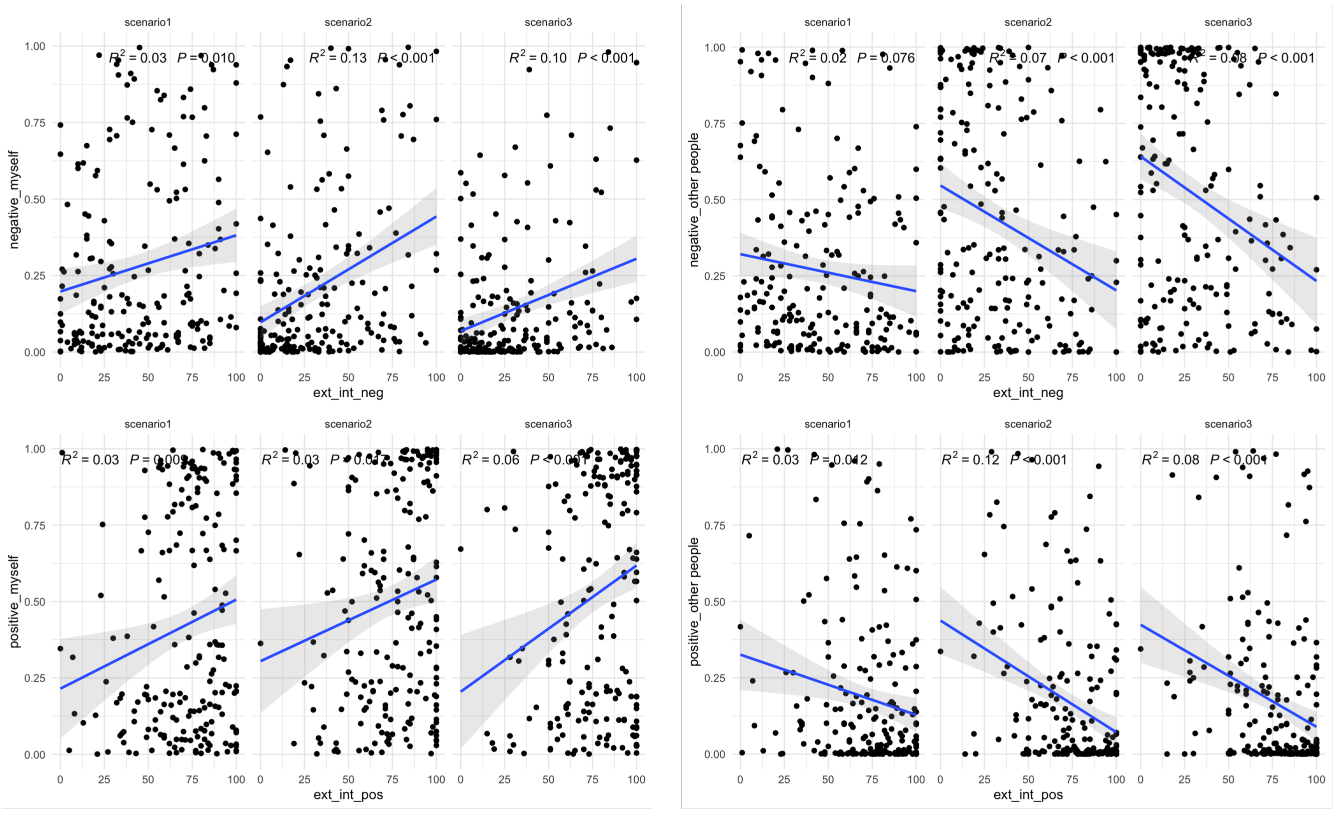

Figure S4: **Associations between explicit cause ratings and classifier label probabilities for free-text descriptions of causes from the learning training task (study 1).** X axes, explicit ratings of cause types following each scenario, on the external-internal response dimension. Y axes, classifier output probabilities for post-scenario free text descriptions of causes, for the labels [caused by] "myself" and [caused by] "other people".

