

# **Latent cause inference during extinction learning in trauma-exposed individuals with and without PTSD**

Agnes Norbury<sup>1</sup>, Hannah Brinkman<sup>1</sup>, Mary Kowalchuk<sup>1</sup>, Elisa Monti<sup>1</sup>, Robert H Pietrzak<sup>2,3</sup>, Daniela Schiller<sup>1,4</sup>, Adriana Feder<sup>1</sup>

<sup>1</sup>Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>2</sup>Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

<sup>3</sup>United States Department of Veterans Affairs National Center for Posttraumatic Stress Disorder, Clinical Neurosciences Division, VA Connecticut Healthcare System, West Haven, CT, USA

<sup>4</sup>Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

## Abstract

Problems in learning that sights, sounds, or situations that were once associated with danger have become safe (extinction learning) may explain why some individuals suffer prolonged psychological distress following traumatic experiences. Although simple associative learning models have been unable to provide a convincing account of how and why this learning fails, it has recently been proposed that this may be explained by individual differences in beliefs about the causal structure of the environment. Here, we tested two competing hypotheses as to how differences in causal inference might be related to trauma-related psychopathology, using extinction learning data collected from clinically well-characterized individuals with varying degrees of post-traumatic stress ( $N=56$ ). Latent cause modelling revealed that individuals with more severe PTSD were more likely to assign observations from conditioning and extinction stages to a single underlying cause. Specifically, multivariate analysis incorporating multiple PTSD and depression symptom dimensions revealed a negative relationship between tendency to infer multiple causes were active in the environment and re-experiencing symptom severity. We interpret these results as providing evidence of a primary deficit in discriminative learning in participants with more severe PTSD re-experiencing symptoms. Specifically, in these individuals, a greater tendency to attribute all stimulus configurations to the same underlying cause resulted in greater uncertainty about stimulus-outcome associations, and impeded learning both that certain stimuli were safe, and that certain stimuli were no longer dangerous. Better understanding of the role of causal inference in trauma-related psychopathology may have relevance for the refinement of cognitive therapies for these disorders.

## Introduction

Post-traumatic stress disorder (PTSD) can be thought of a disorder of inappropriate fear, driven by a failure to update expectations when objects or contexts that were once associated with danger become safe (Lissek & van Meurs, 2015). However, simple associative accounts of learning are unable to account well for why such fear persists – particularly in the face of prolonged exposure (extinction) training, or when considering relapse (spontaneous return of fear) (Dunsmoor et al., 2015). Recently, a novel computational account of extinction learning – latent cause modelling – has been proposed by Gershman, Niv, and colleagues (Gershman & Niv, 2010, 2012). This account posits that during learning, individuals do not simply learn to associate different stimuli or contexts with outcomes, but rather that they attempt to draw inferences about the underlying environmental causes that are responsible for groups of observations (i.e., stimuli and outcomes together). For example, an experimental animal may learn to infer that on different days, or when a different experimenter is present, painful stimuli are unlikely to be presented – rather than having to gradually update their stimulus-outcome associations during every new conditioning or extinction learning session. Individual differences in this inference process regulate whether an individual decides that the same cause is responsible for their current observations (and therefore that the original fear memory should be updated), or whether a new underlying cause is responsible (and therefore the original memory is left intact) (Gershman et al., 2017). According to this account, the inappropriate fear responses observed in post-traumatic stress syndromes could result from two different underlying mechanisms: 1) failure to retrieve a successfully formed extinction memory, as a result of inferring that a different cause is operating in the environment, or 2) failure to successfully form an extinction memory in the first place.

Computationally, the first case can be formalized as heightened tendency to segment ongoing experience into different causal clusters during extinction learning. Simulation evidence suggests that this would be reflected in faster learning during initial extinction training (due to lower conflict between conditioning and extinction trials), but greater vulnerability to relapse or spontaneous return of fear (e.g., if contextual cue changes mean that the old fear memory is retrieved, rather than the new extinction memory) (Gershman & Niv, 2012; Gershman et al., 2015). Indeed, tendency to infer more causes are active across conditioning and extinction episodes has been previously been shown to predict stronger return of physiological fear responses during next-day recall testing in healthy humans (Gershman & Hartley, 2015).

However, a body of evidence also suggests that individuals with PTSD and other anxiety disorders show deficits in aversive processing that may be pre-requisites for successful extinction learning: in particular in the ability to discriminate between safe and danger-associated stimuli, in the context of potential aversive outcomes (pain or monetary loss). For example, both higher arousal to non pain/loss-associated stimuli during initial learning and greater physiological and self-reported aversion responses to all stimuli during extinction training are reliably observed in groups of individuals with anxiety disorders, compared to healthy controls (Duits et al., 2015; Marin et al., 2020). Further, heightened transfer of negative expectations to stimuli that are perceptually similar to fear-associated shapes or sounds has been observed in individuals with post-traumatic stress and anxiety (Lissek & van Meurs, 2015; Kaczurkin et al., 2016; Norbury et al., 2018). Intuitively,

reduced ability to distinguish between (or poorer internal representation of) which stimuli were associated with which outcomes might result in a tendency to assign all observations to a single underlying cause. Importantly, a single underlying cause with poor distinction between different sets of observations could be reflected in both negative expectations for all stimuli (even those never associated with danger), and impeded extinction learning (due to greater uncertainty about stimulus-outcome configurations) (Gershman & Niv, 2012). Therefore, it is possible that the inappropriate negative expectations associated with PTSD are the result of either *heightened* or *reduced* tendency to believe that different causes are responsible for observations during exposure to extinction.

Here, we sought to test these competing hypotheses by investigating latent cause inference during extinction learning in a group of clinically well-characterized trauma-exposed individuals with a range of experience of post-traumatic stress symptoms ( $N=56$ ). Specifically, we investigated whether trauma-exposed individuals with more severe PTSD symptoms would show a pattern of behaviour best explained by a greater or lower tendency to infer novel environmental causes, when compared to trauma-exposed individuals with less severe or no post-traumatic stress. We were particularly interested in whether differences in inference across aversive conditioning and extinction learning were related to individual difference in avoidance symptoms, as inappropriate avoidance behaviour is thought to be a core mechanism maintaining resistance to extinction in anxiety disorders (Arnaudova et al., 2017; Pittig et al., 2020). Following recent theoretical developments that favour modelling psychological disorders as consisting of complex associations of interacting symptoms and other psychosocial factors (Borsboom, 2017), individual differences in latent cause inference were also related to multiple psychological symptom dimensions using network analysis (Greene et al., 2018; Armour et al., 2017; Fritz et al., 2018; de Haan et al., 2020).

The findings presented here represent the first evidence that individual differences in latent cause inference detected using a simple remotely-administered extinction learning paradigm are related to current psychological symptom severity. Ultimately, better understanding of how individual differences in causal inference contribute to maladaptive learning in anxiety and post-traumatic stress may have relevance for the refinement of cognitive and learning-based therapies for these disorders (Moutoussis et al., 2017).

## Results

### Participants

Participants were  $N=56$  adults with DSM-5 Category A trauma exposure to the World Trade Center (WTC) disaster (mean age  $53 \pm 7.0$ ,  $N=19$  female; Table 1). For all participants, current and lifetime experience of WTC-related PTSD and other psychiatric disorders was assessed by clinical interview. Recent experience of PTSD and depression symptoms was further assessed by self-report questionnaires (the PTSD Checklist for DSM-5 [PCL-5], and The Beck Depression Inventory version II [BDI-II]).  $N=42$  (75%) participants currently met DSM-5 criteria for full or subthreshold WTC-related PTSD (mean PCL-5 total score  $40.4 \pm 11.0$ ), and  $N=14$  (25%) participants were highly resilient to WTC trauma (no current or lifetime diagnosis of PTSD or other DSM-5 Axis-1 disorder; mean PCL-5 total score  $1.6 \pm 1.7$ ).  $N=35$  (63%) participants were assessed as experiencing clinically-significant levels of avoidance behaviour (of people, places, or things that reminded them of their WTC experience) during the past month, on the basis of clinical interview. Participants with full or subthreshold PTSD also reported moderate levels of depression symptoms (PTSD group, mean BDI-II total score  $14.5 \pm 9.6$ ; resilient group, mean BDI-II total score  $0.43 \pm 1.1$ ). However, PTSD is highly heterogeneous (Armour et al., 2015; Contractor et al., 2017), and PTSD symptom subscores showed continuous variation across our participants as a group (Table S1) – justifying the use of dimensional approach to PTSD symptomatology for our analysis.

### Extinction learning task

All participants completed an online extinction learning task, analogous in structure to that employed in a previous analysis of latent cause inference during extinction in healthy individuals (Gershman & Hartley, 2015). This task consisted of a three phases: an initial aversive conditioning phase (in context A), extinction learning phase (in context A), and further extinction learning (in novel context B) (Figure 1a). Importantly, the conditioned stimulus (CS) associated with the aversive loss outcome (US) – the CS+ – was only partially reinforced ( $P(\text{US}|\text{CS}+) = 1/3$ ), and the transition to extinction ( $P(\text{US}|\text{CS}+) = 0$ ) was unsignalled. This design maximises uncertainty about whether CS+ trials during extinction should be grouped with unreinforced conditioning phase CS+ trials, implying a common cause is responsible for both kinds of observations, or instead that the change in contingencies indicates it is likely that a new cause is active in the environment.

