Learning training as a cognitive restructuring intervention

Agnes Norbury[1], **Quentin Dercon**[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: `quentin.dercon.22@ucl.ac.uk`

Running head: Learning training as cognitive restructuring

ABSTRACT

Background. 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.

Methods. We investigated the effects of completing 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 causal attribution tendencies. Studies were conducted online using randomized-controlled experimental designs (N=200 & N=164), and data were analysed using hierarchical Bayesian models.

Results. We found that both learning training and the restructuring intervention decreased tendencies to make unhelpful attributions and increased tendencies to make self-enhancing attributions. Across two studies, changes in attribution tendencies were associated with higher learning rates during learning training, an effect specific to learning about different kinds of event attribution. Contrary to expectation, we found 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.

Conclusions. We suggest that personalized training, in conjunction with feedback based on interpretable computational model output, may provide 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') (*1*, *2*). 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 (*3*, *4*). Demonstrating this using data from traditional randomized-controlled trials involving psychotherapy treatment programs (e.g., cognitive-behavioural therapy (CBT)) is challenging, given the multiple types of interventions delivered in each program coupled with a lack of the fine-grained resolution needed to infer temporal dependencies between changes in beliefs and symptoms (*3*, *5*). 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 (*6*–*11*). That said, the degree of conceptual overlap between self-report measures of cognitive skills and symptoms themselves (the 'jangle' fallacy) makes disentangling changes in the former from overall treatment response or residual symptom burden considerably more difficult (*6*, *12*).

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 (*13*–*15*). Combining cognitive-behavioural measures with randomized allocation of therapy-like interventions in high-throughput testing can provide an 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 (*16*, *17*). 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 (a) training in learning to identify different kinds of causal attributions (a learning task intervention) and (b) practice in identifying and challenging unhelpful attributions of events in their own lives (a brief cognitive restructuring intervention). Cognitive therapy can be considered a process of learning (*13*), and it has been suggested that individuals with greater capacity for learning during treatment show greater benefits (*18*). On this basis, we initially hypothesized that individual differences in learning task performance would be related to individual differences in response to a brief cognitive restructuring intervention.

Instead, across two studies, we found evidence that both the learning task training and the brief cognitive restructuring intervention affected causal attribution tendencies, shifting them 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 a better understanding of the learning task state space. Contrary to expectations, 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 for augmenting psychological treatments with strategies aimed at boosting learning and memory of treatment content, and for whom this might be most effective for (*19*, *20*).

RESULTS

We report results of two cross-sectional studies with similar overall designs (Figure 1). 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.

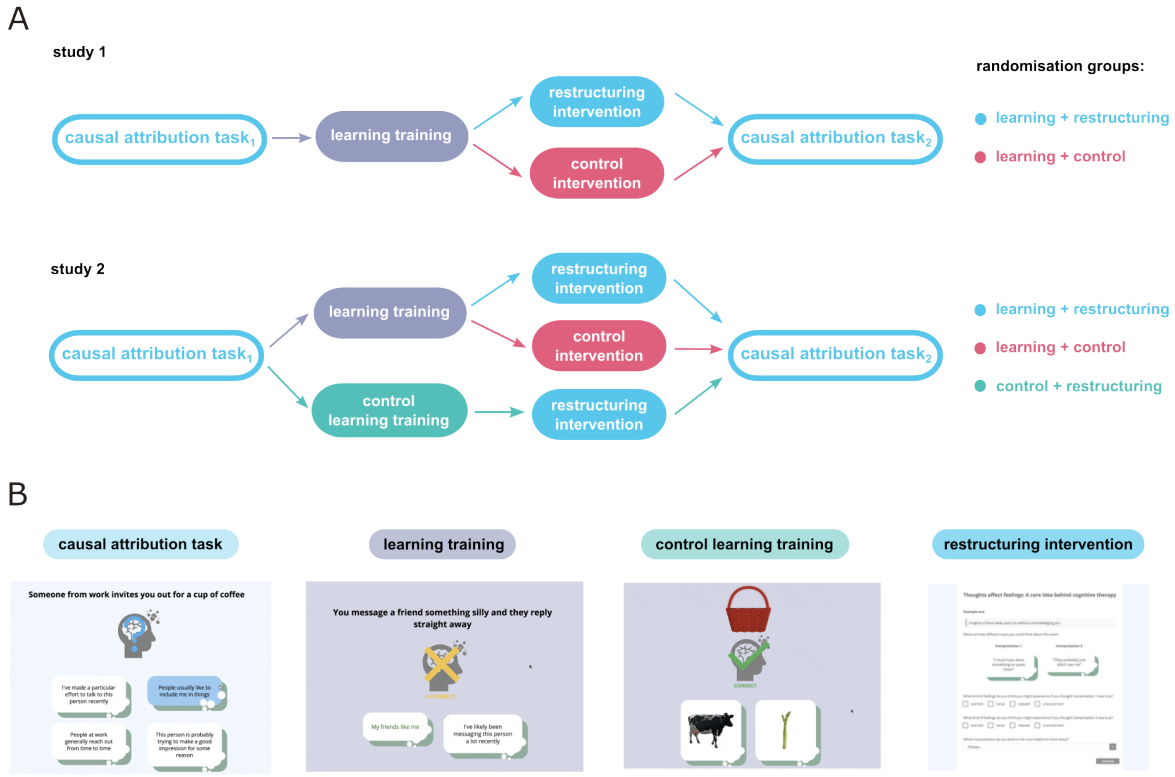

Figure 1: Overview of study designs and measures. **A** Experimental designs and randomisation conditions for each study. In both studies, a cognitive-behavioural measure of causal attribution tendencies (the causal attribution task), was completed pre- and post-completion of two types of intervention. 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, over a single experimental session (around one hour in length). **B** Representative screenshots 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 the kinds of explanations thought to be correct for a hypothetical person in a particular mood state, given explicit feedback. The control learning task, identical in structure, 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 then apply it to recent events from their own lives. Further screenshots and demonstrations of the tasks and interventions are available on the study GitHub repository.

Participants

Participants for both studies were recruited from an online research participation platform (Prolific (*21*)) and are described in Table 1.

In both studies, samples showed 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 PHQ-9 = 32% & 27%; proportion of participants with significant social anxiety according to the miniSPIN = 48% & 46%; Figure S1).

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

We first examined 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 (see Methods).

In study 1 all participants completed the learning training task, so here 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% credible interval (CrI) = (-0.70, -0.26)]), and an increased tendency to attribute positive events to general or global causes (posterior estimate=0.50 [90% CrI = (0.11, 0.90)]) (Figure 2*A-B*; Table S1). Of interest, the group means for each parameter showed some evidence of shifts between time-points, with participants showing slightly higher mean endorsement of internal and global attributions of positive events at the second measurement (Figure 2). These group-level shifts could represent common effects of completing the cognitive restructuring and control interventions on attribution tendency. However, as the control intervention made no reference to how interpretations of events might affect mood, or reappraisal strategies, this is unlikely. An alternative explanation is that these effects are due to completion of the learning training task by all study participants, since this directly involves learning to recognise different kinds of attributions.

We tested this idea directly in study 2. Importantly, this study included a control learning task, as well as cognitive restructuring and control intervention conditions. To formally test whether completion of the learning task resulted in group-level changes in attribution tendencies, we augmented the analysis model for these data such that post-intervention (time 2) attribution tendencies could be influenced by learning training condition, as well as restructuring intervention condition (see Methods).

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 (ELPD) for left-out causal attribution task data, $\mathrm{ELPD}_{\mathrm{diff}} = -0.4$, but of less than 5x than

		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%)
Neurodivergence	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%)
PHQ-9 (/27)	Mean (SD)	7.3 (6.2)	6.3 (5.8)
DAS-SF (/36)	Mean (SD)	19.2 (4.6)	18.6 (4.8)
miniSPIN (/12)	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 explained 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 attention-deficit/hyperactivity disorder (ADHD)*'. Categories for previous mental health treatment were talking therapy (including CBT), medication, self-guided (e.g., workbooks or apps), or other. The 9-item patient health questionnaire (PHQ-9) assesses depressed mood; the short-form dysfunctional attitudes scale (DAS-SF) assesses dysfunctional beliefs; and the 3-item social phobia inventory (miniSPIN) assesses social anxiety.

the standard error (SE) of the estimate: $\mathrm{SE}_{\mathrm{diff}} = 6.8$), suggesting that this indeed had an additional impact on changes in attribution tendencies.

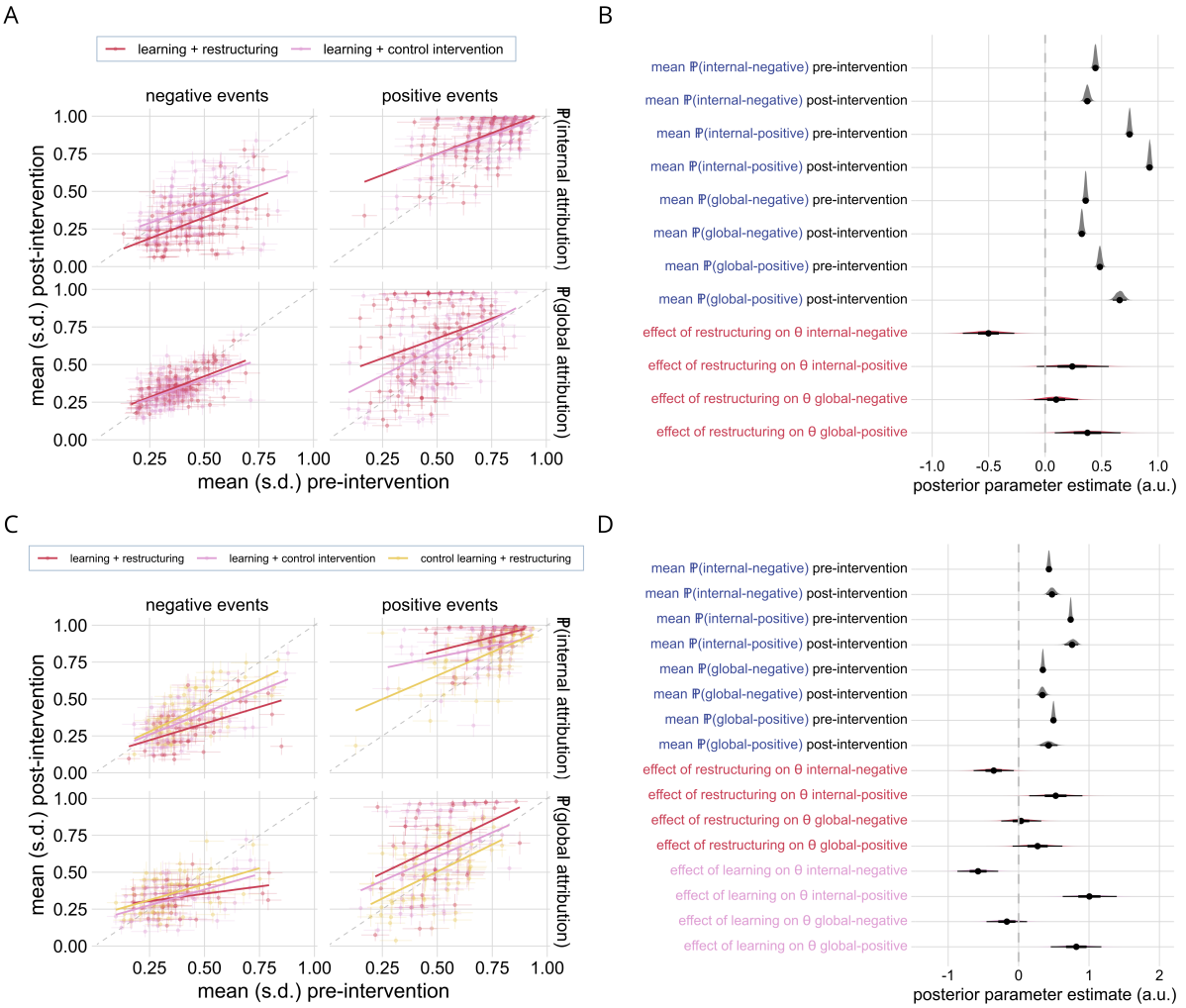

Figure 2: Independent effects of learning training and brief cognitive restructuring on causal attribution. **A** Posterior mean (and standard deviation (SD)) parameter estimates for the causal attribution task for each participant at time 1 (pre-intervention) and time 2 (post-intervention) by randomisation 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 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, where $\mathbb{P}$ denotes probability. Thick inner lines represent 50% and thin outer lines represent 90% quantile-based CrIs (i.e., 90% of the probability density contained within the interval). 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 standardised mean differences (SMDs). **C** The same plot as (*A*), for study 2 participants (N=164). **D** The same plot as (*C*), for study 2 participants; $\mathbb{P}$ denotes probability. 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) revealed 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 2*C*). 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 = -0.32 [90% CrI = (-0.57, -0.06)]), and increased tendency to attribute positive events to internal causes (posterior estimate = 0.65 [90% CrI = (0.19, 1.11)]) (Figure 2*D*, 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 further decreased internal attribution of negative events, as well as increased internal and global attribution of positive events (posterior estimates = -0.51 [90% CrI = (-0.77, -0.26)], 1.24 [90% CrI = (0.77, 1.72)], 1.03 [90% CrI = (0.58, 1.47)]; Table S2).