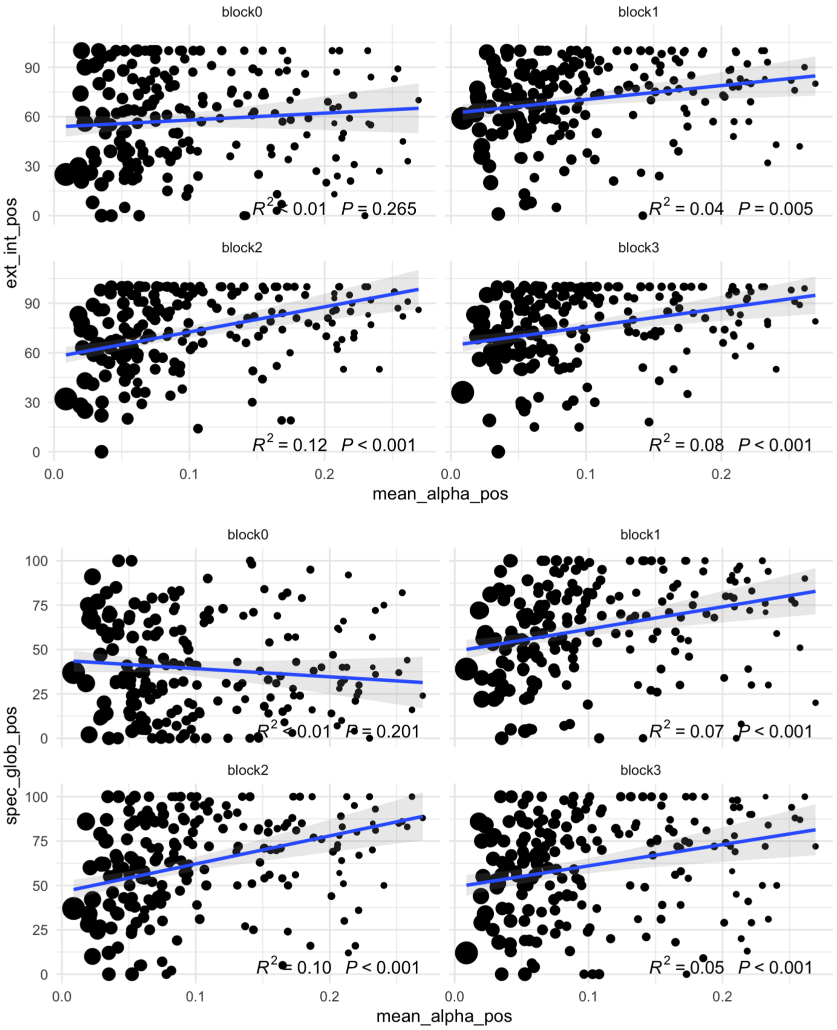

Figure S5: **Correlations between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and within-task explicit cause type ratings.** Pre-task ratings are ratings provided by participants prior to starting the task, during which they are asked to reflect on the causes behind a recent negative and positive event from their own lives, which would not be expected to relate to within-task learning rates. Scenarios 1-3 represent ratings of 'correct' causes following each task scenario, on an external-internal dimension scale (events were caused…"completely by other people or circumstances", "completely by myself') and specific-global dimension scale (events were caused…"by things related to the specific circumstances", "by things that affect all areas of my life"). Posterior α_{pos} estimates are summarised by the mean of the posterior distribution for each participant, with point weight representing the precision of estimation (1/posterior SD).