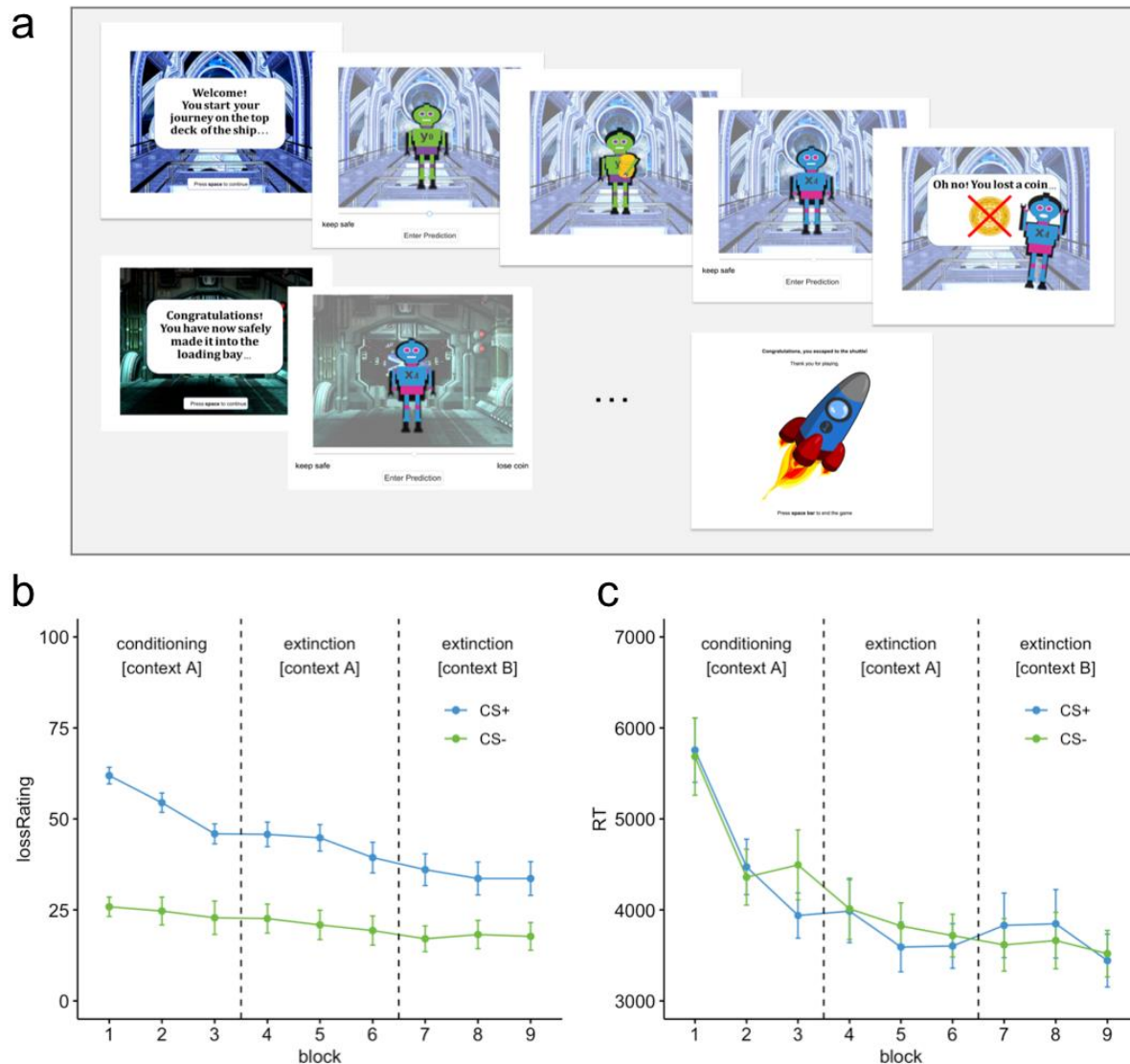
**Manipulation check.** In order to check if participants performed as expected on the task, loss expectancy ratings were analysed by repeated-measures ANOVA (Figure 1b). This analysis revealed significant effects of CS type and task stage on mean loss prediction ratings ( $F_{1,947}=247.1$ ,  $p<2.2\text{e-}16$ ;  $F_{2,947}=27.5$ ,  $p=2.5\text{e-}12$ ; respectively) – with higher ratings for CS+ compared to unconditioned (CS-) stimuli (mean difference  $22.9$  [SE  $1.5$ ], 95% CI  $20.1$ – $25.8$ , points on a scale ranging 1–100), and decreases in mean loss expectancy ratings between conditioning and extinction stages (mean decrease in rating from conditioning to initial extinction stages was  $7.1$  [SE  $1.8$ ] points, 95% CI  $3.0$ – $11.3$ ; mean decrease in rating from initial extinction to further extinction training stages was  $6.1$  [SE  $1.8$ ] points, 95% CI  $1.9$ – $10.3$ ; both differences  $p<0.002$ , Tukey LSD corrected). There was also a significant interaction between CS type and task stage ( $F_{2,947}=6.52$ ,  $p=0.002$ ), with a greater decrease in ratings for CS+ stimuli (mean decrease in CS+ ratings from conditioning to further extinction

training stage of 19.7 [SE 2.5], 95% CI 12.4–26.9,  $p=2.0\text{e-}13$ ; mean decrease in CS- ratings of 6.8 [SE 2.5], 95% CI -0.4–14.0,  $p=0.08$ ). This indicates that overall participants entered greater loss expectancy ratings for the aversively conditioned (CS+) compared to non aversively-conditioned (CS-) stimuli, and decreased their ratings of CS+, but not CS-, stimuli over the course of the task (i.e., when these stimuli began to be presented in extinction).

Response times (median time taken to enter loss prediction ratings) were entered into a similar analysis model. There was a significant effect of experiment stage on response times ( $F_{2,947}=61.8$ ,  $p<2.0\text{e-}16$ ), with quicker response times at later stages in the task (mean decrease in RT from conditioning to initial extinction stages 995ms [SE 111], 95% CI 735–1256,  $p<0.001$ ; no difference in RT from initial to further extinction training stages,  $p>0.4$ ) – but there was no evidence for an effect of CS type or interaction between CS type and experiment stage on response timing (both  $p>0.4$ ). Median response times remained  $>3000\text{ms}$  during the later stages of the task, indicating preservation of relatively considered responding throughout (Figure 1c).

Age	53 (6.9)
Gender ( <i>N</i> female)	19 (34%)
Race ( <i>N</i> )	
Black or African American	6 (11%)
Asian	4 (7%)
Native American	1 (2%)
White or Caucasian	37 (66%)
Other	2 (4%)
Ethnicity ( <i>N</i> )	
Hispanic/Latinx	12 (21%)
Education level ( <i>N</i> )	
Graduated high school (or equivalent)	5 (9%)
Part college	17 (30%)
Graduated 2-year college	5 (9%)
Graduated 4-year college	14 (25%)
Graduate or professional school	15 (27%)
Profession on 11/09/2001 ( <i>N</i> )	
Traditional emergency services responder	23 (41%)
Non-traditional responder or survivor	33 (59%)
PCL-5 total score	30.7 (19.4)
BDI-II total score	10.9 (10.3)
Psychoactive medication ( <i>N</i> )	
SSRI/SNRI (stable dose)	3 (6%)
NDRI (stable dose)	3 (6%)
sedative (night-time use only)	3 (6%)
Additional lifetime trauma history	
<i>N</i> trauma categories endorsed (0–13)	4.9 (2.5)
Childhood physical abuse ( <i>N</i> )	16 (29%)
Childhood sexual abuse ( <i>N</i> )	14 (25%)
Adulthood sexual trauma ( <i>N</i> )	7 (13%)

**Table 1. Summary of demographic and clinical variables for study participants (*N*=56).** Values represent mean (SD) unless otherwise specified. Race/ethnicity and medication status categories are non mutually-exclusive; *N*=8 (14%) individual participants were currently taking a stable dose (>3 months) of a psychoactive medication. PCL-5, PTSD Checklist for DSM-5; BDI-II, Beck Depression Inventory version II; SSRI, selective serotonin reuptake inhibitor; SNRI, serotonin/noradrenaline reuptake inhibitor; NDRI, noradrenaline/dopamine reuptake inhibitor. For further information on PTSD and depression subscore ranges and distributions, see Table S1. All study participants had DSM-5 Category A trauma exposure to the World Trade Center disaster in 2001. For details about additional lifetime trauma categories, and how these were defined, see Table S2.



**Figure 1. Data from the online extinction learning task demonstrated that participants learned to discriminate between conditioned and unconditioned stimuli, and to decrease loss expectancy ratings for conditioned stimuli following the transition to extinction.** **a** Depiction of trials from the online extinction learning task. Participants were told that they were travelling through different zones of a spaceship, and needed to escape with enough space coins to power their journey home. Unfortunately, the coins needed to be carried by helper robots, some of whom were unreliable. On each trial, participants encountered a robot and rated how likely they think that robot would be to lose one of their coins using a sliding bar (participants were informed that their ratings would not change the outcome they observed, but that their predictions should be as accurate as possible in order to aid future space travellers).  $P(\text{lose a coin}|\text{CS}^+)$  was  $1/3$  during initial conditioning, and reduced to 0 during extinction training stages, ( $P(\text{lose a coin}|\text{CS}^-)$  was always 0). The transition between conditioning and initial extinction learning stages was unsignalled, but the final stage of the task (further extinction training in a novel context B) occurred following transition to a different ‘zone’ of the ship (signalled by a change of background image). **b** Mean loss expectancy ratings across participants, by CS type and task stage. **c** Median response times (RTs) to input ratings, by CS type and task stage. Error bars represent standard error of the mean. CS+, aversively conditioned (loss-associated) stimulus; CS-, non loss-associated stimulus.