Therefore, at the group level, both completion of the restructuring intervention and completion of learning training task impacted causal attribution tendencies for everyday events—with both intervention components resulting in a decreased tendency to choose unhelpful and increased tendency to choose self-enhancing interpretations.

Learning rates from the learning training task and changes in self-enhancing attributions

If learning is critical to the effects described above, we might reasonably expect that the effects of the learning task intervention to 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). 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 (Figure S2), we focused our analysis on learning estimates for positive events. Specifically, positive learning rates from the learning task were then 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 relationships between point estimates (posterior parameter means) from separately modelled learning and causal attribution task data. We then carried out a formal test of association 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 regression weights (see Methods).

Associations between separately modelled learning and attribution task data. 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 3*A*,*D*; for pre- and post-intervention parameter estimates see Figure S3). These relationships were not evident for learning rates de-

rived from the control learning task ($Rs = 0.14$ & 0.10; Figure 3*D*).

There was no convincing evidence that the strength of these correlations differed 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.

Results of the first joint models provided strong evidence of positive relationships between positive learning rate (α_{pos}) estimates and changes in internal and global attributions of positive events, across intervention conditions, in study 1 participants (β_{LEARN} internal-positive = 0.56 [90% CrI = (0.34, 0.88)], β_{LEARN} global-positive = 0.46 [90% CrI = (0.27, 0.72)], Figure 3*B*, Table S3). These effects were replicated in study 2 data (β_{LEARN} internal-positive = 0.29 [90% CrI = (0.17, 0.45)], β_{LEARN} global-positive = 0.26 [90% CrI = (0.13, 0.42)]) - but were not evident for learning rates estimated from the control learning task (β_{CONTROL} internal-positive = 0.01 [90% CrI = (-0.02, 0.03)], β_{CONTROL} global-positive = 0.01 [90% CrI = (-0.01, 0.03)], Figure 3*E*, 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 some weak 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 ($\beta_{\text{LEARN+CR}}$ internal-positive = 0.23 [90% CrI = (-0.002, 0.42)]), but there was no evidence for this effect in study 2 ($\beta_{\text{LEARN+CR}}$ internal-positive = -0.07 [90% CrI = (-0.23, 0.08)]). In neither study was there any convincing evidence for an additional influence of α_{pos} estimates on change in global-positive attributions in restructuring group participants (study 1: $\beta_{\text{LEARN+CR}}$ global-positive = 0.10 [90% CrI = (-0.06, 0.28)],Figure 3*C*, Table S5, study 2: $\beta_{\text{LEARN+CR}}$ global-positive = 0.09 [90% CrI = (-0.04, 0.24)], Figure 3*F*, Table S6). Therefore we found no strong evidence in favour of a selective interaction between faster learning on the learning training task and response to the cognitive restructuring intervention.

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

We next 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. Full analysis of learning task data (choice accuracy, response times, explicit-cause ratings and free-text cause descriptions) is available in the Supplementary Results.

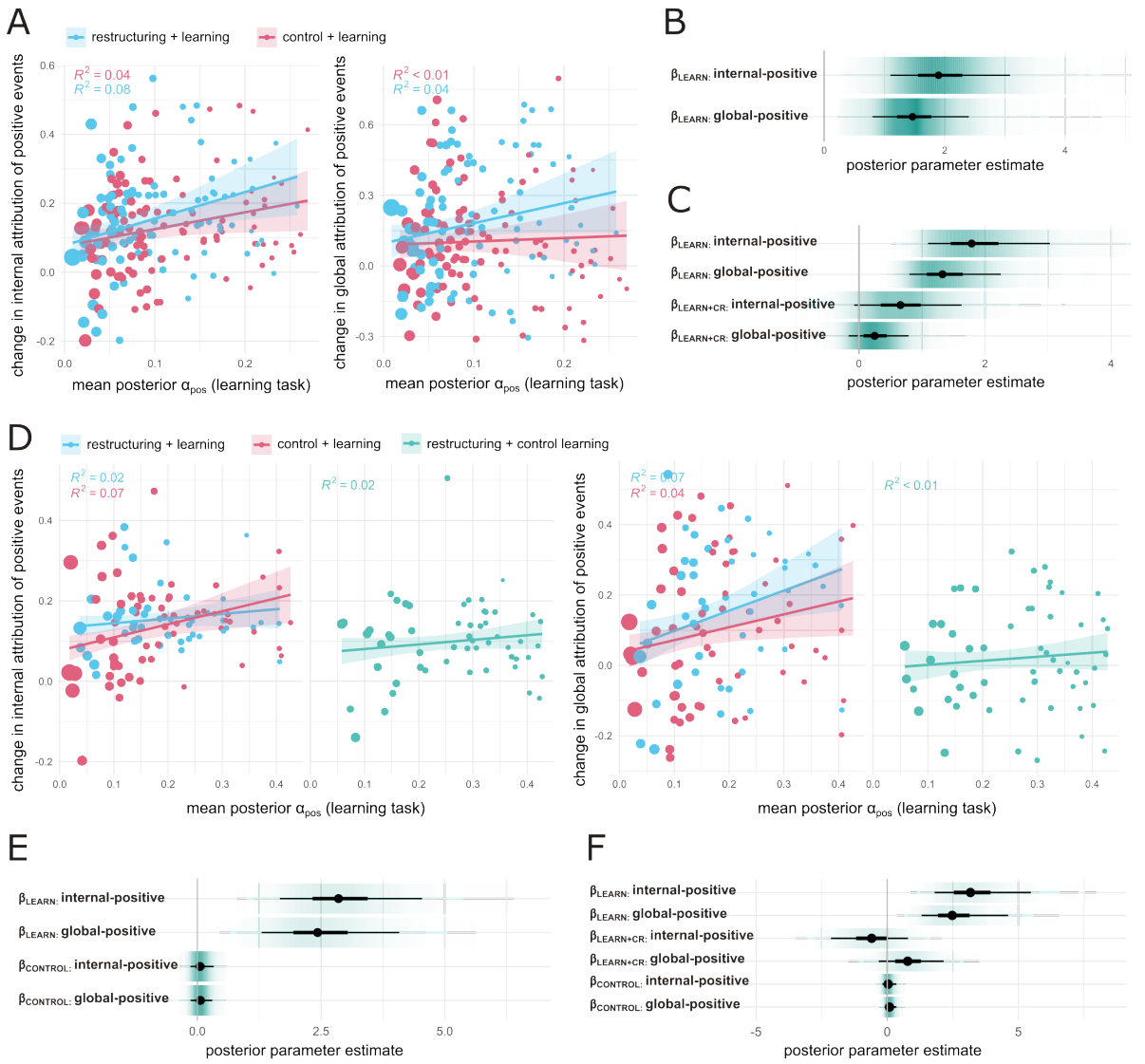

Figure 3: Changes in self-enhancing attributions were positively associated with learning rate estimates from the learning training task, but this effect was not greater in participants who completed cognitive restructuring. **A** Correlations between posterior mean estimates for positive learning rate from the 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 the estimated posterior precision of α_{pos} (i.e., 1/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 predictor (i.e., α_{pos}) and outcome (i.e., mean posterior parameter) SDs, making them roughly equivalent to standardized regression coefficients. Black lines represent 50 & 90% posterior CrIs, and shading represents the 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 ($\beta_{\text{LEARN+CR}}$). **D** The same plot as A, for study 2 participants. **E** The same plot as *B*, for study 2 participants. β_{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 brief cognitive restructuring in addition to learning training.

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: $Rs = 0.20 - 0.33, p \leq 0.033$, Figure S6). 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: $F1, 112.7 = 4.1, p < 0.05$).

Relationships between positive learning rates and free-text cause description label probabilities. In study 1, there was strong evidence that posterior mean estimates of α_{pos} were positively 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 S7). 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).

In summary, in a reinforced setting, participants who learned more quickly to select self-enhancing attributions were also able to better describe the types of causes reinforced as correct during the task. This suggests that participants with higher learning rate estimates (α_{pos}) may have had a 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.

DISCUSSION

Theories of cognitive restructuring suggest it is a process based on learning (*13*). Individual differences in learning and memory of therapy content may be a moderator of symptom change during treatment (*18*, *19*). Inspired by recent demonstrations that clinically relevant inference processes can be reliably measured using computerised learning tasks (*22*, *23*), 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 were specifically related to change in attribution tendencies following the restructuring intervention. Instead, 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 was related to how quickly participants updated 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 with a more intuitive understanding of these dimensions 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 training. For example, in non-clinically depressed participants, a single session of app-based reappraisal training 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 (*24*). Similarly, three weeks of online training 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 (*25*). A recent meta-analysis also found evidence that cognitive bias training (where participants are typically presented with ambiguous everyday scenarios and trained to resolve them in favour of neutral or positive interpretations) reduced symptoms of anxiety and depression compared to some control conditions (*26*).

A novel aspect of the learning task described here is the use of a 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 (*27*) and can alter learning (*16*). One advantage of tasks that can measure attribution biases along multiple dimensions—in conjunction with interpretable computational models—is that this information can be fed back to users over time. 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 (*20*).

The major limitations of the studies presented here are that participation was not restricted to individuals currently experiencing clinically-significant levels of psychological symptoms, and that the brief restructuring intervention used here was not a real-world (i.e., clinically validated) 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 (*28*, *29*).