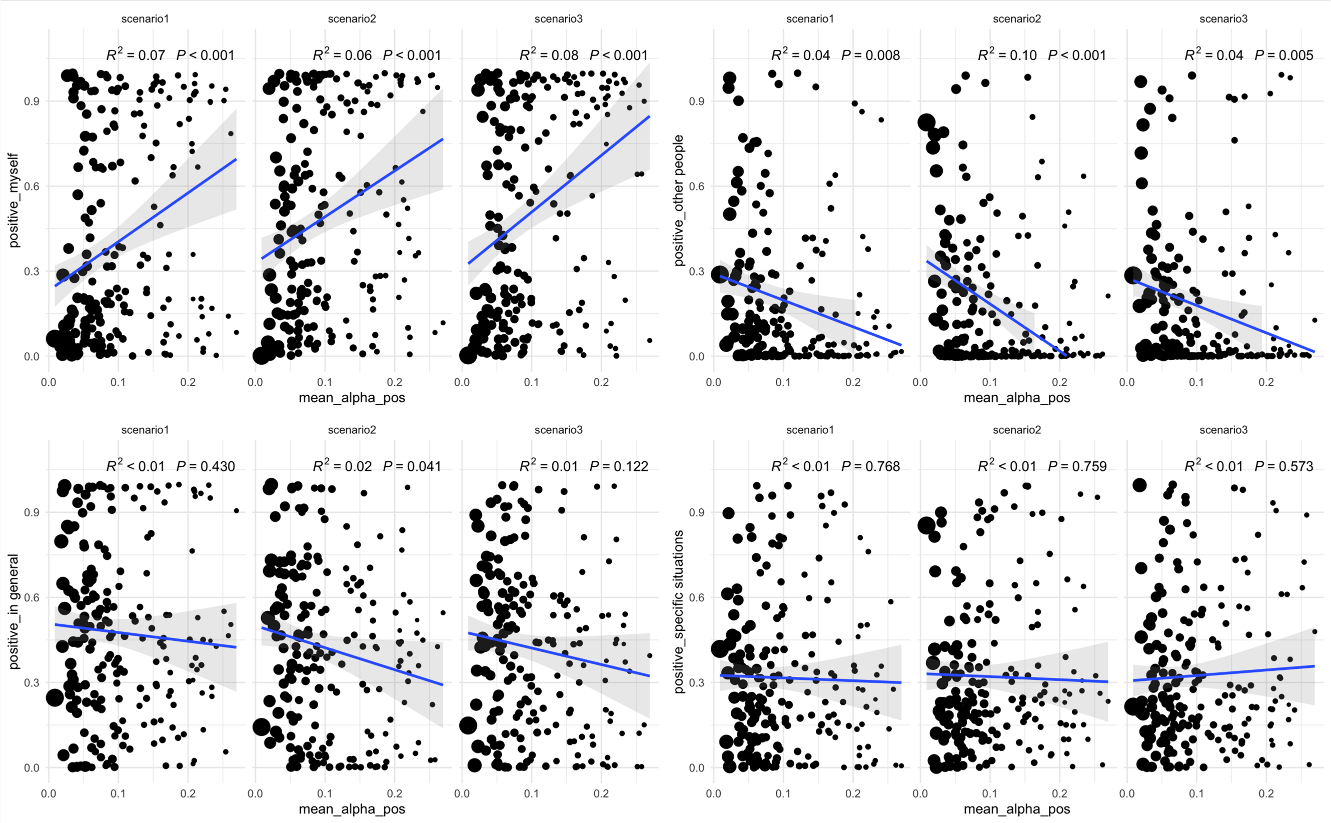

Figure S6: **Correlations between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and label classification probabilities of free-text descriptions of correct cause types (study 1)** Scenarios 1-3 represent classifier output for free-text descriptions of the kinds of 'correct' causes in each preceding task scenario. Posterior α_{pos} estimates are summarised by the mean of the posterior distribution for each participant, with point weight representing the precision of estimation (1/posterior SD).

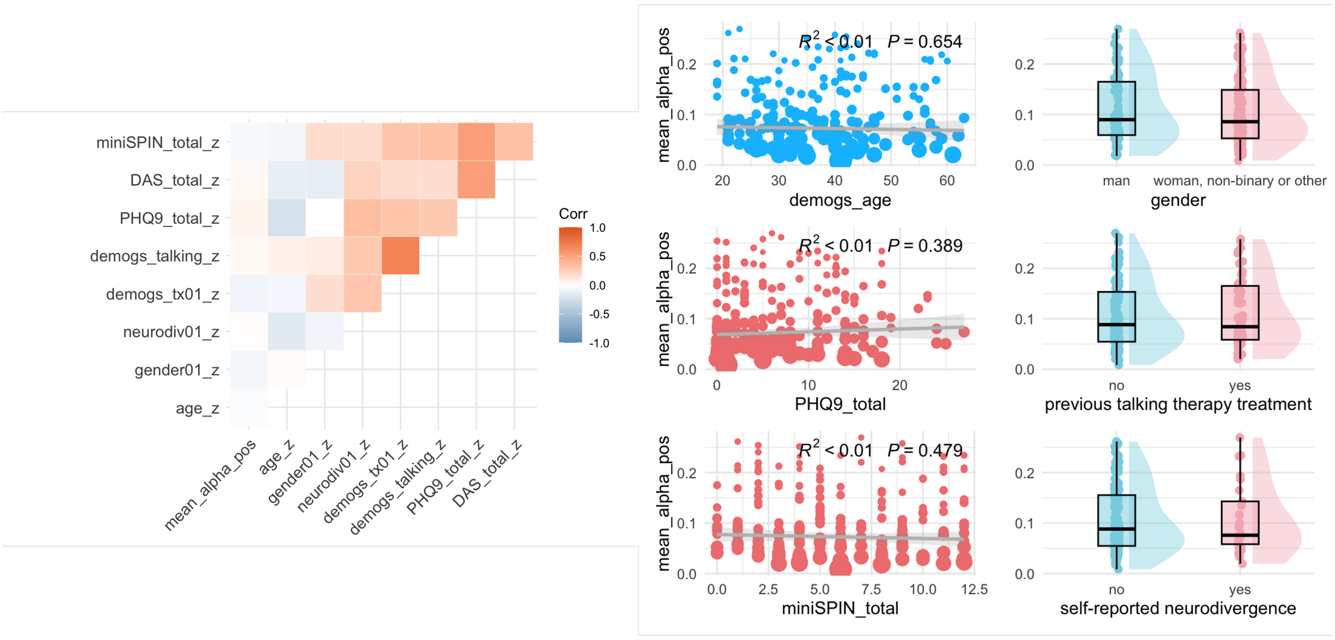

Figure S7: **Relationships between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and self-reported participant demographic and clinical information (study 1).**