## Latent cause modelling of extinction task data

Loss expectancy ratings data from the initial two stages of the task were analysed using the computational model of latent cause inference previously published by Gershman and colleagues (Gershman & Niv, 2012; Gershman & Hartley, 2015). In brief, the model posits that during learning, an individual attempts to infer which underlying (latent) cause is responsible for their observations, based on a combination of their previous experience and prior beliefs about causal structure of the environment (see Figure 2; Methods). Prior beliefs about the causal complexity of the environment are governed by a single concentration parameter,  $\alpha$ . Here, the key model output submitted to further analysis was the likelihood for each participant of a model where  $\alpha$  was allowed to be  $> 0$  (favouring multiple causes), compared a model where  $\alpha=0$  (single cause responsible for all observations) – quantified as a log Bayes Factor (logBF; higher values=greater likelihood of a multi-cause model).

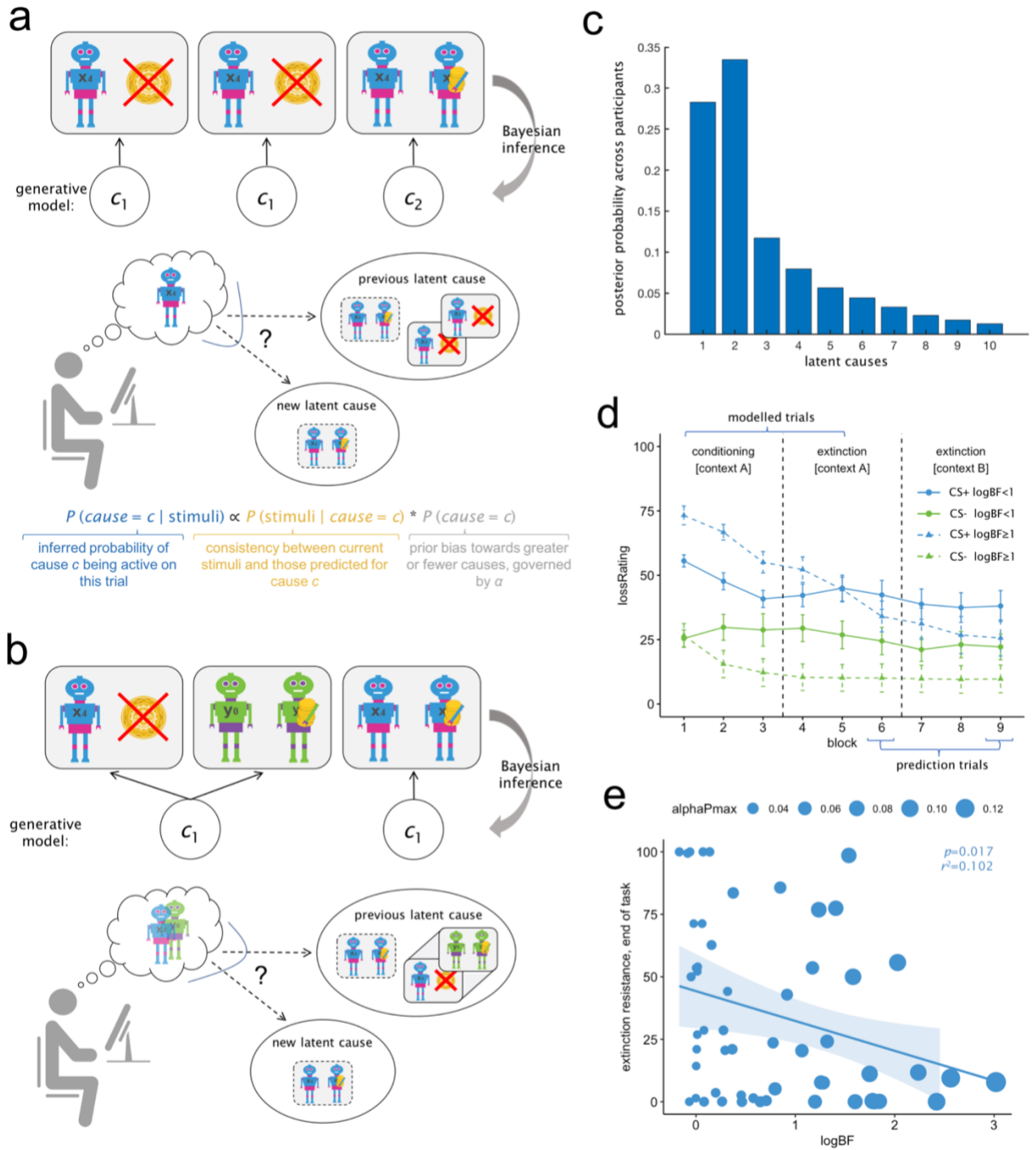
**Model fit.** In order to assess how well the model accounted for our data, observed loss expectancy ratings were plotted against model-predicted output for each trial, based on the posterior probability distribution of  $\alpha$  values for each participant (Figure S1). Across all participants, the mean correlation between actual and predicted loss expectancy ratings was 0.459 (SD 0.25). Permutation difference testing revealed that the mean  $r$  value for actual *vs* predicted loss ratings in our sample was significantly greater than that generated by fitting the latent cause model to randomly shuffled data (mean  $r$  for shuffled data=0.118; difference=0.312,  $p<0.001$ ; Figure S2). Goodness-of-fit ( $r$ ) values did not differ between PTSD and resilient individuals (PTSD group, mean  $r=0.477$ ; resilient group, mean  $r=0.405$ ;  $p>0.4$ , Welch’s two-sample  $t$  test), and were not related to logBF values ( $p>0.8$ , Spearman’s rank correlation test).

**Looking inside the model.** Inspection of the posterior distribution over latent causes for all subjects indicated that participants mainly assigned observations to one or two latent causes (Figure 2c; across participants, marginal probability of a third cause was 0.117). Similar to the pattern observed in Gershman and Hartley (2015), individuals whose behaviour favoured a single cause account appeared to learn more slowly across both acquisition and extinction stages (shallower curves for participants with  $\log\text{BF}<1$ , Figure 2d). Formal comparison by fitting a simple linear slope to individual CS+ ratings over the course of the modelled period revealed significantly shallower gradients in the lower logBF group ( $\log\text{BF}<1$ , mean gradient=-2.67 [SD 8.5];  $\log\text{BF}\geq 1$  mean gradient=-7.21 [SD 7.1];  $p=0.034$ , Wilcoxon signed-rank test). Notably, the single cause group also appeared to discriminate less between conditioned (loss-associated) and unconditioned (non loss-associated) stimuli. Over the course of the modelled period, individuals with lower logBF values distinguished less between CS+ and CS- stimuli in terms of their loss expectancy ratings ( $\log\text{BF}<1$ , mean difference in rating=18.2 [SD 24.7];  $\log\text{BF}\geq 1$  mean difference in rating=43.4 [SD 27.0];  $p<0.001$ , Wilcoxon signed-rank test). This difference was driven by both lower loss expectancy ratings for loss-conditioned (CS+) stimuli, and higher expectancy ratings for non loss-associated (CS-) stimuli, in the low logBF group ( $\log\text{BF}<1$ : mean CS+ rating= $46.2 \pm 14.3$ , mean CS- rating= $28.0 \pm 23.6$ ;  $\log\text{BF}\geq 1$ : mean CS+ rating= $58.3 \pm 13.0$ , mean CS- rating= $15.0 \pm 21.5$ ;  $p=0.002$ ,  $p=0.035$ , respectively; Wilcoxon signed-rank tests). This suggests that the slower extinction learning in individuals with greater likelihood of a single cause model (lower logBF values) might be a result

of more similar observation representations across different trial types (CS+ [reinforced], CS+ [unreinforced], and CS- trials) in these individuals (Figure 2b).

**Relationship to extinction resistance and safety learning.** LogBF values were not related to extinction resistance (mean residual CS+ loss expectancy rating) at the end of the modelled period, or end of the initial extinction learning stage (block 5,  $\beta=-4.2$ ,  $p=0.267$ ; block 6,  $\beta=-6.7$ ,  $p=0.132$ ; linear regressions weighted by certainty in posterior  $\alpha$  estimate), but were significantly related to extinction resistance at the end of the task (block 9,  $\beta=-11.9$ ,  $p=0.017$ ; Figure 2e, Figure S3a). This suggests that latent cause inference during initial learning might predict future resistance to extinction training – with higher likelihood of inferring a single cause during initial learning associated with persistence of loss expectancy for CS+ stimuli many trials into extinction. Interestingly, logBF values were also significantly negatively associated with CS- ratings at these three task stages ( $\beta=-11.5$ ,  $-9.9$ ,  $-9.2$  for block 5, block 6, block 9, respectively; all  $p<0.02$ ), although these associations appear significantly non-linear (Figure S3b). This suggests that latent cause inference may also relate to either heightened generalization of aversive consequences from CS+ to CS- stimuli, or failure of discriminative safety learning for CS- stimuli, over the course of the task.

An alternative explanation is that these relationships are due to a common non-specific effect, such as poorer working memory function, lower attentional performance, or more perseverative response style in individuals with lower logBF estimates. In order to test this hypothesis, we used data from the Cogstate neurocognitive test battery that were available for a subset of participants. Although this test is likely underpowered (only  $N=24$  individuals had Cogstate data), we found no evidence of a relationship between Cogstate composite scores (derived from multiple tasks indexing attentional and working memory function), and logBF estimates ( $\beta=0.010$ ,  $p=0.338$ ). In order to further test whether lower logBF scores might reflect tendency towards a stimulus-independent or inattentive response style, we also examined whether median response times during each stage of the task were related to logBF estimates. Interestingly, there was marginal evidence of a *negative* relationship between logBF and median response times during conditioning and extinction learning stages ( $\beta=-552$ ,  $p=0.054$ ;  $\beta=-469$ ,  $p=0.055$ ) – with longer median response times associated with lower logBF scores (extinction recall:  $\beta=-341$ ,  $p=0.214$ ; Figure S3c). This may indicate greater uncertainty about stimulus-outcome associations in lower BF individuals during initial learning and extinction training.