There are also some methodological limitations of our studies. First, for consistency with our previous work (*17*), we interpret 90% CrIs excluding zero as indicative of a meaningful contribution of a given parameter to the behavioural dimension of interest. This may be more prone to false positives than wider intervals, though we note 95% CrIs also excluded zero for all single-signed credible intervals we report in the text (see Supplementary Tables). More pertinently, although inference procedures for the learning task model were well-calibrated, we do not provide empirical data regarding the test-retest reliability of these measures. This limits our ability to infer reliable individual differences in learning between participants. We were unable to investigate individual differences in learning about negative events on the learning training task (a dimension that may be particularly relevant for depression), given our sample was mostly at ceiling for this response dimension. It is also important to note that our single-session experimental design, whilst supporting fast and high-throughput measurement in a relevant sample of individuals, may result in increased likelihood of motivational biases or demand effects influencing our primary dependent measures (i.e., participants updating their responses on the second attribution task in line with previously reinforced 'correct' responses on the learning training task, or the perceived purpose of the study). It will therefore be vital to determine in future work whether effects observed here are evident over longer timescales, and if they generalise to 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, remotely-delivered treatments where the potential for impact is large, but where initializing engagement and high attrition rates are acute problems (*30*). 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" (*31*). It is possible that using cognitive tasks with interpretable model-based output, and, critically, feeding this information back to users can help address this need. The utility of these approaches needs to be established in empirical studies, ideally with participation from all 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 (*32*).

METHODS

DATA AND CODE AVAILABILITY STATEMENT

Code for implementing all tasks and analyses described here, alongside anonymized study data is available on 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

Participants were recruited from an online research participation platform (Prolific (*21*)), and required to be resident in the UK, 18-65 years old, and fluent in English. Power analyses for both studies are available in the Supplementary Methods.

STUDY DESIGN

The design of each study is described in Figure 1*A*. Upon recruitment to each study, participants were assigned to a study arm using a random number generation-based procedure. All studies took place online over a single session, of approximately one hour.

MEASURES

Causal attribution task

A full description of the causal attribution task, including task development, design optimization, and measurement properties can be found in Norbury *et al.* (2024) (*17*). Of note, output parameters from the associated analysis model have excellent identifiability and test-retest reliability, and have previously been found to be associated with self-reported negative self-beliefs and current depression symptom severity.

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 happened to them.

In each of two equivalent versions of the task (one pre- and one post-intervention), participants were presented with 32 event scenarios (16 positive and 16 negative events, randomly interleaved), divided into two blocks. Event scenarios differed across the two task versions. 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) (*33*).

Learning training task

The learning training task was developed as a measure of how easily participants can learn to select different kinds of causal attributions, in a reinforced setting. In contrast to the causal attribution task, the learning training task used a third person framing. Specifically, participants were told that they would be learning about how a hypothetical person in a particular mood might reason about the causes behind events. For each scenario, it was their job to learn to select the correct kinds of explanations for that person in that mood, via trial and error. Participants were provided with explicit instructions stressing the differences between the learning and attribution tasks, and required to pass a multiple-choice post-instructions quiz before proceeding (for full details, see screenshots available on the study GitHub repository). After each scenario, participants were asked to provide ratings and brief free-text description of the kinds of causes that were correct in that scenario (see Supplementary Methods).

Control learning task

The control learning task was exactly matched in trial type and reinforcement structure to the causal learning task. Participants were told that 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 (see Supplementary Methods). 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 (*1*) and consisted of information about a cognitive model of mood, 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 (*34*), 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. The full content of each intervention is available on the study GitHub repository.

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 (see Supplementary Methods).

Analysis

All analyses were carried out in R version 4.1.2 (R Core Team, 2021).

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` (see Supplementary Methods).

Classification of learning task free-text data

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 natural language processing (NLP) classification pipeline (Facebook's BART-MNLI-LARGE transformer model (*35*)), with the non mutually-exclusive candidate labels ["myself", "other people", "in general", "specific situations"]. Output probabilities for each candidate label were further analyzed as above.

Hierarchical Bayesian modelling

General methods. Model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling as implemented in Stan 2.21.0 (*36*), using RStan 2.21.3 (Stan Development Team, 2021). All models used generic weakly-informative priors (see Supplementary Methods). We report quantile-based 90% CrIs for consistency with results reported in our previous work on similar data (*17*), though the respective quantiles for 95% intervals can be found in the Supplementary Tables.

Hierarchical Bayesian analysis of causal attribution task data

Modelling of causal attribution task data followed the approach previously described in Norbury *et al.* (2024) (*17*), using an analysis model for which task design was previously optimised (see Supplementary Methods). Group-level parameters described potential effects of allocation to the restructuring intervention on individual-level parameter estimates at time 2, with priors for these parameters centred on 0. For study 2, this model was augmented to include potential effects of allocation to the learning training task condition.

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, 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 (see Supplementary Methods). Under 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.

Associating separately modelled causal attribution and learning task data parameters

As simple first-pass check, we examined correlations between point estimates (posterior means) of each parameter, weighted by the posterior precision (i.e., 1/SD) of the predictor variable (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

To formally test for associations between parameters, we constructed a series of joint models of causal attribution and learning task data (*37*, *38*). For the first joint models, the causal attribution task analysis model (Supplementary Methods) 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 ($\phi_{p,2}$) via the inclusion of β weight parameters (β_{LEARN}; see Hopkins *et al.* (2021) (*23*) and Haines *et al.* (2020) (*39*)). These β weights can interpreted similarly as in a standard regression model, with the group-level intervention effects (e.g., ϕ_{CR} for the cognitive restructuring (CR) intervention) now representing the intercept.

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

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}, β_{CONTROL}):

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

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

For study 1:

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{\text{CR}} + \alpha_{\text{pos}} * \left(\beta_{\text{LEARN}} + \beta_{\text{LEARN+CR}}\right), & \text{if CR intervention + learning task} \\ \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if control intervention + learning task} \end{cases}
\end{aligned} \tag{3}$$

For study 2:

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{\text{CR}} + \phi_{\text{LEARN}} + \\ \quad \alpha_{\text{pos}} * \left(\beta_{\text{LEARN}} + \beta_{\text{LEARN+CR}}\right), & \text{if CR intervention + learning task} \\ \phi_{\text{LEARN}} + \alpha_{\text{pos}} * \beta_{\text{LEARN}}, & \text{if control intervention + learning task} \\ \phi_{\text{CR}} + \alpha_{\text{pos}} * \beta_{\text{CONTROL}} & \text{if CR intervention + control learning task} \end{cases}
\end{aligned} \tag{4}$$

For all joint models, the priors for β effects were centred on zero (e.g., $\beta_{\text{LEARN}} \sim N(0, 1)$).

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. QD is supported by a Wellcome Trust PhD studentship. 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. 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.

Disclosures

AN, QD and RJD declare 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.

Acronyms

ADHD	Attention-Deficit/Hyperactivity Disorder
CBT	Cognitive-Behavioural Therapy
CrI	Credible Interval
CR	Cognitive Restructuring
DAS-SF	Dysfunctional Attitudes Scale, Short-Form
ELPD	Expected Log Pointwise Predictive Density
MCMC	Markov Chain Monte Carlo
miniSPIN	mini Social Phobia Inventory, 3-item
NLP	Natural Language Processing
PHQ-9	Patient Health Questionnaire, 9-item
RT	Reaction Time
SD	Standard Deviation
SE	Standard Error
SBC	Simulation-Based Calibration
SMD	Standardised Mean Difference

REFERENCES

1. A. T. Beck, A. J. Rush, B. F. Shaw, G. Emery, *Cognitive Therapy of Depression* (Guilford Press, New York, 1979).
2. D. A. Clark, Cognitive Reappraisal. *Cognitive and Behavioral Practice* **29**, 564–566, DOI (2022).
3. L. Lorenzo-Luaces, R. E. German, R. J. DeRubeis, It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, Psychological Interventions for Depression **41**, 3–15, DOI (2015).
4. L. Lorenzo-Luaces, J. R. Keefe, R. J. DeRubeis, Cognitive-Behavioral Therapy: Nature and Relation to Non-Cognitive Behavioral Therapy. *Behavior Therapy* **47**, 785–803, DOI (2016).
5. A. E. Kazdin, Understanding how and why psychotherapy leads to change. *Psychotherapy Research* **19**, 418–428, DOI (2009).
6. N. E. Hundt, J. Mignogna, C. Underhill, J. A. Cully, The relationship between use of CBT skills and depression treatment outcome: a theoretical and methodological review of the literature. *Behavior Therapy* **44**, 12–26, DOI (2013).
7. D. R. Strunk, S. N. Hollars, A. D. Adler, L. A. Goldstein, J. D. Braun, Assessing Patients' Cognitive Therapy Skills: Initial Evaluation of the Competencies of Cognitive Therapy Scale. *Cognitive Therapy and Research* **38**, 559–569, DOI (2014).
8. L. L. Hawley *et al.*, Cognitive-Behavioral Therapy for Depression Using Mind Over Mood: CBT Skill Use and Differential Symptom Alleviation. *Behavior Therapy* **48**, 29–44, DOI (2017).
9. N. B. Gumport, L. Dong, J. Y. Lee, A. G. Harvey, Patient Learning of Treatment Contents in Cognitive Therapy. *Journal of Behavior Therapy and Experimental Psychiatry* **58**, 51–59, DOI (2018).
10. N. R. Forand *et al.*, Efficacy of Guided iCBT for Depression and Mediation of Change by Cognitive Skill Acquisition. *Behavior Therapy* **49**, 295–307, DOI (2018).
11. I. D. Schmidt, B. J. Pfeifer, D. R. Strunk, 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**, 446–456, DOI (2019).
12. L. Lorenzo-Luaces, Identifying active ingredients in cognitive-behavioral therapies: What if we didn't? *Behaviour Research and Therapy* **168**, 104365, DOI (2023).
13. M. Moutoussis, N. Shahar, T. U. Hauser, R. J. Dolan, Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Computational Psychiatry* **2**, 50–73, DOI (2018).
14. A. M. Reiter, N. A. Atiya, I. M. Berwian, Q. J. Huys, Neuro-cognitive processes as mediators of psychological treatment effects. *Current Opinion in Behavioral Sciences*, Computational cognitive neuroscience **38**, 103–109, DOI (2021).
15. Q. J. Huys, M. Browning, M. P. Paulus, M. J. Frank, Advances in the computational understanding of mental illness. *Neuropsychopharmacology* **46**, 3–19, DOI (2021).
16. Q. Dercon *et al.*, A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychological Medicine* **54**, 327–337, DOI (2023).

17. A. Norbury, T. U. Hauser, S. M. Fleming, R. J. Dolan, Q. J. M. Huys, Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Science Advances* **10**, eadk3222, DOI (2024).

18. S. J. E. Bruijniks, R. J. DeRubeis, S. D. Hollon, M. J. H. Huibers, 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**, 668–692, DOI (2019).

19. A. G. Harvey *et al.*, Improving Outcome of Psychosocial Treatments by Enhancing Memory and Learning. *Perspectives on Psychological Science* **9**, 161–179, DOI (2014).

20. C. L. Nord *et al.*, A transdiagnostic meta-analysis of acute augmentations to psychological therapy. *Nature Mental Health* **1**, 389–401, DOI (2023).

21. S. Palan, C. Schitter, Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* **17**, 22–27, DOI (2018).

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

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

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

25. R. R. Morris, S. M. Schueller, R. W. Picard, Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of Medical Internet Research* **17**, e72, DOI (2015).

26. L. A. Fodor *et al.*, Efficacy of cognitive bias modification interventions in anxiety and depressive disorders: a systematic review and network meta-analysis. *The Lancet Psychiatry* **7**, 506–514, DOI (2020).

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

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

29. M. J. H. Huibers, L. Lorenzo-Luaces, P. Cuijpers, N. Kazantzis, On the Road to Personalized Psychotherapy: A Research Agenda Based on Cognitive Behavior Therapy for Depression. *Frontiers in Psychiatry* **11**, 607508, DOI (2021).

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

31. S. E. Knowles *et al.*, Qualitative meta-synthesis of user experience of computerised therapy for depression and anxiety. *PloS One* **9**, e84323, DOI (2014).