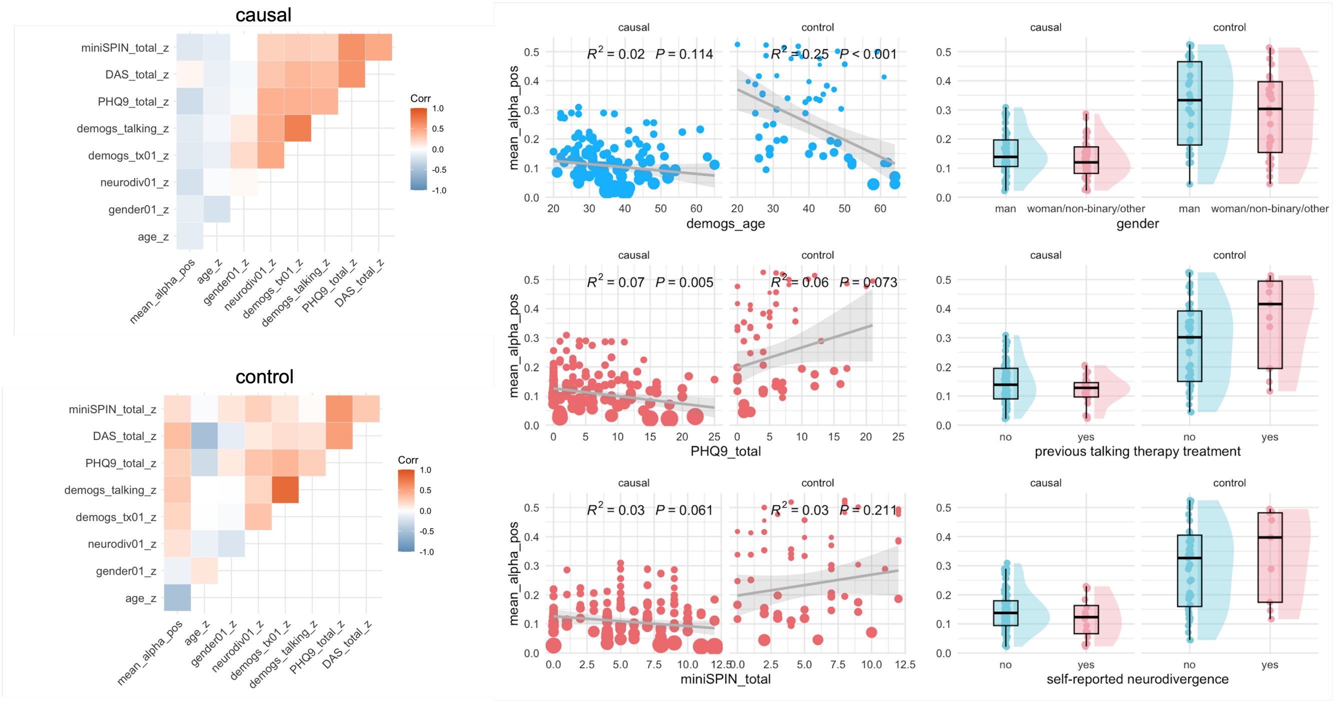

Figure S8: **Relationships between mean posterior estimates of learning rates for positive events from the learning training and and control learning tasks (α_{pos}) and self-reported participant demographic and clinical information (study 2).**