**Figure 2. Latent cause modelling of extinction task data revealed that trauma-exposed individuals whose extinction task data were better explained by a single cause model discriminated less between conditioned and unconditioned stimuli during initial learning, and showed greater resistance to extinction.** **a** The model posits that during learning an individual attempts to infer which latent cause is responsible for their observations, based on their previous experience of the task and prior beliefs about causal structure of the environment. On each trial, individuals may infer that a previous cause is responsible for their observations, or that something in the underlying task structure has changed, and observations should be assigned to a new cause. The probability of assigning an observation to a new cause is proportional to the dissimilarity between the current stimuli and those predicted for the current cause, and individual preference towards simpler or more complex causal structures (governed by a single parameter,  $\alpha$ ). According to one account, individuals with fear-learning disorders may be more likely to assign extinction trial observations to a new underlying cause, rendering them susceptible to extinction

relapse (spontaneous return of fear) if, for some reason, they infer that the original cause is active again (e.g., when times passes or contextual cues change). **b** Under an alternative account, individuals with fear-learning disorders may have a fundamental deficit in distinguishing trials involving aversively conditioned (CS+) and unconditioned (CS-) stimuli (e.g., due to over-generalization of aversive information, or hampering of safety learning by hyperarousal). Disparate stimuli and outcomes (CS+ and CS- trials) may be clustered together in the inferred causal structure of the environment, leading to greater uncertainty about the relationships between stimuli and outcomes, and therefore slower learning of stimulus values during both initial conditioning and later extinction stages. **c** The marginal probability distribution of latent causes averaged across all participants indicated that most participants inferred that one or two causes were responsible for their observations across conditioning and extinction stages (other causes had relatively low posterior probabilities). **d** The likelihood that an individual's internal model of the task contained more than one cause can be quantified as the log Bayes' Factor (logBF) for a model where  $\alpha > 0$ , compared to single cause model (where  $\alpha = 0$ ). For illustration, behavioural data is displayed separately for individuals for whom model comparison favoured a model with more than one cause (logBF  $\geq 1$ , dotted lines,  $N=20$ ), and individuals for whom model comparison found no strong evidence for a multi-cause model (logBF  $< 1$ , solid lines,  $N=36$ ). The latter group tended to learn more slowly across the task (flatter curves) and showed less discrimination in loss expectancy ratings between CS+ and CS- stimuli. Error bars represent standard error of the mean. **e** Lower logBF values (calculated from conditioning and extinction stage data only) were associated with higher resistance to extinction scores (residual CS+ loss expectancy ratings) at the end of the task. Regression line and  $p$  value represents linear model fit, weighted by posterior certainty in  $\alpha$  parameter estimates (alphaPmax; higher certainty=larger dot size). Panels *a* and *b* are adapted from a figure in Gershman et al., (2015).

## Relationship between latent cause inference and PTSD symptoms

We next sought to examine whether our index of latent cause inference was related to experience of PTSD symptoms, across our sample of trauma-exposed individuals.

**Relationship with avoidance symptoms.** In individual linear regression models weighted by posterior certainty in alpha parameter estimates, logBF values were significantly negatively related to PCL-5 avoidance symptoms ( $\beta=-0.89$ ,  $p=0.045$ ), and non-significantly related to PCL-5 total symptom severity score ( $\beta=-5.6$ ,  $p=0.069$ ), Figure 3a. Specifically, individuals with lower logBF values, indicating greater likelihood of a single cause model across conditioning and extinction learning, reported greater levels of avoidance symptoms. In order to test for evidence of a non-specific relationship between psychological symptom levels and parameter estimates, logBF values were also compared to BDI-II total depression symptom scores ( $\beta=-1.9$ ,  $p=0.234$ ).

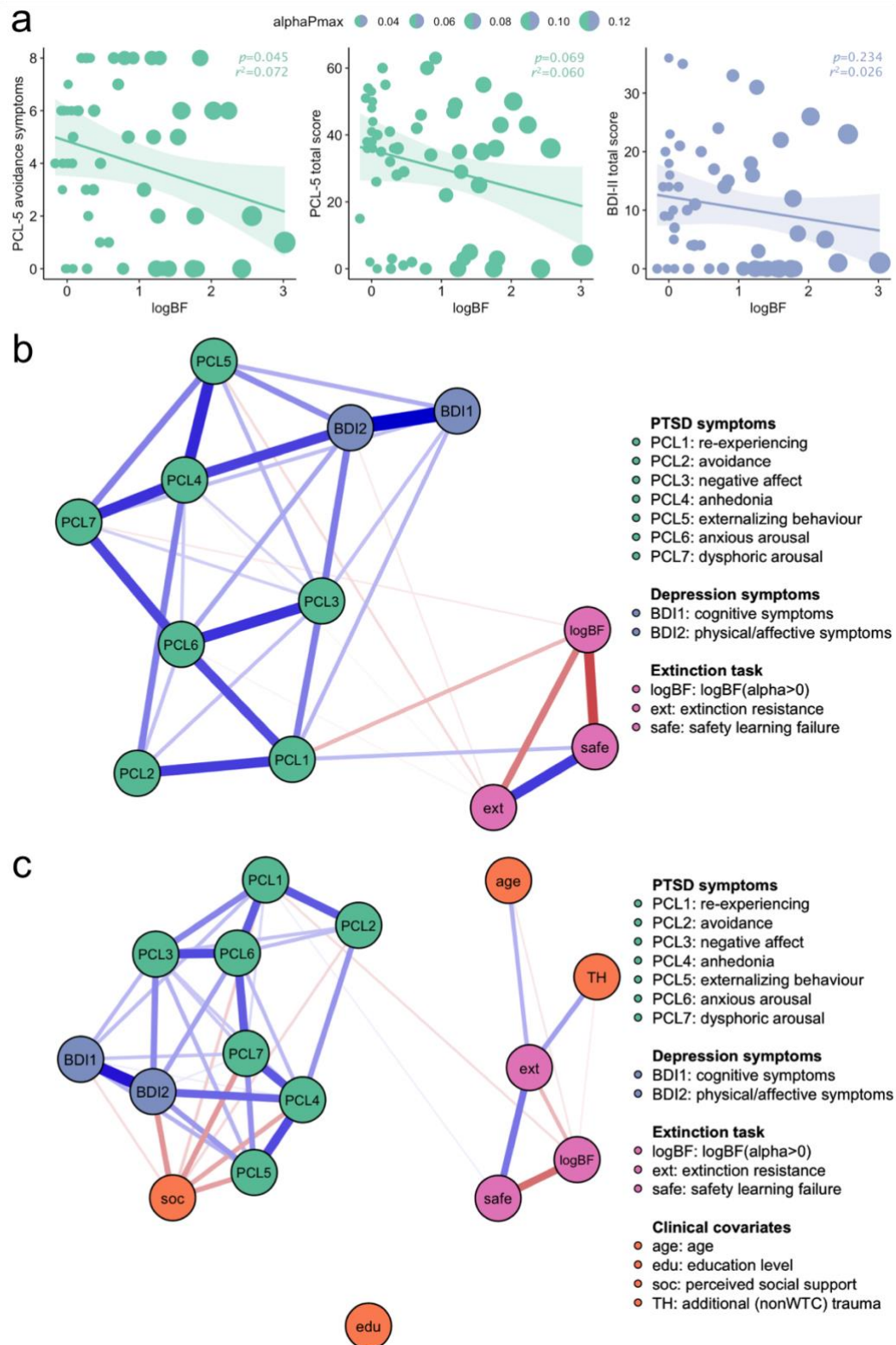
## Network analysis of extinction task parameters, PTSD, and depression symptoms.

Multivariate relationships between latent cause inference, task performance variables, and all measured PTSD and depression symptom dimensions were next explored using network analysis. Briefly, relationships between variables were represented as undirected Gaussian graphical models. Each node in the network represents a variable, and connections between nodes (edges) represent statistical relationships between variables after accounting for values of all other variables in the network (i.e., partial correlation coefficients; Epskamp & Fried, 2018). In order to generate a sparse and interpretable network, regularization was applied to the underlying partial correlation matrix during network estimation. This is a conservative approach designed to remove spurious connections, such that any retained edges can be thought of as contributing meaningfully to the overall variance (according to simulation studies, edges discovered by this method are likely to represent edges in the true network, but some true edges may be missing; Epskamp, 2018; Epskamp & Fried, 2018).

Whilst accounting for individual differences in extinction resistance and safety learning failure at the end of the task, greater severity of PTSD re-experiencing symptoms was associated with lower logBF values (regularized edge weight  $-0.089$ , bootstrapped 95% CI for edge value  $=-0.224-0$ , Figure 3b). As re-experiencing symptoms were positively connected to avoidance symptoms (regularized edge weight  $0.262$ , bootstrapped 95% CI  $= 0.053-0.414$ ), this suggests that the relationship between logBF and avoidance behaviour may be mediated by more intense re-experiencing symptoms (intrusive thoughts, nightmares, flashbacks, and emotional and physiological reactivity to trauma-related cues). There was also a negative connection between logBF and dysphoric arousal PTSD symptoms (difficulty concentrating and sleep disturbance; regularized edge weight  $-0.026$ , bootstrapped 95% CI  $= -0.153-0$ ). Sample weights and bootstrapped 95% CIs for all network edges are displayed in Figure S4. Simulation-based power analysis revealed acceptable network recovery properties at  $N=56$ . Across 1000 simulations, the median correlation between true and recovered networks at this sample size was  $0.765$  (IQR  $0.13$ ). Median sensitivity (accurate discovery of present edges) was  $0.694$  (IQR  $0.14$ ), and specificity (accurate discovery of absent edges) was  $0.767$  (IQR  $0.17$ ).