32. C. Seiferth *et al.*, How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. *Nature Mental Health* **1**, 542–554, DOI (2023).

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

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

35. M. Lewis *et al.*, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019, Preprint on *arXiv*.

36. B. Carpenter *et al.*, Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, 1–32, DOI (2017).

37. B. M. Turner, B. U. Forstmann, B. C. Love, T. J. Palmeri, L. Van Maanen, Approaches to Analysis in Model-based Cognitive Neuroscience. *Journal of Mathematical Psychology* **76**, 65–79, DOI (B 2017).

38. N. Haines, PhD thesis, The Ohio State University, 2021, LINK.

39. N. Haines *et al.*, Anxiety Modulates Preference for Immediate Rewards Among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis. *Clinical Psychological Science* **8**, 1017–1036, DOI (2020).

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

41. DuncanLab, *OMINDS*, version 0.1.5, 2022, *GitHub*.

42. K. Kroenke, R. L. Spitzer, J. B. Williams, The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606–613, DOI (2001).

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

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

45. I. A. Cristea *et al.*, The effects of cognitive behavior therapy for adult depression on dysfunctional thinking: A meta-analysis. *Clinical Psychology Review* **42**, 62–71, DOI (2015).

46. T. B. Üstün *et al.*, Developing the World Health Organization Disability Assessment Schedule 2.0. *Bulletin of the World Health Organization* **88**, 815–823, DOI (2010).

47. J. E. J. Buckman *et al.*, Socioeconomic Indicators of Treatment Prognosis for Adults With Depression. *JAMA Psychiatry* **79**, 406–416, DOI (2022).

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

49. A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, P.-C. Bürkner, Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC. *Bayesian Analysis* **16**, 667–718, DOI (2021).

50. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432, DOI (2016).

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

52. S. Talts, M. Betancourt, D. Simpson, A. Vehtari, A. Gelman, *Validating Bayesian Inference Algorithms with Simulation-Based Calibration*, 2020, Preprint on *arXiv*.

53. M. Modrák *et al.*, Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity. *Bayesian Analysis*, 1–29, DOI (2023).

54. N. D. Daw, in *Decision Making, Affect, and Learning* (Oxford University Press, 2011), DOI.

Supplementary Material

Supplementary Methods

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.* (2024) (*17*) for full details) and was conducted using G*Power 3.1 (*40*). We 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 $\alpha = 0.05$). 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.* (2024) (*17*), 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 (*40*) 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).

Learning training task

Participants completed three blocks of twenty 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.* (2024) (*17*)). 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 solvable 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 S2*a*,*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 S4*a*,*c*).

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.

Learning control task

Response options (objects) were 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 (*41*). 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.* human-made-smaller objects for scenario 2, and natural-smaller *vs.* objects human-made-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.

Self-report measures

Symptoms of low mood were measured using the PHQ-9 (*42*). A brief measure of social anxiety symptoms, the miniSPIN (*43*), was also included given our previous observations that social anxiety is relatively elevated in online research participation samples. The DAS-SF, a measure of negative self-beliefs observed in some depressed people (*44*), was included as it has previously been shown to be sensitive to cognitive treatment of low mood (*45*).

The demographic measure included questions about participant gender identity, age, neurodivergence (explained to participants 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 (*46*), and financial, housing, and employment status (as per Buckman *et al.* (2022) (*47*)). 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 (*48*).

Initial statistical analysis of learning task data

Choice accuracy (whether the chosen response option was correct or not) and choice reaction times (RTs) were modelled in linear mixed-effects regression models as

$$\text{accuracy} \sim \text{trialWithinBlock} * \text{eventValence} * \text{scenarioNo} + (1 \mid \text{subID}) \quad \text{(S1)}$$

$$\text{RT} \sim \text{trialWithinBlock} * \text{eventValence} * \text{scenarioNo} + (1 \mid \text{subID}). \quad \text{(S2)}$$

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

$$\text{value} \sim \text{eventValence} * \text{scenarioNo} + (1 \mid \text{subID}). \quad \text{(S3)}$$

Hierarchical Bayesian modelling

Model fit procedure. MCMC chains were initiated with random starting values, and posterior distributions were formed using four chains of 4,000 iterations, with 2,000 discarded warm-up samples (i.e., 8,000 kept iterations per model). Convergence of sampling chains was assessed via inspection of trace plots, and by using standard diagnostics: bulk and tail effective sample size > 100 and split $\hat{R}$ statistics < 1.05 for each parameter (*49*).

Model comparison. 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. Specifically, models were compared in terms of expected log pointwise predictive density ($\text{ELPD}_{\text{diff}}$) using the R package `loo` (*50*). For experimental effects of interest, parameters were assessed using 90% CrIs, with a 90% CrI excluding zero interpreted as representing evidence for a meaningful contribution to posterior parameter estimates.

Hierarchical Bayesian modelling of causal attribution task data

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:

$$\begin{aligned} y_{\text{internal},p,t,v} &\sim \text{Bernoulli}(\theta_{\text{internal},p,t,v}) \\ y_{\text{global},p,t,v} &\sim \text{Bernoulli}(\theta_{\text{global},p,t,v}) \end{aligned} \quad \text{(S4)}$$

where $\theta_{\text{internal},p,t,v}$ and $\theta_{\text{global},p,t,v}$ 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:

$$\begin{aligned}
\begin{bmatrix} \theta_{\text{internal},1,\text{neg}} \\ \theta_{\text{global},1,\text{neg}} \\ \theta_{\text{internal},2,\text{neg}} \\ \theta_{\text{global},2,\text{neg}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} \theta_{\text{internal},\mu,1,\text{neg}} \\ \theta_{\text{global},\mu,1,\text{neg}} \\ \theta_{\text{internal},\mu,2,\text{neg}} \\ \theta_{\text{global},\mu,2,\text{neg}} \end{bmatrix}, \sigma_{\theta,\text{neg}} \right) \\
\begin{bmatrix} \theta_{\text{internal},1,\text{pos}} \\ \theta_{\text{global},1,\text{pos}} \\ \theta_{\text{internal},2,\text{pos}} \\ \theta_{\text{global},2,\text{pos}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} \theta_{\text{internal},\mu,1,\text{pos}} \\ \theta_{\text{global},\mu,1,\text{pos}} \\ \theta_{\text{internal},\mu,2,\text{pos}} \\ \theta_{\text{global},\mu,2,\text{pos}} \end{bmatrix}, \sigma_{\theta,\text{pos}} \right)
\end{aligned} \tag{S5}$$

where $\theta_{\text{internal},\mu,t,v}$ and $\theta_{\text{internal},\mu,t,v}$ 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.

Participant-level parameter estimates were constructed using non-centered reparametrisations in order to separate the hierarchical parameters and lower-level parameters in the prior (*51*). 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 ($\tilde{\phi}_{p,t}$) for all participants p. The between-subjects effects of intervention group were then modelled as:

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{\text{CR}}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2}, & \text{if control intervention + learning task} \end{cases}
\end{aligned} \tag{S6}$$

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., $\phi_{\text{CR}} \sim N(0, 1)$).

For study 2, this model was augmented to include potential group-level effects of allocation the

learning training task condition (ϕ_{LEARN}):

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

$$\phi_{p,2} = \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{\text{CR}} + \phi_{\text{LEARN}}, & \text{if CR intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{\text{LEARN}}, & \text{if control intervention + learning task} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{\text{CR}}, & \text{if CR intervention + control learning task} \end{cases} \tag{S7}$$

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 $\phi_{\sigma,s} \sim \text{Cauchy}(0,1)$. Priors for individual participant deviations from group-level parameter estimates ($\theta_{\text{internal},p,t,\text{neg}}$, $\theta_{\text{internal},p,t,\text{pos}}$, $\theta_{\text{global},p,t,\text{neg}}$, $\theta_{\text{global},p,t,\text{pos}}$) 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 $\text{LKJ}(1)$ 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}, β_{CONTROL}, $\beta_{\text{LEARN+CR}}$), were also specified as $N(0,1)$.

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

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. 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 (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 five times than the standard error of the estimate (*50*); 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:

$$Q_{v,c,t} = Q_{v,c,t} + \alpha_{v,p} * (\text{outcome}_{p,t} - Q_{v,c,t}) \tag{S8}$$

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). Although greater value update rates are not optimal in all settings, given the use of a deterministic reinforcement schedule in our task, here larger posterior α parameter estimates are interpreted as representing faster learning.

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:

$$y_{p,t} \sim \text{CategoricalLogit}(\beta_p * \left[Q_{v,:,t}\right]); \tag{S9}$$

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 ($\text{ELPD}_{\text{diff}} = -124.8$, $\text{SE}_{\text{diff}} = 14.6$).

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 $\phi_{\sigma,s} \sim \text{Cauchy}(0, 1)$. Priors for individual participant deviations from group-level parameter estimates 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 $[0, 20]$.

Simulation-based calibration (SBC) analysis. SBC analysis was used to validate inference procedures for the learning task models (*52*). 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 $[0, D]$ (*53*). 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` (*53*), and are available for the chosen learning model in Figure S10.

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 (*54*).

Supplementary Results

Initial statistical analysis of learning task data

Response accuracy. Choice data for the learning task is shown in Figure S2*a*,*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, $\text{scenario} * \text{trial number}$ 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 statistical evidence of an 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, $\text{valence} * \text{trial number}$ interaction, $\text{valence} * \text{trial} * \text{scenario number}$ 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 S2*a*,*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 ($\text{valence} * \text{trial} * \text{scenario number}$ 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 S4*c*), with this knowledge improving over the task ($\text{valence} * \text{scenario number}$ 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 ($\text{scale} * \text{valence}$ 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 (*35*)) was also able to distinguish ground truth cause types from participants' free text descriptions of each scenario (Figure S4*d*). Specifically, there was statistical evidence of differences in output label probabilities in the expected direction for the internal-external ("myself", "other people") dimension ($\text{event valence} * \text{label}$ 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 ($\text{valence} * \text{label} * \text{scenario number}$ 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 was statistical evidence of 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 ($\text{valence} * \text{label} * \text{scenario number}$ 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 S5), 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, there was limited evidence of any associations (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 S2*c*). 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; $\text{task type} * \text{valence}$ interaction, $F_{1,9662} = 284.7, p < 0.001$), and did not show slower learning over the task for 'positive' events ($\text{task type} * \text{valence} * \text{scenario number}$, $\text{tasktype} * \text{valence} * \text{trial number}$, and $\text{task type} * \text{valence} * \text{trial} * \text{scenario number}$ 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, there was strong statistical evidence that choice times were 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 S2*d*). Control task participants were also not slower to choose response options for 'positive' (red basket) stimuli ($\text{task type} * \text{valence}$ interaction, $F_{1,9662} = 4.1, p < 0.05$).

Ratings values for the control task were substantially less variable and more extreme (Figure S4*c*), suggesting that the response dimensions for this task were more explicit and easily parsed by participants ($\text{task type} * \text{valence}$ interaction in both tasks model, $F_{1,1782} = 33.8, p < 0.001$).

When 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), there was strong evidence that output label probabilities were lower across response dimensions (main effect of task type, $F_{1,162} = 139.6, 46.7, p < 0.001$), as well as some evidence they were not sensitive to trial 'valence' ($\text{task type} * \text{valence} * \text{label}$ 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.

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 S8, Figure S9). Interestingly, whilst in study 1, there was no evidence of a relationship between learning rates and current depression symptom severity (PHQ-9, $R < 0.1$), 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 S8).

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).

Subgroup analysis in participants with clinically elevated depression symptoms. In an exploratory analysis, we investigated whether our main findings held in subsamples of participants who reported clinically elevated depression symptoms on the PHQ-9 (i.e., PHQ-9 total score $\geq$ 10 (*42*); study 1: N=63, study 2: N=42).