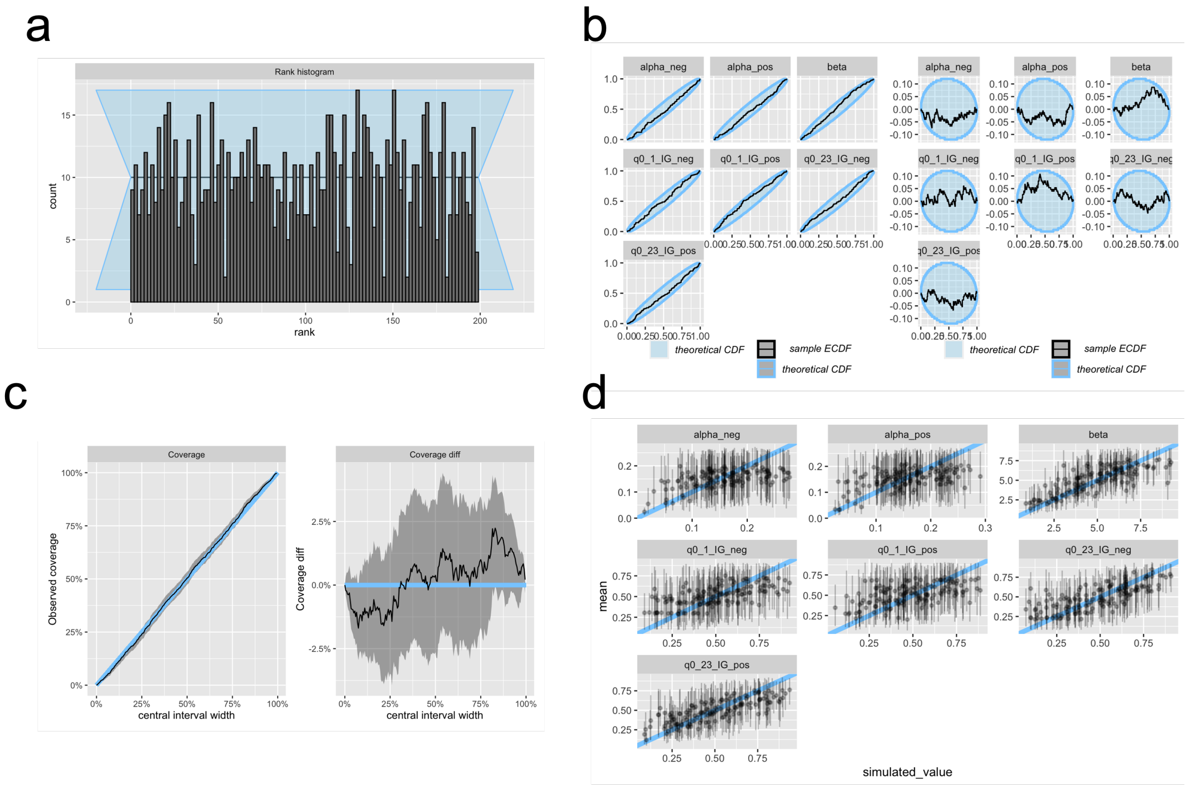

Figure S9: **Simulation-based calibration analysis for the learning training task. a** Rank histogram, a check for uniformity of posterior draw ranks. Horizontal black line, expected average count; blue trapezoid, approximate 95% interval for expected deviations over ranks. **b** (E)CDF, (empirical) cumulative distribution functions for each model parameter. Blue ellipses, regions outlining expected 95% deviations; circular plots show are rotated by 45 for easier visualisation of deviations. **c** Coverage plots, which show the proportion of true variable values that fall within the 95% posterior credible intervals for each parameter. Rank histogram, a check for uniformity of posterior draw ranks. Horizontal black line, expected average count; blue trapezoid, approximate 95% interval for expected deviations over ranks. **d** Simulated and recovered posterior values for independently randomly generated parameter values, for 150 simulated datasets.

Supplementary Tables

	mean	se (mean)	sd	10%	90%	N_{eff}	$\hat{R}$
Mean θ for internal attributions of negative events at time 1	-0.222	0.002	0.061	-0.302	-0.144	778	1.007
Mean θ for internal attributions of negative events at time 2	-0.525	0.003	0.090	-0.640	-0.410	740	1.004
Mean θ for internal attributions of positive events at time 1	1.085	0.003	0.071	0.994	1.176	718	1.002
Mean θ for internal attributions of positive events at time 2	2.511	0.007	0.182	2.278	2.742	715	1.003
Mean θ for global attributions of negative events at time 1	-0.580	0.001	0.051	-0.645	-0.514	1313	1.003
Mean θ for global attributions of negative events at time 2	-0.730	0.002	0.066	-0.813	-0.646	1299	1.001
Mean θ for global attributions of positive events at time 1	-0.062	0.003	0.066	-0.150	0.019	648	1.003
Mean θ for global attributions of positive events at time 2	0.661	0.006	0.163	0.455	0.865	869	1.000
Effect of restructuring on θ internal-negative at time 2	-0.476	0.004	0.132	-0.645	-0.306	894	1.006
Effect of restructuring on θ internal-positive at time 2	0.331	0.009	0.254	0.004	0.665	746	1.004
Effect of restructuring on θ global-negative at time 2	0.069	0.002	0.090	-0.049	0.183	1561	1.003
Effect of restructuring on θ global-positive at time 2	0.493	0.008	0.240	0.185	0.798	844	1.004