**Network analysis incorporating other clinical and demographic covariates.** When additional covariates (age, education level, cumulative trauma history, and perceived level of social support) were added to the network, the negative connection between logBF values and re-experiencing symptoms was retained (regularized edge weight -0.041, bootstrapped 95% CI for edge value -0.175–0, Figure 3c). There were also negative connections between age and lifetime trauma history and logBF scores. Specifically, individuals who were older, and participants who reported greater cumulative lifetime trauma tended to have lower logBF estimates (regularized edge weights -0.029, -0.020; bootstrapped 95% CIs -0.272–0, -0.234–0; respectively). Sample weights and bootstrapped 95% CIs for all network edges are displayed in Figure S5.

In order to test whether the inclusion of additional covariates significantly influenced relationships between psychological symptoms and contextual extinction task behaviour, we removed the clinical covariate nodes from the adjacency matrix for the fuller network (Figure 3b), and compared the remaining values to those estimated from the reduced network (Figure 3c; Methods). The sum of edges between PTSD/depression symptoms and extinction task variables was reduced from 5.02 to 4.16 by inclusion of the four additional covariates – in other words, controlling for these additional variables reduced the inter-connectivity of these variables by 17%. The overall structure of the network was robust to inclusion of the additional covariates, as correlation between the edge weights derived from covariate controlled and non-covariate controlled networks was high (Spearman's  $\rho=0.955$ ). The biggest change between covariate-controlled and non-covariate controlled networks was a reduction in connection strength between extinction learning and safety failure estimates (decrease in edge strength from 0.264 to 0.178; mean change in edge strength between networks was 0.016). This appeared to be mediated via by inclusion of the age and lifetime trauma history covariates, which were both positively related to extinction resistance (higher extinction resistance at the end of the task in older participants, and those with a greater cumulative trauma load; Figure 3c). However, simulation-based power analysis revealed that the structure of the full covariate-controlled network reported here should be interpreted with caution, as it is likely underpowered: with an  $N$  of 150 or more required for satisfactory sensitivity and specificity in true network recovery.



**Figure 3. Trauma-exposed individuals with greater tendency to infer a single cause was responsible for observations across conditioning and extinction stages reported more severe PTSD, but not depression, symptoms.** **a** Bivariate relationships between latent cause inference and avoidance, total PTSD, and total depression symptoms. logBF represents log Bayes' Factor for a model with more than one cause ( $\alpha > 0$ ), compared to a single cause model ( $\alpha = 0$ ). PCL-5, PTSD Checklist for DSM-5; BDI-II, Beck Depression Inventory, version II. *p* values

represent the results of linear regression models, weighted by posterior certainty in the value of  $\alpha$  (alphaPmax; higher certainty=larger dot size). **b** Regularized network model incorporating clinical symptom dimensions (seven PCL-5 PTSD and two BDI-II depression symptom dimensions) and extinction task performance measures (extinction resistance, or mean residual CS+ loss expectancy at the end of the task; safety learning failure, or mean CS- loss expectancy at the end of the task; and logBF, indexing latent cause inference across initial conditioning and extinction learning trials). **b** Regularized network model incorporating PTSD and depression symptoms, extinction task performance measures, and clinically-relevant covariates: specifically age, self-reported education level, perceived social support (MOS-SS total score), and additional lifetime trauma history. Connections between nodes (edges) indicate those retained following LASSO regularization, a conservative approach which favours a sparse network structure and removes spurious edges. Blue edges represent positive connections, and red edges negative connections. Greater line width and stronger colour intensity represent greater edge strength, with edge weights plotted using the same scale in order to be comparable across networks.



## Discussion

Here, we provide the first evidence that individual differences in latent cause inference, as measured during a simple behavioural extinction learning paradigm, may be related to experience of psychological symptoms following trauma. Specifically, we found that trauma-exposed individuals whose patterns of behavioural responses were associated with greater likelihood of a generative model with a single underlying cause exhibited greater resistance to extinction training in future trials, poorer safety learning, and higher levels of avoidance symptoms. Network analysis incorporating multiple PTSD and depression symptom clusters indicated that the association between latent cause inference and avoidance was likely mediated via greater severity of PTSD re-experiencing symptoms (intrusive thoughts, nightmares, flashbacks, and emotional/physiological reactivity to reminders). Importantly, this analysis also controlled for individual differences in task performance (extinction resistance and safety learning failure), indicating that our model-based index had additional explanatory power over raw behavioural scores with respect to prototypical post-traumatic symptoms.

Strengths of the data presented here include a clinically well-characterised sample: all participants completed an in-depth clinical interview, as well as providing self-reported measures of current symptom levels, additional lifetime trauma exposure, and other relevant sociodemographic information. All participants also had exposure to the same primary (index) trauma (the World Trade Center disaster in 2001). Participants reported a range of current PTSD symptom levels, from no/minimal symptoms (resilient) to severe cases (mean PCL-5 total score  $31 \pm 19$ ) – however it should be noted that due to the length of time passed since the index trauma, this represented a chronic disease course for all symptomatic individuals. Although  $N=56$  is a relatively modest sample size for a behavioural study, simulation-based power analysis for the network model revealed satisfactory sensitivity and specificity estimates for a discovery analysis.

Considering model parsimony (only two free parameters) and the continuous nature of the response variable, the latent cause model provided a generally good account of participants' loss expectancy data – however the extent to which this was true varied across subjects (Figure S1). Bivariate associations between model output and behavioural and clinical measures were therefore weighted by how informed estimates of the model parameter governing causal clustering were by task data (i.e., peakiness of the posterior probability density function for alpha values). In weighted models, likelihood of a multi-cause model ( $\log\text{BF}[\alpha > 0]$ ) was negatively associated with extinction resistance (residual CS+ loss expectancy ratings) at the end of the task (Figure 2). LogBF values were also strikingly negatively related to failure of safety learning (loss expectancy ratings for CS- stimuli, in the absence of any association with the loss outcome) at all stages of the task (Figure S3). Individuals with lower logBF values tended to show slower response times across conditioning and initial extinction task stages – which may indicate greater uncertainty about stimulus values during initial learning.

We interpret these results as suggesting that failures of extinction learning in PTSD may relate to a primary deficit in extinction memory formation (associated with a tendency towards causal 'overgeneralization'), rather than failure to retrieve a successfully formed extinction memory (associated with a tendency towards causal hyper-segmentation). Specifically, in individuals with

trauma-related psychopathology, reduced discrimination between CS+ and CS- during initial learning may result in a tendency to classify all observations (CS+, CS-, US and US omission) as being produced by the same underlying cause (Figure 2b). This results in uncertainty about specific stimulus-outcome associations, and therefore hampers learning of both initial CS-US associations (successful danger and safety learning during conditioning), and learning that these contingencies have changed (under extinction). This explanation is consistent with the behavioural pattern observed in individuals whose behaviour supported greater likelihood of a single cause model (Figure 2d).

An alternative explanation for our findings is that logBF values and symptom measures may both related to some other relevant individual difference, such as poorer working memory – which might predict less discriminative task performance –, or a more habitual or perseverative response style – which might predict continuing to enter high loss expectancy values under extinction. Although we did not find any evidence that logBF values were associated with performance scores on a battery of neurocognitive tests probing general executive function, this data was only available in a subset of individuals ( $N=24$ ), and should be considered in the context of evidence of executive dysfunction in PTSD (Scott et al., 2015). Further, although ratings can be considered a relatively ‘pure’ measure of values or beliefs, they typically exhibit more exaggerated response functions than implicit measures (such as physiological recordings) during experimental tests of fear-conditioning (Holt et al., 2014), and may be more susceptible to certain forms of response bias. For example, it is possible that the loss expectancy data collected here are sensitive to demand characteristics (participants entering responses they believe are desired by the experimenter), and that perception of these characteristics may differ between patient and healthy samples (Orne, 1962). Future work should therefore include both attentional checks (catch trials) during task performance, and explicit questions about stimulus value, and beliefs about task structure and purpose.

It is also important to stress that, in order to facilitate remote administration, the ‘aversive’ outcome used in this task is highly unlikely to evoke ‘fear’ in the same way as stimuli used in previous work (e.g., painful electric shock). Although previous experimental tasks have successfully used monetary or game points loss in place of more primary aversive outcomes to discover differences in learning related to self-reported anxiety (e.g., Norbury et al., 2018; Wise & Dolan, 2020), more evidence is needed that the outcome employed here is engaging the kind of cognitive processes relevant to learning and extinction related to traumatic experience. Future studies using this framework will therefore attempt to increase emotional engagement with the task, by using more immersive graphics and overall taking a more gamified approach to task presentation and will also explicitly probe aversiveness of the loss outcome to study participants (see Nord et al., 2017; Wise & Dolan, 2020, for a successful examples of this strategy).