In the study 1 subsample, we replicated simple bivariate associations between learning rate estimated on the learning training task and change in mean posterior internal attribution of positive events on the causal attribution task (whole group r=0.361, $p < 0.001$; for correlation with change in global attribution of positive events, r=0.141, p=0.113). From the joint model of learning and causal attribution task data, we also found evidence of a meaningful contribution of learning rate estimates to change in internal and global attributions of positive events in this subsample (β_{LEARN} internal-positive = 0.48 [90% CI = (0.24, 0.87)], β_{LEARN} global-positive = 0.27 [90% CI = (0.13, 0.50)]). We considered the sample too small to further break down by intervention condition.

In the study 2 subsample, likely at least part due to the small number of available data points, we did not replicate the simple bivariate association between learning rate estimated on the learning training task and change in mean posterior internal attribution of positive events on the causal attribution task (r=0.141, p=0.280) and found only very weak statistical evidence of an association with change in global attribution of positive events (r=0.244, p=0.083). From the joint model of learning and causal attribution task data, however, we again found evidence of a meaningful contribution of learning rate estimates to change in internal and global attributions of positive events, for the learning training but not control learning task (β_{LEARN} internal-positive=0.52 [90% CI = (0.22, 0.94], β_{LEARN} global-positive = 0.47 [90% CI = (0.20, 0.86)]; β_{CONTROL} internal-positive = -0.01 [90% CI = (-0.08, 0.08)], β_{CONTROL} global-positive = -0.05 [90% CI = (-0.13, 0.04)]).

Together, these data provide preliminary evidence to suggest that one of our key findings—that learning rates during specific learning training are positively associated with changes in attribution tendencies—may hold in clinical populations, though we emphasise that these analyses are likely underpowered.

Supplementary Figures

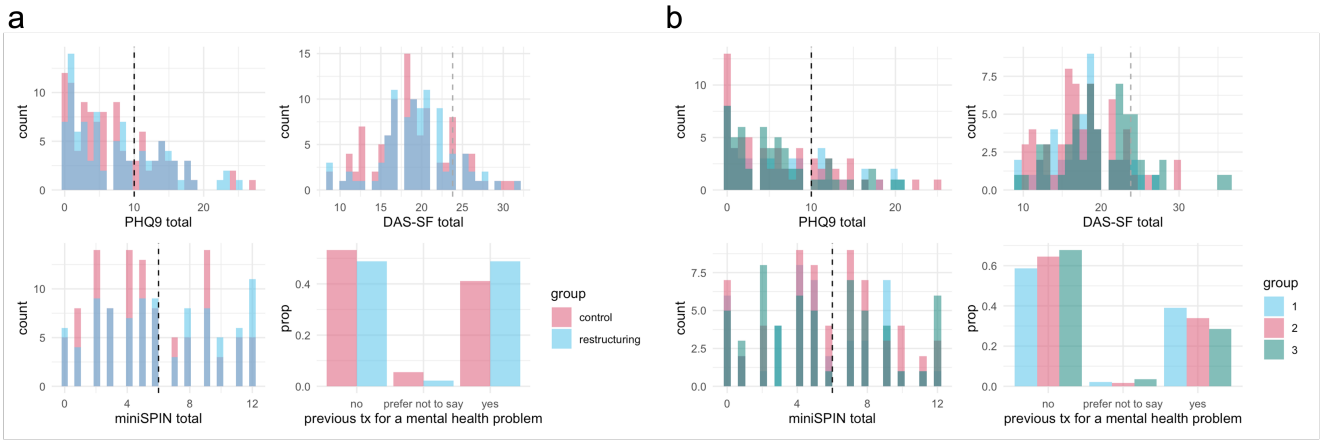

Figure S1: Distribution of self-reported clinical scores for both studies. **a** Study 1 participants. The restructuring group represents participants randomized to the cognitive restructuring intervention, with the control group representing participants randomized the control (emotion-focused) intervention. 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. The 9-item patient health questionnaire assesses depressed mood; the short-form dysfunctional attitudes scale assesses dysfunctional beliefs; and the 3-item social phobia inventory assesses social anxiety. Black dotted lines represent previously-published cut-off scores for clinically-significant levels of symptoms on the PHQ-9 and miniSPIN. 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.

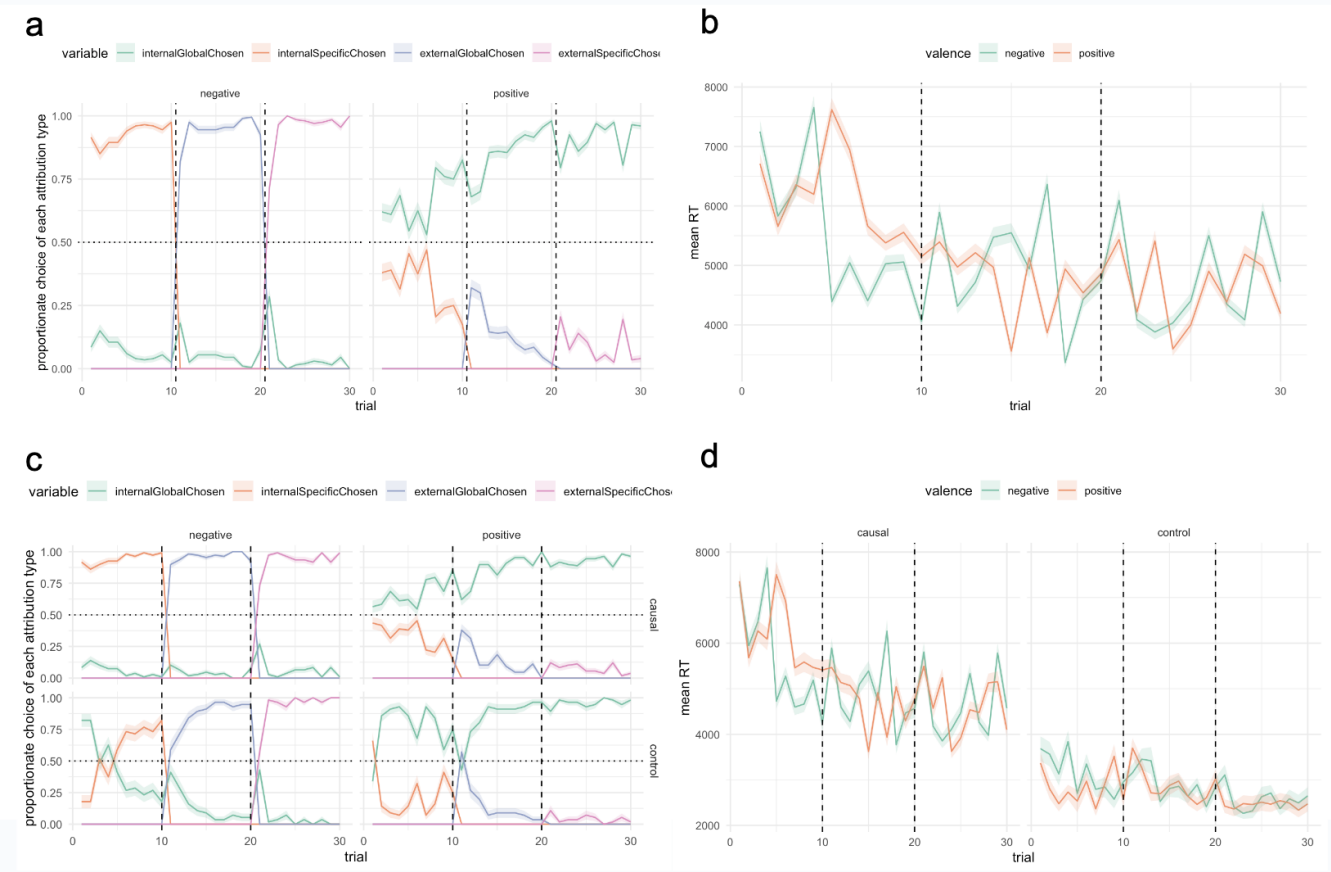

Figure S2: Choice accuracy and response time data for the learning training tasks. **a** Study 1 learning task choice accuracy data. Participants were instructed that they would learn about three different scenarios, each of which represented a different kind of mood a person could be in. For each scenario, 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.

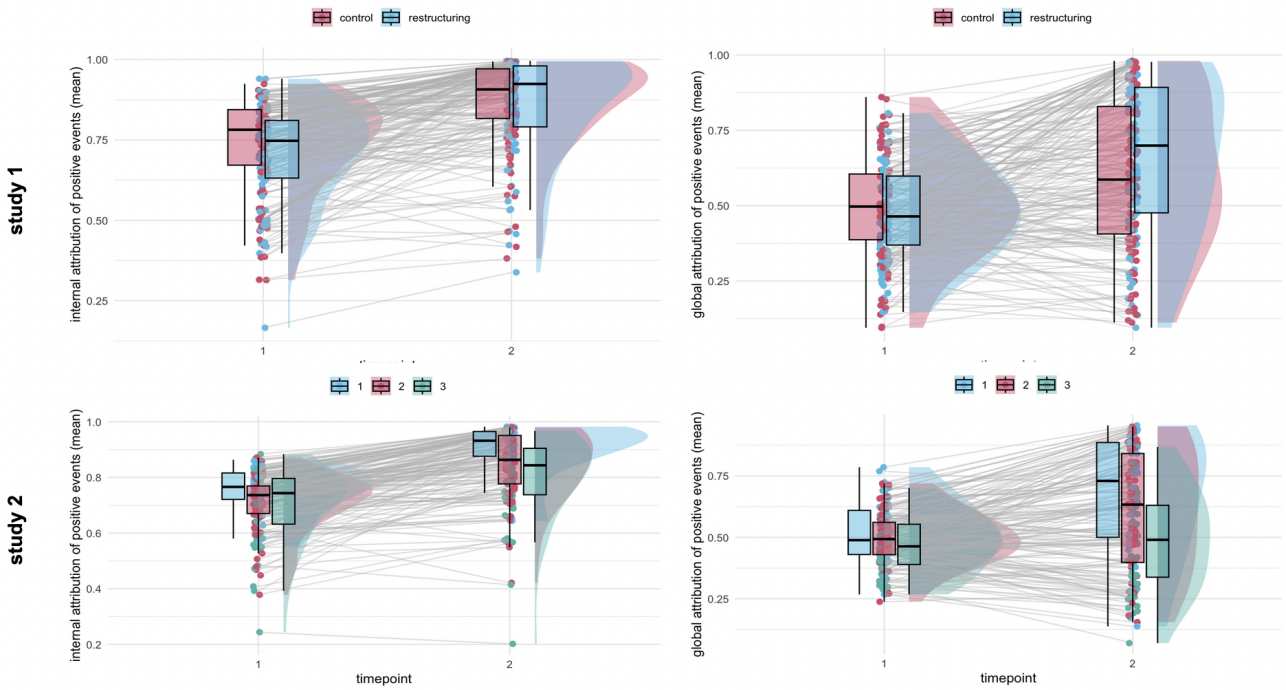

Figure S3: Posterior mean parameter estimates for internal and global attribution of positive events at pre- and post-intervention time-points. Top row, data from study 1. Bottom row, data from study 2. In all plots, time-point 1 is pre-intervention, and time-point 2 is post-intervention. In study 1, interventions consisted of the learning training task plus either brief cognitive restructuring or a control intervention. In study 2, group 1 completed the learning training task and brief restructuring intervention, group 2 completed the learning training task and control intervention, and group 3 completed the control learning task and brief restructuring intervention.