Table S1: **Hierarchical Bayesian model results for effects of cognitive restructuring on causal attribution tendencies in study 1 data** Mean, posterior mean; se (mean), standard error of the posterior mean. 10%, 90%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size (an estimate of the number of independent draws from the posterior distribution of the estimand of interest); $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains (if all chains are at equilibrium, $\hat{R}$ will be 1). All values are raw (untransformed) parameter estimates (for transformation constraints applied to main text figures see Supplementary Methods).

	mean	**se (mean)**	**sd**	**10%**	**90%**	N_{eff}	$\hat{R}$
Mean θ for internal attributions of negative events at time 1	-0.285	0.002	0.067	-0.370	-0.200	1295	1.001
Mean θ for internal attributions of negative events at time 2	-0.106	0.005	0.180	-0.339	0.121	1266	1.002
Mean θ for internal attributions of positive events at time 1	1.018	0.002	0.074	0.924	1.113	1641	1.001
Mean θ for internal attributions of positive events at time 2	1.074	0.009	0.307	0.671	1.462	1184	1.001
Mean θ for global attributions of negative events at time 1	-0.627	0.002	0.067	-0.713	-0.543	1582	1.002
Mean θ for global attributions of negative events at time 2	-0.625	0.004	0.162	-0.834	-0.420	1499	1.001
Mean θ for global attributions of positive events at time 1	-0.022	0.001	0.064	-0.103	0.061	1815	1.002
Mean θ for global attributions of positive events at time 2	-0.295	0.010	0.304	-0.675	0.101	920	1.001
Effect of restructuring on θ internal-negative at time 2	-0.299	0.004	0.149	-0.490	-0.108	1509	1.002
Effect of restructuring on θ internal-positive at time 2	0.634	0.007	0.268	0.291	0.973	1319	1.000
Effect of restructuring on θ global-negative at time 2	0.003	0.003	0.135	-0.166	0.178	1734	1.000
Effect of restructuring on θ global-positive at time 2	0.326	0.008	0.260	-0.011	0.653	1009	1.001
Effect of learning training on θ internal-negative at time 2	-0.491	0.004	0.150	-0.682	-0.298	1430	1.002
Effect of learning training on θ internal-positive at time 2	1.147	0.008	0.268	0.803	1.492	1002	1.006
Effect of learning training on θ global-negative at time 2	-0.149	0.003	0.136	-0.329	0.025	1732	1.001
Effect of learning training on θ global-positive at time 2	0.939	0.009	0.261	0.604	1.267	935	1.004

Table S2: **Hierarchical Bayesian model results for effects of cognitive restructuring and learning training on causal attribution tendencies in study 2 data.** Mean, posterior mean; se (mean), standard error of the posterior mean. 10%, 90%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size (an estimate of the number of independent draws from the posterior distribution of the estimand of interest); $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains (if all chains are at equilibrium, $\hat{R}$ will be 1). All values are raw (untransformed) parameter estimates (for transformation constraints applied to main text figures see Supplementary Methods).

	mean	**se (mean)**	**sd**	**10%**	**90%**	N_{eff}	$\hat{R}$
Effect of restructuring on θ internal-negative at time 2	-0.477	0.004	0.134	-0.651	-0.304	1122	1.002
Effect of restructuring on θ internal-positive at time 2	0.246	0.006	0.219	-0.038	0.529	1265	1.005
Effect of restructuring on θ global-negative at time 2	0.071	0.002	0.092	-0.046	0.186	1660	1.000
Effect of restructuring on θ global-positive at time 2	0.417	0.005	0.193	0.169	0.659	1447	1.004
$Beta_{LEARN}$, effect of α_{pos} on θ internal-positive at time 2	0.538	0.015	0.177	0.344	0.745	148	1.045
$Beta_{LEARN}$, effect of α_{pos} on θ global-positive at time 2	0.439	0.013	0.146	0.282	0.615	121	1.043

Table S3: **Joint hierarchical model 1 results for study 1 data**. *Mean*, posterior mean; *se (mean)*, standard error of the posterior mean. *10%, 90%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains). All values are raw (untransformed) parameter estimates, except β values which are in units of α_{pos} (which ranges [0,1]), which have been transformed to a similar range as other intervention effects by /100.