Finally, effect sizes reported here are modest – with our measure of latent cause inference explaining around 7-8% of variance in self-reported PTSD symptoms. However, these associations persisted under a conservative (regularized) analysis approach, which tends to shrink connections weights (Epskamp et al., 2018), and after controlling for multiple other psychological symptom dimensions and clinically-relevant covariates. It was striking that logBF values were also associated with cumulative trauma history (independently from age), as previous work suggests that lifetime trauma

load is an important predictor of vulnerability to post-traumatic stress (Karam et al., 2014; Feder et al., 2016). It will be necessary to test if these relationships persist in future replication samples, and if sensitivity can be increased by the various improvements to task design discussed above. A further important step will be to undertake longitudinal assessments, in order to investigate both reliability of model-based causal inference metrics, and directionality of relationships with evolving symptom dynamics. If these challenges can be overcome, this may further our understanding of the role of high-level dysfunctional beliefs in the development and maintenance of post-traumatic stress, and perhaps even give insight into how such beliefs might be better targeted by psychological therapies (Moutoussis et al., 2017).

## **Acknowledgments**

This work was supported by CDC-NIOSH U01 awards to AF (grant numbers OH011473 and OH010729), and a NARSAD Young Investigator award from the Brain and Behavior Research Foundation to AN (grant number 28604).

## Methods and Materials

### Participants

Participants were World Trade Center disaster survivors and rescue/recovery workers, recruited from two ongoing studies at the Trauma and Resilience Program at the Icahn School of Medicine at Mount Sinai. All participants had DSM-5 Category A trauma exposure (defined as “actual or threatened death or serious injury”, American Psychiatric Association, 2013) during the WTC disaster, as determined by clinical interview. Participants from these studies fell into two broad categories: individuals who currently met diagnostic criteria for full or subthreshold PTSD, and individuals who were assessed as never having met criteria for PTSD or any other DSM-5 Axis-1 disorder. As per previous studies from our research group (e.g., Chen et al., 2020), subthreshold PTSD was defined according to two commonly-used sets of criteria, adapted for DSM-5 (see McLaughlin et al., 2015). Specifically, this required the presence of one Criterion B (intrusion) symptom, plus 1) either three symptoms from across Criteria C (avoidance) and D (negative alterations in mood or cognition), or two symptoms from Criterion E (alterations in arousal and physiological reactivity); or 2) one criterion C or D symptom plus one Criterion E symptom.

For both studies, participants with a lifetime history of a primary psychotic or bipolar disorder, alcohol or other substance use disorder within the prior three months, current uncontrolled medical illness, disorder of the central nervous system, or history of head injury were not recruited.

Participants were further excluded if they were currently taking antipsychotic, mood stabilising, or opioid medications, or reported day-time use of sedative medications. Participants taking other psychoactive medications (e.g. SSRIs) were required to be on a stable dose (same medication and dose for at least three months, prior to taking part). Both studies received ethical approval from the Institutional Review Board at the Icahn School of Medicine at Mount Sinai and all participants provided informed, written consent. All data analysed here were collected between 06/03/19 and 23/02/20, i.e., prior to the onset of the Covid-19 pandemic in the USA.

### Demographic and clinical measures

**PTSD and depression symptoms.** All participants completed an in-depth clinical interview with an experienced Trauma and Resilience Program team member. For  $N=25$  participants this consisted of the Structure Clinical Interview for DSM-5 (SCID-5) and Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) (Williams et al., 2015; Weathers, Blake, et al., 2013), and for  $N=31$  participants this was the Mini-International Neuropsychiatric Interview for DSM-5 (MINI-5) (Sheehan et al., 1998). Additionally, all participants completed the PCL-5 and BDI-II self-report measures of PTSD and depression symptoms (Weathers, Litz, et al., 2013; Beck et al., 1996). In order to reduce number of measurements relative to number of observations (participants), individual PTSD symptoms were parsed into seven dimensions (re-experiencing, avoidance, negative affect, externalizing behaviour, anxious arousal, and dysphoric arousal symptoms clusters), which have previously been demonstrated to provide the best fit to PTSD symptoms data across various independent samples of trauma-exposed individuals (Armour et al., 2015, 2016). Depression symptoms as measured on the BDI-II were divided into ‘cognitive’ and ‘physical/affective’ subdimensions on the basis of results of

a longitudinal dynamic network analysis of diverse samples of depressed individuals by Bringmann and colleagues (Bringmann et al., 2015).

**Trauma history.**  $N=25$  participants provided information about additional lifetime trauma exposure using the Traumatic Life Events Questionnaire (TLEQ; Kubany et al., 2000), and  $N=31$  using the Trauma History Screen (THS; Carlson et al., 2011). Both questionnaires assess exposure to common types of intense stressors across the lifetime. Items probing common trauma categories across both measures were used to derive a non frequency-weighted trauma history score for the presence/absence of 13 common traumatic experiences (e.g. life-threatening illness or injury, natural disaster, sexual trauma, sudden death of a close friend or loved one – see Table S2 for full details). Stressful events probed only by one of the two scales (e.g. sudden abandonment by partner, miscarriage) were not included. Although items are not completely interchangeable, and so total trauma scores used here should be interpreted with caution, THS and TLEQ scores are moderately strongly correlated ( $r=0.73-77$ ), and median agreement between measures by trauma category was previously found to be 72-78% in community samples of adults (Carlson et al., 2011). Total score on this measure was used for further analysis on the basis of evidence that the burden of trauma is cumulative across lifespan and this predicts likelihood of PTSD from WTC trauma (Karam et al., 2014; Feder et al., 2016).

**Demographic information, social support, and cognitive function.** Participants completed a demographic questionnaire where they provided information about their age, gender, race/ethnicity, and highest level of educational achievement. All participants also completed the Medical Outcomes Survey brief Social Support questionnaire (MOS-SS; Sherbourne & Stewart, 1991) – a measure of perceived social support – which was included in the analysis as social support has previously been shown to play a vital role in resilience to psychopathology following trauma (Fritz et al., 2018; Stevens & Jovanovic, 2019). A subset of individuals ( $N=24$ ) completed the Cogstate battery, a set of computerised tests probing general executive function that have been shown to be sensitive to mild cognitive impairment (Maruff et al., 2009). Specifically, participants completed the identification test (attention), detection test (psychomotor function), Groton maze learning tasks (executive function and memory), international shopping list test (verbal learning and memory), one card learning test (visual learning), and the one-back task (working memory). Primary outcome measures for each test were combined to form composite cognitive function scores ( $\bar{x}$  scored across participants), as described in the Cogstate manual.

### **Extinction learning task**

In order to test feasibility for future remote work, the extinction learning task was administered online. On each trial, participants were presented with an image of a robot and asked to rate their expectancy of that particular losing a coin using a sliding bar with the anchors “keep safe” [0] and “lose coin” [100]. After participants input their prediction, the outcome of that trial was displayed (a simple animation representing either loss of a coin, the aversive outcome, or retention of coins, the neutral outcome; Figure 1a). Participants were informed at the start of the task that their ratings would have no impact on the outcomes they observed, but should be made as accurately as possible, in order to aid future space travellers. Following instructions screens outlining the cover story and images of example trials, a short true/false quiz tested understanding of the task contingencies.

Participants scoring less than 3/3 correct were routed back to the start of the instructions screens to try again.

Each stage of the experiment (conditioning [in context A], initial extinction training [in context A], and further extinction training [in novel context B]) consisted of 3 blocks of 10 trials (90 trials total). CS+ stimuli were reinforced on 1/3 of trials in conditioning phase only: the ratio of trial types was 4:2:4 CS+(unreinforced):CS+(reinforced):CS- during conditioning, and 4:0:4 during both extinction stages. Responses were not speeded (trials did not progress until a rating had been entered), and participants had the option to take a break for as long as they liked at the end of each block of 10 trials. The identities of CS+ and CS- stimuli (defined by a combination of robot colours and markings) were counterbalanced across participants and all participants experienced the same pseudorandom trial sequence (and therefore identical outcomes).

The original purpose of including novel context B in the task design was to test for the possibility of detecting renewal effects related to contextual cue changes. However, in the analysis presented here we do not further investigate renewal or recall effects as 1) we observed differential responding to CS+ and CS- stimuli at the end of the initial learning stage, indicating incomplete learning, and 2) the lack of temporal delay between the two extinction training phases means we are unlikely to be tapping into cognitive mechanisms underlying recall (Orederu & Schiller, 2018).

Stimuli were generated using royalty-free vector art from the pixabay database (<https://pixabay.com/>), and were checked for distinctiveness under common forms of colour blindness using the Color Oracle simulator (<https://colororacle.org/>). The task was programmed in javascript using tools from the jsPsych library (de Leeuw, 2015), and deployed using Pavlovia (<https://pavlovia.org/>). Task code and stimuli are available at the github repository for this project <https://github.com/agnesnorbury/latent-cause-PTSD>.

## Analysis

All statistical analyses (except latent cause modelling) were carried out in R version 3.6.1 (R Core Team, 2019). Analysis code and full version information for R packages is available at <https://github.com/agnesnorbury/latent-cause-PTSD>.

**Contextual extinction task data.** Effects of within-task manipulations (effects CS type and task stage on ratings and response input times) were explored using repeated-measures ANOVA of linear mixed-effect models with error terms included as random effects (lmer function from the R package lmerTest). Within-subjects factors in the model were CS type (CS+ *vs* CS-) and task stage (conditioning, initial extinction learning in original context A, and further extinction learning in novel context B). Follow-up contrasts to explore the direction of effects used the Tukey LSD correction for multiple comparisons.