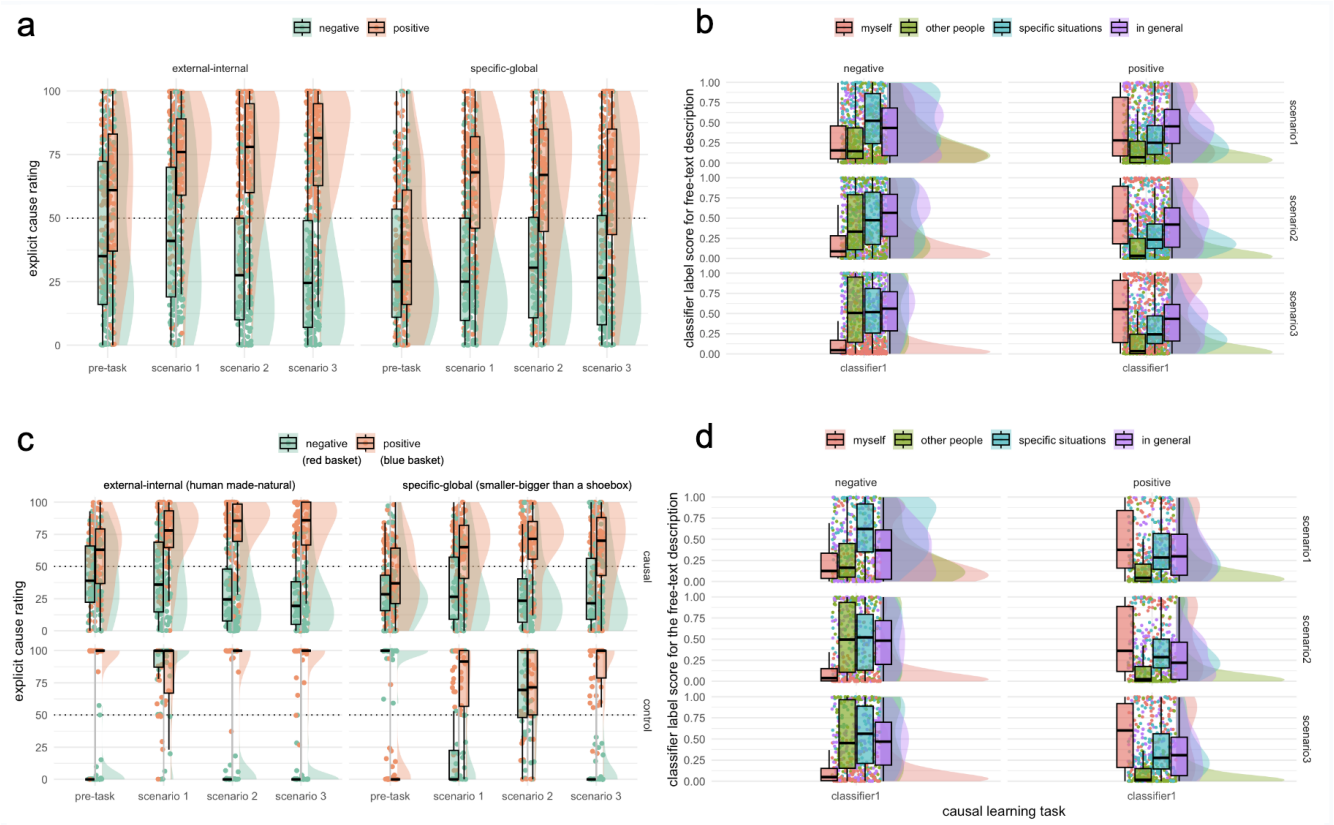

Figure S4: Explicit ratings and free-text description data from the learning training tasks. **a** Within-task explicit cause ratings data for study 1. After each scenario, participants were asked to rate the kinds of causes of events that were thought to be correct, along two separate dimensions (external-internal 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 (*35*)), 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). **c** The same plot as a, for study 2 data, for the causal attribution learning training task (top row), and control learning task (bottom row). Labels in brackets describe the equivalent dimensions in the control learning task. **d** The same plot as b, for study 2 data (causal learning task participants only).

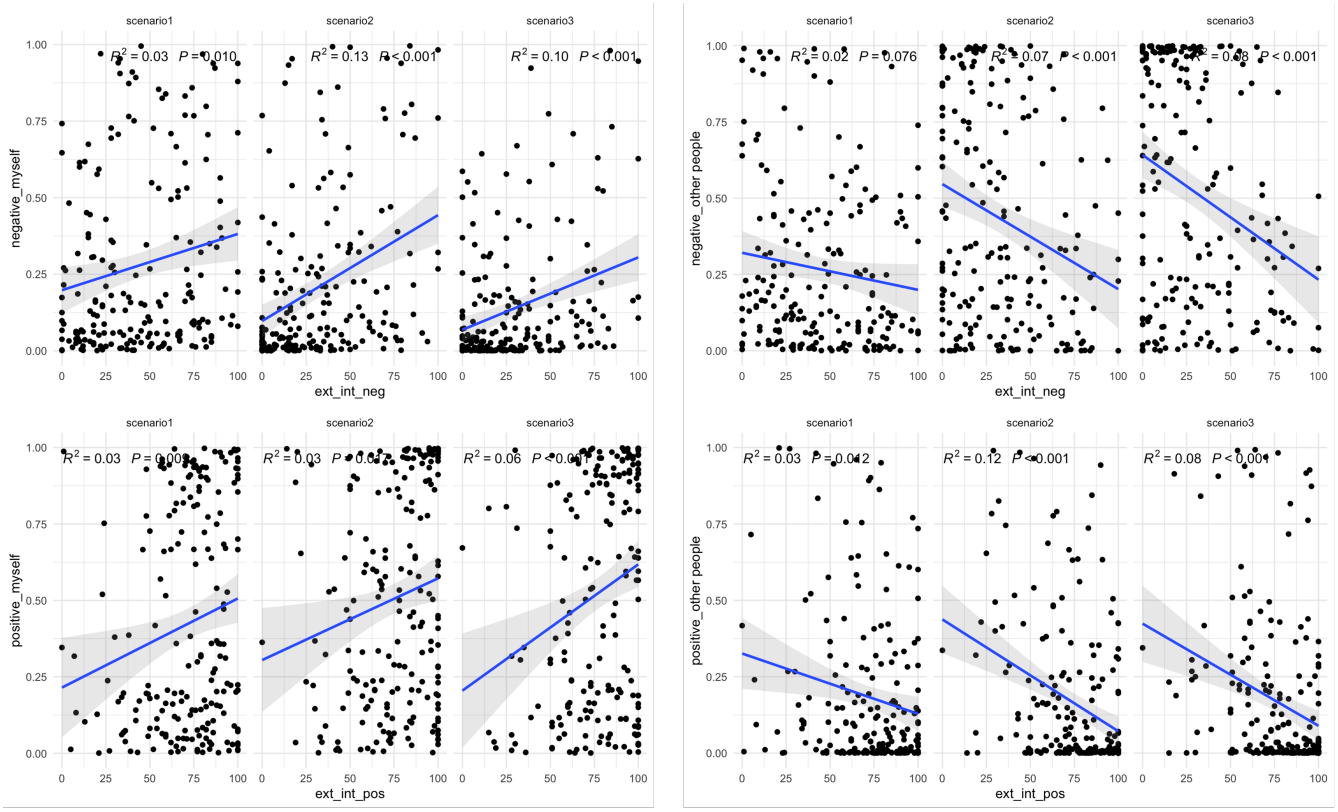

Figure S5: Associations between explicit cause ratings and classifier label probabilities for free-text descriptions of causes from the learning training task (study 1). X axes, explicit ratings of cause types following each scenario, on the external-internal response dimension. Y axes, classifier output probabilities for post-scenario free text descriptions of causes, for the labels [caused by] "myself" and [caused by] "other people".

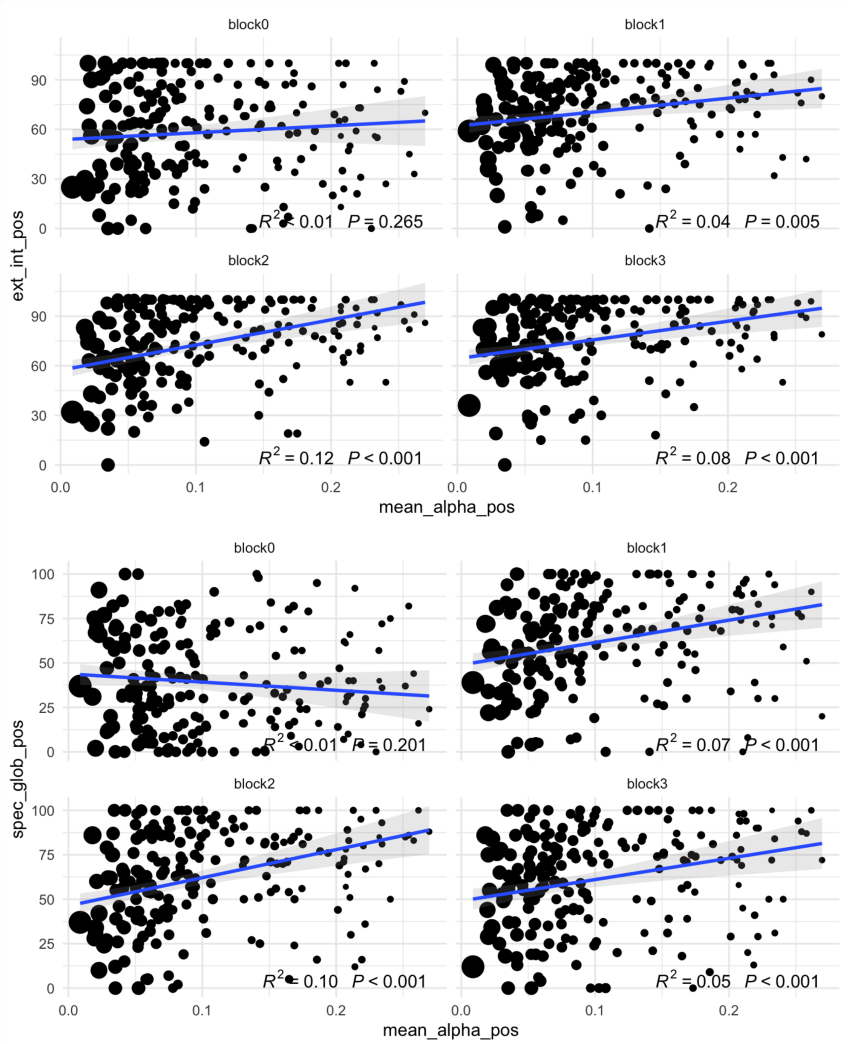

Figure S6: Correlations between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and within-task explicit cause type ratings. Pre-task ratings are ratings provided by participants prior to starting the task, during which they are asked to reflect 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" or "completely by myself") and specific-global dimension scale (events were caused "by things related to the specific circumstances" or "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 posterior precision of the estimate (1/SD).