	mean	**se (mean)**	**sd**	**10%**	**90%**	N_{eff}	$\hat{R}$
Effect of restructuring on θ internal-negative at time 2	-0.321	0.004	0.154	-0.517	-0.127	1901	1.001
Effect of restructuring on θ internal-positive at time 2	0.677	0.006	0.299	0.294	1.062	2612	1.000
Effect of restructuring on θ global-negative at time 2	0.032	0.003	0.141	-0.144	0.211	1943	1.000
Effect of restructuring on θ global-positive at time 2	0.347	0.006	0.277	-0.007	0.694	1987	1.001
Effect of learning training on θ internal-negative at time 2	-0.520	0.004	0.161	-0.730	-0.316	1849	1.000
Effect of learning training on θ internal-positive at time 2	-0.888	0.016	0.529	-1.555	-0.217	1087	1.002
Effect of learning training on θ global-negative at time 2	-0.139	0.003	0.145	-0.324	0.044	1957	1.000
Effect of learning training on θ global-positive at time 2	-0.959	0.021	0.556	-1.653	-0.260	689	1.003
$Beta_{LEARN}$, effect of α_{pos} on θ internal-positive at time 2	0.291	0.004	0.088	0.190	0.403	618	1.005
$Beta_{LEARN}$, effect of α_{pos} on θ global-positive at time 2	0.257	0.004	0.088	0.155	0.372	472	1.007
$Beta_{CONTROL}$, effect of α_{pos} on θ internal-positive at time 2	0.007	0.001	0.015	-0.010	0.025	465	1.013
$Beta_{CONTROL}$, effect of α_{pos} on θ global-positive at time 2	0.007	0.001	0.014	-0.009	0.023	635	1.008

Table S4: **Joint hierarchical model 1 results for study 2 data**. *Mean*, posterior mean; *se (mean)*, standard error of the posterior mean. *10%, 90%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains). All values are raw (untransformed) parameter estimates, except β values which are in units of α_{pos} (which ranges $[0, 1]$), which have been transformed to a similar range as other intervention effects by /100.

	mean	**se (mean)**	**sd**	**10%**	**90%**	N_{eff}	$\hat{R}$
Effect of restructuring on θ internal-negative at time 2	-0.478	0.004	0.134	-0.651	-0.310	1340	1.000
Effect of restructuring on θ internal-positive at time 2	-0.334	0.021	0.462	-0.931	0.249	490	1.021
Effect of restructuring on θ global-negative at time 2	0.071	0.002	0.092	-0.046	0.189	2003	1.001
Effect of restructuring on θ global-positive at time 2	0.179	0.010	0.386	-0.321	0.664	1498	1.004
$Beta_{LEARN}$, effect of α_{pos} on θ internal-positive at time 2	0.587	0.010	0.193	0.375	0.832	404	1.005
$Beta_{LEARN}$, effect of α_{pos} on θ global-positive at time 2	0.482	0.009	0.160	0.309	0.692	344	1.008
$Beta_{LEARN+CR}$, additional effect of α_{pos} on θ internal-positive at time 2 in restructuring group	0.215	0.007	0.161	0.021	0.422	470	1.021
$Beta_{LEARN+CR}$, additional effect of α_{pos} on θ global-positive at time 2 in restructuring group	0.093	0.003	0.102	-0.026	0.219	1242	1.004

Table S5: **Joint hierarchical model 2 results, for study 1 data**. *Mean*, posterior mean; *se (mean)*, standard error of the posterior mean. *10%, 90%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains). All values are raw (untransformed) parameter estimates, except β values which are in units of α_{pos} (which ranges $[0, 1]$), which have been transformed to a similar range as other intervention effects by /100.

	mean	**se (mean)**	**sd**	**10%**	**90%**	N_{eff}	$\hat{R}$
Effect of restructuring on θ internal-negative at time 2	-0.317	0.004	0.151	-0.510	-0.123	1859	1.001
Effect of restructuring on θ internal-positive at time 2	1.058	0.015	0.553	0.335	1.773	1304	1.002
Effect of restructuring on θ global-negative at time 2	0.017	0.003	0.139	-0.161	0.198	1795	1.000
Effect of restructuring on θ global-positive at time 2	-0.163	0.013	0.487	-0.781	0.458	1470	1.000
Effect of learning training on θ internal-negative at time 2	-0.512	0.004	0.158	-0.721	-0.314	1847	1.000
Effect of learning training on θ internal-positive at time 2	-0.718	0.017	0.565	-1.440	-0.009	1092	1.001
Effect of learning training on θ global-negative at time 2	-0.148	0.004	0.146	-0.330	0.042	1649	1.002
Effect of learning training on θ global-positive at time 2	-1.369	0.028	0.572	-2.101	-0.644	426	1.007
$Beta_{LEARN}$, effect of α_{pos} on θ internal-positive at time 2	0.357	0.007	0.127	0.221	0.519	343	1.016
$Beta_{LEARN}$, effect of α_{pos} on θ global-positive at time 2	0.281	0.007	0.112	0.161	0.420	270	1.018
$Beta_{LEARN+CR}$, additional effect of α_{pos} on θ internal-positive at time 2 in restructuring group	-0.066	0.003	0.098	-0.187	0.052	1005	1.003
$Beta_{LEARN+CR}$, additional effect of α_{pos} on θ global-positive at time 2 in restructuring group	0.087	0.002	0.082	-0.008	0.189	1302	1.002
$Beta_{CONTROL}$, effect of α_{pos} on θ global-positive at time 2	0.005	0.001	0.018	-0.014	0.027	519	1.006
$Beta_{CONTROL}$, effect of α_{pos} on θ internal-positive at time 2	0.011	0.001	0.015	-0.006	0.029	407	1.010