Extinction resistance was defined as mean CS+ loss expectancy rating under extinction, measured at the end of both the initial extinction [context A] and further extinction learning [context B] task stages). Absolute values were used for CS+ ratings (as opposed to difference in values between CS+ and CS- stimuli), as – in contrast to other types of data such as GSR or BOLD signals – expectancy ratings have an absolute meaning. Further, experimental evidence suggests that individuals with PTSD may over-generalize negative information from conditioned to unconditioned stimuli (Lissek

& van Meurs, 2015; Kaczkurkin et al., 2016) – which might result in inappropriately low difference-based values for these quantities (e.g. in the case where loss expectancy ratings are high for both CS+ and CS- stimuli).

**Latent cause modelling of extinction task data.** Latent cause modelling of loss expectancy ratings data was carried out using code associated with Gershman & Niv (2012) (available at <https://github.com/sjgershm/LCM>), run in MATLAB R2019a (MathWorks Inc., 2019).

Briefly, the model assumes that on each trial, participants compute the posterior probability that cause  $c$  generated the observed stimuli, using Bayes' rule (equation 1).

$$P(\text{cause} = c \mid \text{stimuli}) \propto P(\text{stimuli} \mid \text{cause} = c) * P(\text{cause} = c) \quad (1)$$

I.e., the inferred probability of cause  $c$  being active on a given trial, given observation of the trial stimuli (compound observation consisting of conditioned stimuli and outcomes, see Figure 2a), is proportional to the likelihood of the current cause (consistency between current stimuli and predicted stimuli for cause  $c$ ), multiplied by a prior term that indexes an individual's preference for simpler or more complex causal structures. This prior biases the model to assign trials to a given cause in proportion to the number of trials previously assigned to that cause, and to a new cause with a probability proportional to the value of the free parameter alpha (i.e., the distribution over states is modelled using a Chinese Restaurant Process with concentration parameter  $\alpha$  – see Gershman & Niv, 2012; Gershman et al., 2015). Smaller values of alpha bias individuals towards simpler clusterings, observations tend to be assigned to the same cause, and larger values towards more complex clusterings, where observations are assigned to different causes (when  $\alpha=0$ , all observations are assigned to a single cause, and when  $\alpha \rightarrow \infty$ , every observation is assigned to a unique cause). As the learner has some uncertainty about the stimulus configuration associated with each state, the output on each trial is not a single cause, but a posterior probability distribution across potential underlying causes.

As per Gershman and Hartley (2015) (where the modelled response variable was GSR amplitude on each trial), we used a linear observation function with single scaling parameter beta to fit model-generated  $P(\text{loss})$  on each trial to observed loss probability ratings (equation 2).

$$\text{lossRating} = \beta * \sum_c P(\text{loss} \mid \text{cause} = c) * P(\text{cause} = c \mid \text{stimuli}) + \epsilon \quad (2)$$

where  $\epsilon$  is a noise parameter drawn from a normal distribution with mean=0 and variance=1.

Values of parameters  $\alpha$  and  $\beta$  that maximized the likelihood of an individuals' ratings data were found using the non-linear optimization method described in (Gershman & Niv, 2012; Gershman & Hartley, 2015). Briefly, this is a sampling approach based on summing over sets of hypothetical state sequences or particles (particle filtering). Here, we used  $M=500$  particles. Values of parameters were estimated for each participant separately (i.e., non-hierarchically). The maximum number of latent causes an individual could infer over was 10, and  $\alpha$  values were allowed to range [0,10] ( $N=50$  linearly spaced values were evaluated). Loss expectancy ratings data were  $z$  scored within-participants prior to analysis.

Following Gershman and Hartley (2015), latent cause modelling was applied to conditioning and initial extinction training data only. The last block of initial extinction learning trials was held out, so

that model output would be unbiased by trials used to calculate extinction resistance at the end of this stage. The key output of the model submitted to further analysis was the likelihood of a model where  $\alpha$  was allowed to be  $>0$  (i.e., with multiple inferred causes), compare to a model where  $\alpha=0$  (single underlying cause). This likelihood was computed as a log Bayes Factor, and is referred to as logBF throughout the manuscript. A log Bayes Factor of  $\geq 1$  is generally interpreted as representing strong evidence in favour of the comparator hypothesis (here, in favour of a multi-cause model) (Kass & Raftery, 1995). In order to assess goodness of model fit, actual *vs* predicted ratings generated by the model on each trial were compared for each participant using Pearson correlations. Permutation difference testing with 10,000 random assignments was used to compare goodness of fit ratings derived from actual data and  $N=1000$  randomly shuffled dummy datasets (Figure S2).

**Bivariate relationships between latent cause inference and behavioural/clinical data.** In order to account for individual differences in uncertainty about posterior estimates of the key internal model parameter ( $\alpha$ ), logBF estimates were bivariate related to behavioural and clinical measures using weighted least squares regression, with regression weights set equal to peak of the posterior probability distribution over alpha values. As probability distributions sum to 1, this peak value is inversely proportional to the width of spread of the probability density, which represents uncertainty about the posterior estimate (or how informative task data were in updating the uniform density prior). This ensures that greater weight in the analysis is allocated to estimates from participants for whom this quantity was more confidently derived (Figure S1b).

**Network analysis of latent cause inference, clinical, and sociodemographic data.** Network analysis was used to quantify inter-relations between individual differences in latent cause inference, as indexed using the online game, and individual differences in PTSD symptomatology at that point in time, taking into account other relevant clinical and demographic factors. Network estimation methodology followed previously published procedures for performing network analysis using regularized Gaussian graphical models (Epskamp & Fried, 2018; Fried & Burger, 2019).

Specifically, networks were constructed using regularized Gaussian graphical estimation implemented in the R package qgraph, version 1.6.5 (Epskamp et al., 2012). Under this approach, nodes represents observed variables, and connections between nodes (edges) represent unique pairwise associations (partial correlation coefficients) after conditioning on all other variables in the dataset. In order to deal with the relatively small number of observations (participants) compared to potential network connections, least absolute shrinkage and selection operator (LASSO) regularization (Tibshirani, 1996) was applied to the underlying partial correlation matrix. This form of regularization causes weak edge estimates to shrink to zero, resulting in a sparse, conservative network structure (Friedman et al., 2008). In order to obtain an optimal sparse estimate of the partial correlation matrix, the LASSO tuning parameter governing degree of regularization was selected using the extended Bayesian Information Criterion (Foygel & Drton, 2010). Clinical measures exhibited skewed distributions across the sample (see histograms in Table S1), and so were transformed using the nonparanormal function (huge.npn from the R package huge; Zhao et al., 2012) prior to network estimation. All networks reported here are based on npn-transformed Pearson correlations, and are unthresholded.



Two nested networks were estimated: one consisting of PTSD/depression symptoms scores and behavioural task variables, and one with PTSD/depression symptoms, behavioural task variables, and other clinically-relevant covariates (age, education level, perceived level of social support, and additional lifetime trauma history). Extinction resistance, calculated as mean CS+ loss expectancy rating at the task, and safety learning failure, calculated as mean loss expectancy for the CS- at the end of the task, were included in the networks as well as logBF in order to ascertain if the latent cause model parameter was more closely related to symptoms than these simple behavioural performance indices. Effects of including the additional clinically-relevant covariates in the symptom and extinction learning network were assessed using the approach taken in Armour et al. 2017, i.e., by deleting the relevant covariates from the adjacency matrix of the fuller network, then subtracting this modified (covariate-controlled) symptom network from the original network (without covariates).

Network robustness was assessed using the R package bootnet, version 1.4.3 (Epskamp et al., 2018). Specifically, edge weight accuracy was assessed using nonparametric bootstrapping (observations in the data were resampled with replacement to create 2000 plausible new datasets). These bootstrapped estimates can be used to estimate the sampling distribution of network edge weights, and therefore construct 95% confidence intervals (CIs), such that in 95% of the cases the CI will contain the true value of the parameter. *NB*, these CIs are not intended to be used to draw inferences about the significance of an edge in the network (distributions of LASSO-regularized parameters are substantially non-normal therefore approximate *p*-values are not easy to obtain via bootstrap sampling). The *presence* of an edge in a LASSO-regularized model implies that it is sufficiently strong to be included in the model. However, a wide CI means that it may be hard to interpret the strength of that edge (Epskamp et al., 2018). Due to our relatively small *N* (and likely instability of exact edge strength estimates), we do not report differences in edge strength between nodes, or node centrality indices. Power of the network analysis was assessed following the approach described by Faelens et al., using the netSimulator function from bootnet (Faelens et al., 2019). Specifically, we used the discovered (“true”) network structure to simulate 1000 networks at various sample sizes (*N*=56, 100, 150, 500), and then examined how well the simulated networks recovered the true network structure. This was assessed using three metrics: correlation between the simulated and true networks, sensitivity (rate of discovery of network connections present in the true network), and specificity (rate of discovery of network connections absent in the true network).

### **Data availability statement**

De-identified raw data for the contextual extinction task is available at <https://github.com/agnesnorbury/latent-cause-PTSD>. Clinical and demographic data are not freely publicly available due to lack of permission from study participants for public data sharing at the time of original consent.