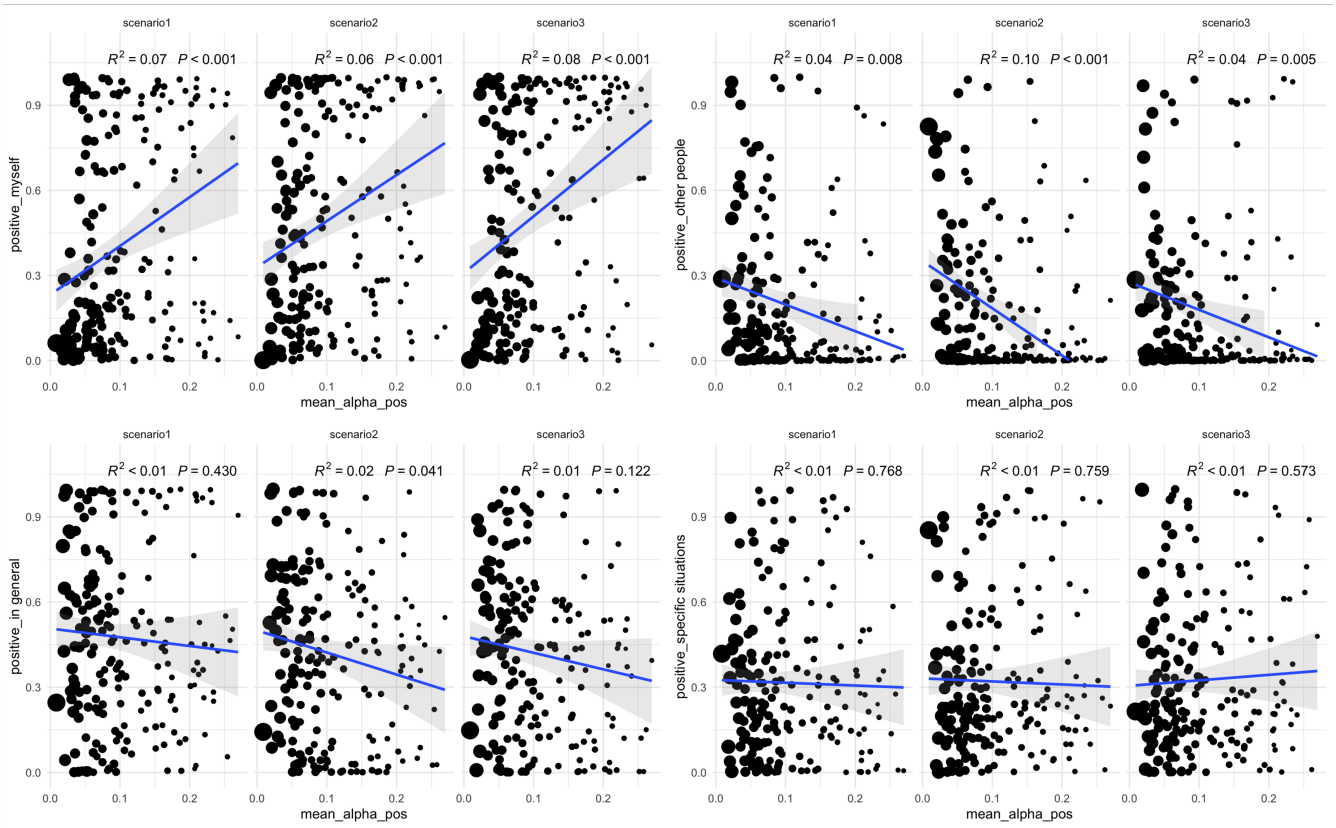

Figure S7: Correlations between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and label classification probabilities of free-text descriptions of correct cause types (study 1) Scenarios 1-3 represent classifier output for free-text descriptions 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).

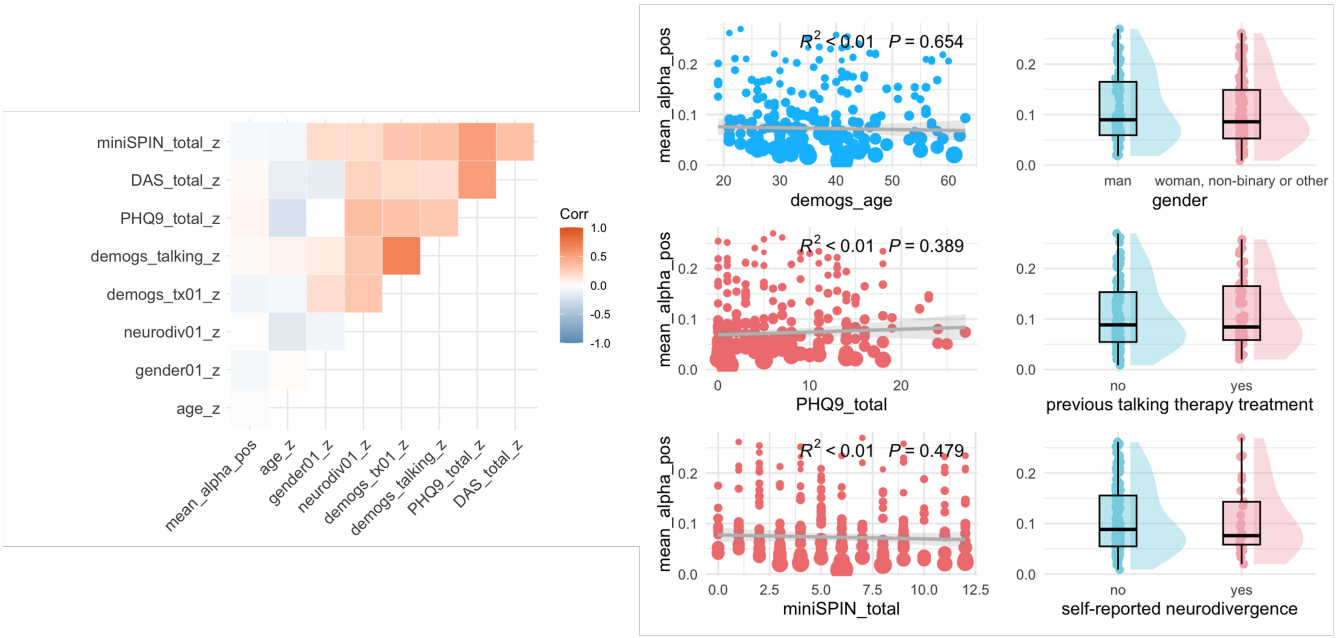

Figure S8: Relationships between mean posterior estimates of learning rates for positive events on the learning training task (α_{pos}) and self-reported participant demographic and clinical information (study 1).

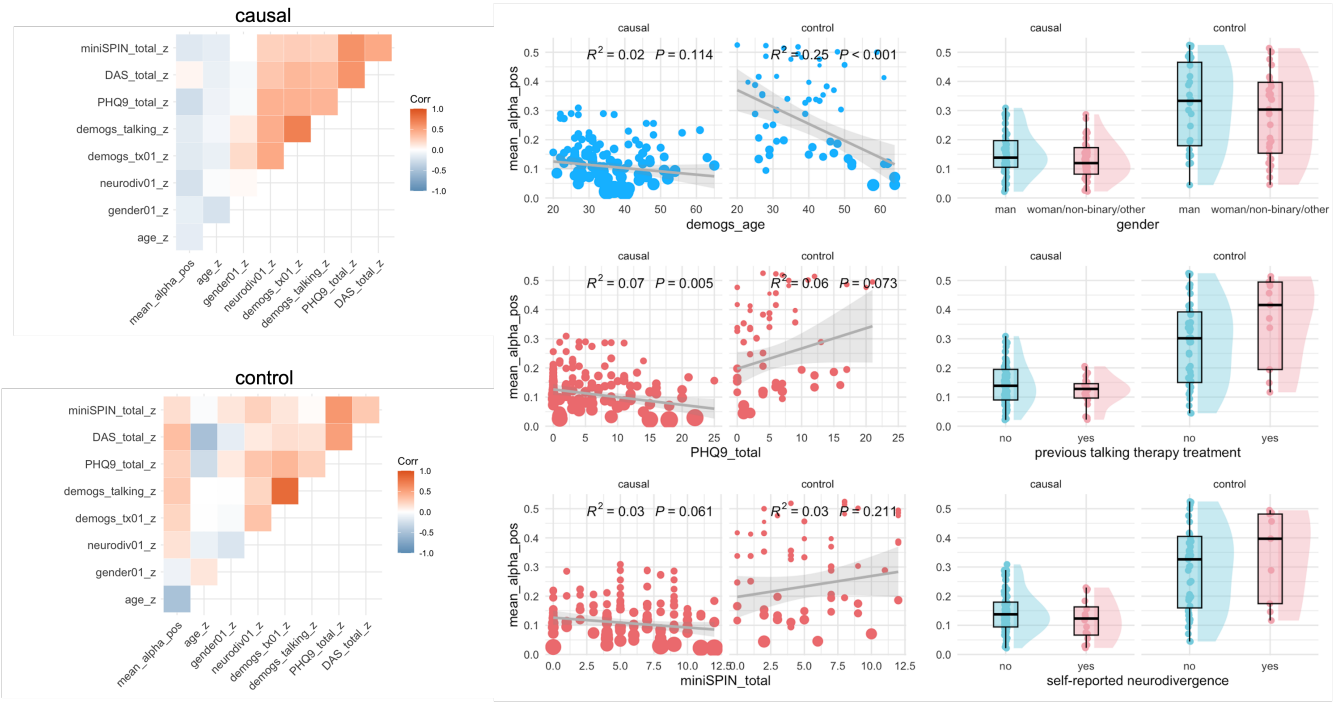

Figure S9: Relationships between mean posterior estimates of learning rates for positive events from the learning training and and control learning tasks (α_{pos}), and self-reported participant demographic and clinical information (study 2).