Table S6: **Joint hierarchical model 2 results, for study 2 data**. *Mean*, posterior mean; *se (mean)*, standard error of the posterior mean. *10%, 90%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains). All values are raw (untransformed) parameter estimates, except β values which are in units of α_{pos} (which ranges $[0, 1]$), which have been transformed to a similar range as other intervention effects by /100.

model	**description**	$ELPD_{diff}$ **(choice data)**	SE_{diff} **(choice data)**
base model	Model of choice data only (pre and post-intervention), as described in Norbury et al., 2023.	-21.5	9.8
joint model 1	Joint model of choice data + β weights representing influence of α_{pos} on post-intervention internal-positive and global-positive parameter estimates	-10.3	9.3
joint model 2	As above, with additional β weights representing influence of α_{pos} on post-intervention internal-positive and global-positive parameter estimates in restructuring condition participants	0.0	0.0

Table S7: **Model comparison results for causal attribution task data likelihood from the original (base) model, compared to joint models of causal attribution and learning task data in study 1**. $ELPD_{diff}$, difference in expected log pointwise predictive density for each model from the best model, which is defined as having zero difference to itself. SE_{diff}, the standard error of this difference.

model	description	*N params*	$ELPD_{diff}$	SE_{diff}
m_qlearning_negpos_1alpha	Q-learning model with separate values for internal-global and non-internal global response options for positive and negative events, single learning rate α, and inverse softmax temperature parameter β as individual-level free parameters	2	-655.7	41.6
m_qlearning_negpos_2alpha	As above, with separate αs for positive and negative events	3	-526.8	36.3
m_qlearning_negpos_2alpha_2q0	As above, with a group-level parameter governing the starting values of internal-global attributions (q0) for positive and negative events, across all scenarios	3	-70.0	14.8
m_qlearning_negpos_2alpha_2q0_init_delta	As above, with q0 applied to the first scenario only, then incremented by a group-level delta parameter for scenarios 2,3	3	-20.3	10.5
m_qlearning_negpos_2alpha_2q0_init_2delta	As above, with scenario 2 q0 = q0 + delta, and scenario 3 q0 = q0 + 2*delta	3	-13.6	8.2
m_qlearning_negpos_2alpha_2q0i	As m_qlearning_negpos_2alpha, but with q0 as an individual-level free parameter applied to all scenarios (q0i)	5	-59.9	13.6
m_qlearning_negpos_2alpha_2q0i_init	As above, with q0i applied to scenario 1 only	5	-428.6	30.9
m_qlearning_negpos_2alpha_2q0i_init_delta	As above, with starting value for scenarios 2,3 defined as q0i + a group-level delta parameter	5	**-2.4**	**7.3**
m_qlearning_negpos_2alpha_2q0i_init_2delta	As above, with scenario 2 q0 = q0i + delta, and scenario 3 q0 = q0i + 2*delta	5	**0.0**	**0.0**
m_qlearning_negpos_2alpha_2q0i1_2q0i23*	As m_qlearning_negpos_2alpha_2q0i, with separate individual-level free parameters governing q0 for scenario 1 and scenarios 2,3	7	**-5.0**	**9.1**

Table S8: **Model comparison results for causal attribution learning task data from study 1.** $ELPD_{diff}$, difference in expected log pointwise predictive density for each model from the best model, which is defined as having zero difference to itself. SE_{diff}, the standard error of this difference. Models with an ELPD difference of greater than several times the SE of the estimate are usually taken to indicate better predictive performance. *N params*, number of individual-level free parameters. Bold font, best models with roughly equivalent performance. *, model taken forward for subsequent analyses based on results of simulation-based calibration and parameter recovery analysis.