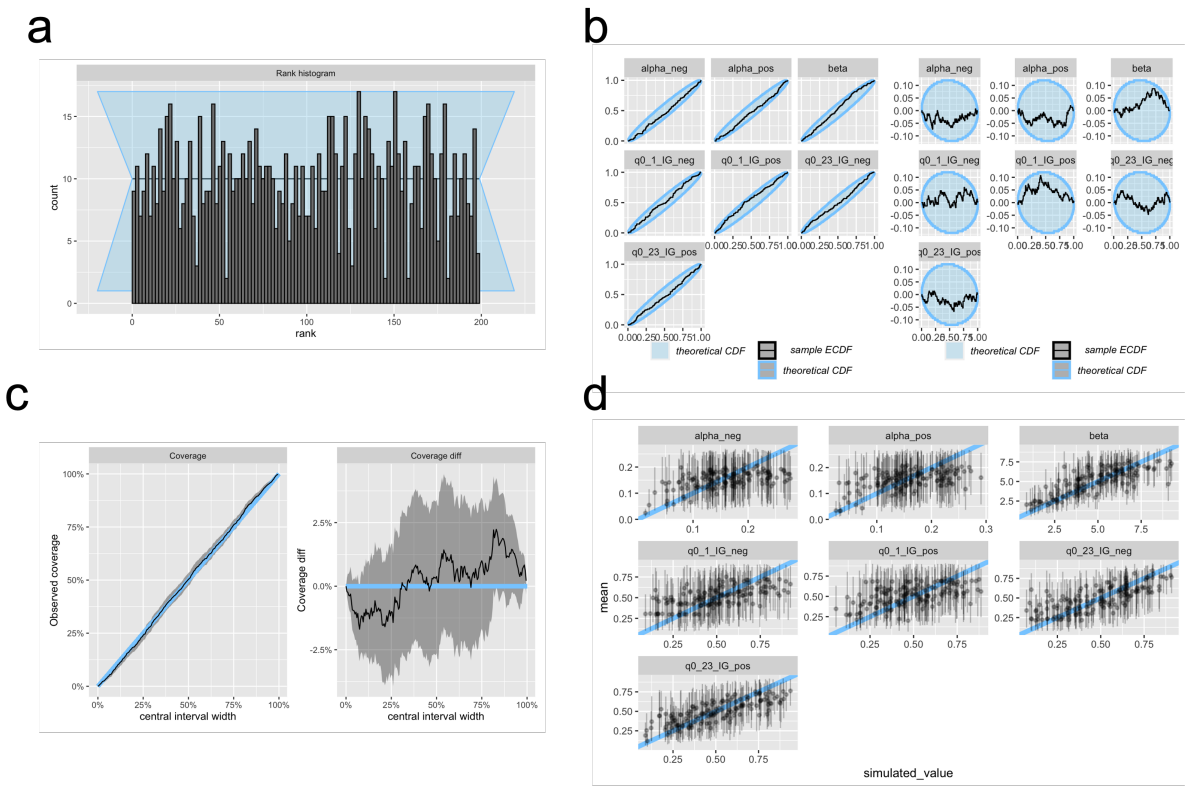

Figure S10: Simulation-based calibration analysis for the learning training task. a Rank histogram, a check for uniformity of posterior draw ranks. Horizontal black line, expected average count; blue trapezoid, approximate 95% interval for expected deviations over ranks. **b** (E)CDF, (empirical) cumulative 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	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Mean θ for internal attributions of negative events at time 1	-0.220	0.059	-0.335	-0.317	-0.125	-0.106	1335
Mean θ for internal attributions of negative events at time 2	-0.520	0.093	-0.704	-0.671	-0.370	-0.339	1557
Mean θ for internal attributions of positive events at time 1	1.084	0.070	0.946	0.970	1.199	1.222	1615
Mean θ for internal attributions of positive events at time 2	2.520	0.185	2.171	2.230	2.834	2.891	1310
Mean θ for global attributions of negative events at time 1	-0.582	0.051	-0.685	-0.667	-0.499	-0.484	2150
Mean θ for global attributions of negative events at time 2	-0.734	0.064	-0.866	-0.842	-0.628	-0.610	2778
Mean θ for global attributions of positive events at time 1	-0.061	0.066	-0.187	-0.168	0.046	0.069	1413
Mean θ for global attributions of positive events at time 2	0.665	0.166	0.344	0.395	0.942	0.989	1550
Effect of restructuring on θ internal-negative at time 2	-0.479	0.133	-0.740	-0.698	-0.264	-0.219	1925
Effect of restructuring on θ internal-positive at time 2	0.325	0.263	-0.178	-0.100	0.766	0.847	1134
Effect of restructuring on θ global-negative at time 2	0.073	0.089	-0.103	-0.074	0.219	0.253	3020
Effect of restructuring on θ global-positive at time 2	0.498	0.237	0.039	0.114	0.895	0.965	1497

Table S1: Hierarchical Bayesian model results for effects of cognitive restructuring on causal attribution tendencies in study 1 data *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, 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). All values are raw (untransformed) parameter estimates (for transformation constraints applied to main text figures see Supplementary Methods).

	mean	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Mean θ for internal attributions of negative events at time 1	-0.280	0.070	-0.415	-0.393	-0.167	-0.145	2832
Mean θ for internal attributions of negative events at time 2	-0.103	0.189	-0.467	-0.411	0.207	0.269	3510
Mean θ for internal attributions of positive events at time 1	1.044	0.075	0.899	0.921	1.170	1.195	3263
Mean θ for internal attributions of positive events at time 2	1.177	0.329	0.526	0.633	1.715	1.823	3208
Mean θ for global attributions of negative events at time 1	-0.645	0.069	-0.784	-0.760	-0.532	-0.512	3782
Mean θ for global attributions of negative events at time 2	-0.678	0.174	-1.025	-0.966	-0.395	-0.338	3719
Mean θ for global attributions of positive events at time 1	-0.020	0.068	-0.153	-0.131	0.091	0.111	3011
Mean θ for global attributions of positive events at time 2	-0.306	0.314	-0.930	-0.825	0.217	0.314	3274
Effect of restructuring on θ internal-negative at time 2	-0.315	0.157	-0.627	-0.574	-0.060	-0.005	4082
Effect of restructuring on θ internal-positive at time 2	0.648	0.282	0.108	0.185	1.113	1.206	3458
Effect of restructuring on θ global-negative at time 2	0.029	0.143	-0.244	-0.204	0.268	0.315	4123
Effect of restructuring on θ global-positive at time 2	0.336	0.268	-0.183	-0.107	0.776	0.853	3688
Effect of learning training on θ internal-negative at time 2	-0.514	0.157	-0.825	-0.774	-0.260	-0.208	4082
Effect of learning training on θ internal-positive at time 2	1.237	0.285	0.688	0.774	1.720	1.805	3424
Effect of learning training on θ global-negative at time 2	-0.138	0.146	-0.421	-0.377	0.098	0.150	4054
Effect of learning training on θ global-positive at time 2	1.025	0.269	0.485	0.577	1.467	1.546	3521

Table S2: Hierarchical Bayesian model results for effects of cognitive restructuring and learning training on causal attribution tendencies in study 2 data. *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, 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). All values are raw (untransformed) parameter estimates (for transformation constraints applied to main text figures see Supplementary Methods).

	mean	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Effect of restructuring on θ internal-negative at time 2	-0.477	0.131	-0.732	-0.695	-0.261	-0.222	2428
Effect of restructuring on θ internal-positive at time 2	0.234	0.214	-0.182	-0.116	0.587	0.658	2422
Effect of restructuring on θ global-negative at time 2	0.071	0.090	-0.106	-0.077	0.218	0.246	4051
Effect of restructuring on θ global-positive at time 2	0.411	0.196	0.034	0.089	0.736	0.800	2700
β_{LEARN}, effect of α_{pos} on θ internal-positive at time 2	0.562	0.175	0.306	0.337	0.884	0.975	554
β_{LEARN}, effect of α_{pos} on θ global-positive at time 2	0.457	0.141	0.246	0.272	0.715	0.795	458

Table S3: Joint hierarchical model 1 results for study 1 data. *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size. 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	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Effect of restructuring on θ internal-negative at time 2	-0.314	0.151	-0.605	-0.562	-0.065	-0.024	3604
Effect of restructuring on θ internal-positive at time 2	0.678	0.294	0.105	0.196	1.162	1.262	4754
Effect of restructuring on θ global-negative at time 2	0.027	0.139	-0.247	-0.203	0.253	0.295	3760
Effect of restructuring on θ global-positive at time 2	0.351	0.282	-0.196	-0.114	0.818	0.906	3955
Effect of learning training on θ internal-negative at time 2	-0.511	0.153	-0.816	-0.763	-0.265	-0.213	3361
Effect of learning training on θ internal-positive at time 2	-0.900	0.525	-1.930	-1.754	-0.025	0.167	1920
Effect of learning training on θ global-negative at time 2	-0.143	0.144	-0.430	-0.380	0.092	0.131	3381
Effect of learning training on θ global-positive at time 2	-0.960	0.542	-2.033	-1.835	-0.077	0.121	1190
β_{LEARN}, effect of α_{pos} on θ internal-positive at time 2	0.293	0.090	0.149	0.167	0.453	0.497	1175
β_{LEARN}, effect of α_{pos} on θ global-positive at time 2	0.259	0.087	0.116	0.134	0.415	0.450	909
β_{CONTROL}, effect of α_{pos} on θ internal-positive at time 2	0.006	0.016	-0.022	-0.017	0.034	0.043	869
β_{CONTROL}, effect of α_{pos} on θ global-positive at time 2	0.007	0.014	-0.019	-0.014	0.031	0.038	1088

Table S4: Joint hierarchical model 1 results for study 2 data. *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size. 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	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Effect of restructuring on θ internal-negative at time 2	-0.477	0.134	-0.740	-0.696	-0.255	-0.214	1794
Effect of restructuring on θ internal-positive at time 2	-0.385	0.469	-1.317	-1.164	0.371	0.516	1102
Effect of restructuring on θ global-negative at time 2	0.071	0.092	-0.106	-0.080	0.223	0.253	2979
Effect of restructuring on θ global-positive at time 2	0.161	0.387	-0.617	-0.483	0.787	0.921	2232
β_{LEARN}, effect of α_{pos} on θ internal-positive at time 2	0.579	0.187	0.288	0.327	0.922	1.021	711
β_{LEARN}, effect of α_{pos} on θ global-positive at time 2	0.474	0.159	0.225	0.257	0.770	0.849	524
$\beta_{\text{LEARN+CR}}$, additional effect of α_{pos} on θ internal-positive at time 2 in restructuring group	0.229	0.164	-0.042	-0.002	0.516	0.602	1178
$\beta_{\text{LEARN+CR}}$, additional effect of α_{pos} on θ global-positive at time 2 in restructuring group	0.095	0.105	-0.093	-0.062	0.284	0.337	1753

Table S5: Joint hierarchical model 2 results, for study 1 data. *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size. 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	**s.d.**	**2.5%**	**5%**	**95%**	**97.5%**	N_{eff}
Effect of restructuring on θ internal-negative at time 2	-0.318	0.151	-0.614	-0.564	-0.068	-0.023	2959
Effect of restructuring on θ internal-positive at time 2	1.060	0.534	0.022	0.185	1.942	2.110	1653
Effect of restructuring on θ global-negative at time 2	0.029	0.143	-0.256	-0.206	0.262	0.308	2679
Effect of restructuring on θ global-positive at time 2	-0.176	0.495	-1.149	-0.991	0.630	0.800	2272
Effect of learning training on θ internal-negative at time 2	-0.515	0.152	-0.819	-0.767	-0.265	-0.217	2904
Effect of learning training on θ internal-positive at time 2	-0.689	0.550	-1.757	-1.589	0.210	0.379	1681
Effect of learning training on θ global-negative at time 2	-0.137	0.146	-0.425	-0.376	0.104	0.152	2664
Effect of learning training on θ global-positive at time 2	-1.377	0.575	-2.505	-2.320	-0.427	-0.215	1357
β_{LEARN}, effect of α_{pos} on θ internal-positive at time 2	0.353	0.118	0.172	0.195	0.569	0.637	836
β_{LEARN}, effect of α_{pos} on θ global-positive at time 2	0.282	0.103	0.125	0.145	0.462	0.523	693
$\beta_{\text{LEARN+CR}}$, additional effect of α_{pos} on θ internal-positive at time 2 in restructuring group	-0.066	0.094	-0.260	-0.226	0.083	0.111	1815
$\beta_{\text{LEARN+CR}}$, additional effect of α_{pos} on θ global-positive at time 2 in restructuring group	0.089	0.085	-0.060	-0.036	0.236	0.276	1557
β_{CONTROL}, effect of α_{pos} on θ global-positive at time 2	0.007	0.021	-0.026	-0.019	0.039	0.053	438
β_{CONTROL}, effect of α_{pos} on θ internal-positive at time 2	0.012	0.020	-0.016	-0.012	0.043	0.058	339

Table S6: Joint hierarchical model 2 results, for study 2 data. *mean, s.d.*, posterior mean and SD; *2.5%, 5%, 95%, 97.5%*, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size. 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**	$\mathrm{ELPD}_{\mathrm{diff}}$ **(choice data)**	$\mathrm{SE}_{\mathrm{diff}}$ **(choice data)**
base model	Model of choice data only (pre and post-intervention), as described in Norbury *et al.* (2024) (*17*).	-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. $\mathrm{ELPD}_{\mathrm{diff}}$, difference in ELPD for each model from the best model, which is defined as having zero difference to itself. $\mathrm{SE}_{\mathrm{diff}}$, the standard error of this difference.

model	**description**	***params***	$\mathrm{ELPD}_{\mathrm{diff}}$	$\mathrm{SE}_{\mathrm{diff}}$
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. $\mathrm{ELPD}_{\mathrm{diff}}$, difference in expected log pointwise predictive density for each model from the best model, which is defined as having zero difference to itself. $\mathrm{SE}_{\mathrm{diff}}$, 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. *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.