## References

- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association.  
<https://doi.org/10.1176/appi.books.9780890425596>
- Armour, C., Contractor, A., Shea, T., Elhai, J. D., & Pietrzak, R. H. (2016). Factor Structure of the PTSD Checklist for DSM-5: Relationships Among Symptom Clusters, Anger, and Impulsivity. *The Journal of Nervous and Mental Disease*, 204(2), 108–115.  
<https://doi.org/10.1097/NMD.0000000000000430>
- Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, 45, 49–59. <https://doi.org/10.1016/j.janxdis.2016.11.008>
- Armour, C., Tsai, J., Durham, T. A., Charak, R., Biehn, T. L., Elhai, J. D., & Pietrzak, R. H. (2015). Dimensional structure of DSM-5 posttraumatic stress symptoms: Support for a hybrid Anhedonia and Externalizing Behaviors model. *Journal of Psychiatric Research*, 61, 106–113.  
<https://doi.org/10.1016/j.jpsychires.2014.10.012>
- Arnaudova, I., Kindt, M., Fanselow, M., & Beckers, T. (2017). Pathways towards the proliferation of avoidance in anxiety and implications for treatment. *Behaviour Research and Therapy*, 96, 3–13.  
<https://doi.org/10.1016/j.brat.2017.04.004>
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation.
- Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, 16(1), 5–13.  
<https://doi.org/10.1002/wps.20375>
- Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine*, 45(4), 747–757. <https://doi.org/10.1017/S0033291714001809>
- Carlson, E. B., Smith, S. R., Palmieri, P. A., Dalenber, C., Ruzek, J. I., Kimerling, R., Burling, T. A., & Spain, D. A. (2011). Development and Validation of a Brief Self-Report Measure of Trauma Exposure: The Trauma History Screen. *Psychological Assessment*, 23(2), 463–477.  
<https://doi.org/10.1037/a0022294>
- Chen, C., Salim, R., Rodriguez, J., Singh, R., Schechter, C., Dasaro, C. R., Todd, A. C., Crane, M., Moline, J. M., Udasin, I. G., Harrison, D. J., Luft, B. J., Southwick, S. M., Pietrzak, R. H., & Feder, A. (2020). The Burden of Subthreshold Posttraumatic Stress Disorder in World Trade

- Center Responders in the Second Decade After 9/11. *The Journal of Clinical Psychiatry*, 81(1).  
<https://doi.org/10.4088/JCP.19m12881>
- Contractor, A. A., Roley-Roberts, M. E., Lagdon, S., & Armour, C. (2017). Heterogeneity in patterns of DSM-5 posttraumatic stress disorder and depression symptoms: Latent profile analyses. *Journal of Affective Disorders*, 212, 17–24. <https://doi.org/10.1016/j.jad.2017.01.029>
- de Haan, A., Landolt, M. A., Fried, E. I., Kleinke, K., Alisic, E., Bryant, R., Salmon, K., Chen, S.-H., Liu, S.-T., Dalgleish, T., McKinnon, A., Alberici, A., Claxton, J., Diehle, J., Lindauer, R., Roos, C. de, Halligan, S. L., Hiller, R., Kristensen, C. H., ... Meiser-Stedman, R. (2020). Dysfunctional posttraumatic cognitions, posttraumatic stress and depression in children and adolescents exposed to trauma: A network analysis. *Journal of Child Psychology and Psychiatry*, 61(1), 77–87. <https://doi.org/10.1111/jcpp.13101>
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., & Baas, J. M. P. (2015). Updated Meta-Analysis of Classical Fear Conditioning in the Anxiety Disorders. *Depression and Anxiety*, 32(4), 239–253. <https://doi.org/10.1002/da.22353>
- Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking Extinction. *Neuron*, 88(1), 47–63. <https://doi.org/10.1016/j.neuron.2015.09.028>
- Epskamp, S. (2018). Preliminary simulations on the interpretation of cross-sectional Gaussian graphical models. *PsyArXiv*. <https://doi.org/10.31234/osf.io/54xrs>
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195–212.  
<https://doi.org/10.3758/s13428-017-0862-1>
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 048(i04). [https://econpapers.repec.org/article/jssjstsof/v\\_3a048\\_3ai04.htm](https://econpapers.repec.org/article/jssjstsof/v_3a048_3ai04.htm)
- Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, 23(4), 617–634. <https://doi.org/10.1037/met0000167>
- Faelens, L., Hoorelbeke, K., Fried, E., De Raedt, R., & Koster, E. H. W. (2019). Negative influences of Facebook use through the lens of network analysis. *Computers in Human Behavior*, 96, 13–22. <https://doi.org/10.1016/j.chb.2019.02.002>
- Feder, A., Mota, N., Salim, R., Rodriguez, J., Singh, R., Schaffer, J., Schechter, C. B., Cancelmo, L. M., Bromet, E. J., Katz, C. L., Reissman, D. B., Ozbay, F., Kotov, R., Crane, M., Harrison,

- D. J., Herbert, R., Levin, S. M., Luft, B. J., Moline, J. M., ... Pietrzak, R. H. (2016). Risk, coping and PTSD symptom trajectories in World Trade Center responders. *Journal of Psychiatric Research*, 82, 68–79. <https://doi.org/10.1016/j.jpsychires.2016.07.003>
- Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 604–612). Curran Associates, Inc. <http://papers.nips.cc/paper/4087-extended-bayesian-information-criteria-for-gaussian-graphical-models.pdf>
- Fried, E. I., & Burger, J. (2019). *2019-09 Workshop—Network analysis workshop (FLAMES, Ghent)*. <https://doi.org/10.17605/OSF.IO/P527F>
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441. <https://doi.org/10.1093/biostatistics/kxm045>
- Fritz, J., Fried, E. I., Goodyer, I. M., Wilkinson, P. O., & van Harmelen, A.-L. (2018). A Network Model of Resilience Factors for Adolescents with and without Exposure to Childhood Adversity. *Scientific Reports*, 8(1), 15774. <https://doi.org/10.1038/s41598-018-34130-2>
- Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior*, 43(3), 243–250. <https://doi.org/10.3758/s13420-015-0176-z>
- Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife*, 6, e23763. <https://doi.org/10.7554/eLife.23763>
- Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, 20(2), 251–256. <https://doi.org/10.1016/j.conb.2010.02.008>
- Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, 40(3), 255–268. <https://doi.org/10.3758/s13420-012-0080-8>
- Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5, 43–50. <https://doi.org/10.1016/j.cobeha.2015.07.007>
- Greene, T., Gelkopf, M., Epskamp, S., & Fried, E. (2018). Dynamic networks of PTSD symptoms during conflict. *Psychological Medicine*, 48(14), 2409–2417. <https://doi.org/10.1017/S0033291718000351>
- Holt, D. J., Boeke, E. A., Wolthuisen, R. P. F., Nasr, S., Milad, M. R., & Tootell, R. B. H. (2014). A parametric study of fear generalization to faces and non-face objects: Relationship to discrimination thresholds. *Frontiers in Human Neuroscience*, 8, 624. <https://doi.org/10.3389/fnhum.2014.00624>

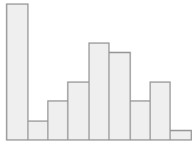
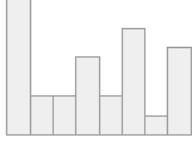
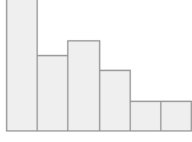
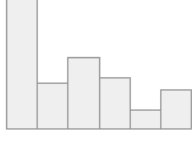
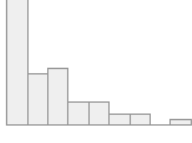
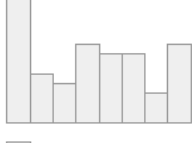
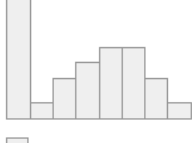
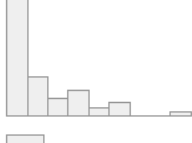
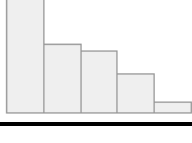
- Kaczurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., Sponheim, S. R., & Lissek, S. (2016). Neural Substrates of Overgeneralized Conditioned Fear in PTSD. *American Journal of Psychiatry*, 174(2), 125–134.  
<https://doi.org/10.1176/appi.ajp.2016.15121549>
- Karam, E. G., Friedman, M. J., Hill, E. D., Kessler, R. C., McLaughlin, K. A., Petukhova, M., Sampson, L., Shahly, V., Angermeyer, M. C., Bromet, E. J., Girolamo, G. de, Graaf, R. de, Demyttenaere, K., Ferry, F., Florescu, S. E., Haro, J. M., He, Y., Karam, A. N., Kawakami, N., ... Koenen, K. C. (2014). Cumulative Traumas and Risk Thresholds: 12-Month Ptsd in the World Mental Health (wmh) Surveys. *Depression and Anxiety*, 31(2), 130–142.  
<https://doi.org/10.1002/da.22169>
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Kubany, E. S., Haynes, S. N., Leisen, M. B., Owens, J. A., Kaplan, A. S., Watson, S. B., & Burns, K. (2000). Development and preliminary validation of a brief broad-spectrum measure of trauma exposure: The Traumatic Life Events Questionnaire. *Psychological Assessment*, 12(2), 210–224. <https://doi.org/10.1037//1040-3590.12.2.210>
- Lissek, S., & van Meurs, B. (2015). Learning models of PTSD: Theoretical accounts and psychobiological evidence. *International Journal of Psychophysiology*, 98(3, Part 2), 594–605.  
<https://doi.org/10.1016/j.ijpsycho.2014.11.006>
- Marin, M.-F., Hammoud, M. Z., Klumpp, H., Simon, N. M., & Milad, M. R. (2020). Multimodal Categorical and Dimensional Approaches to Understanding Threat Conditioning and Its Extinction in Individuals With Anxiety Disorders. *JAMA Psychiatry*, 77(6), 618–627.  
<https://doi.org/10.1001/jamapsychiatry.2019.4833>
- Maruff, P., Thomas, E., Cysique, L., Brew, B., Collie, A., Snyder, P., & Pietrzak, R. H. (2009). Validity of the CogState Brief Battery: Relationship to Standardized Tests and Sensitivity to Cognitive Impairment in Mild Traumatic Brain Injury, Schizophrenia, and AIDS Dementia Complex. *Archives of Clinical Neuropsychology*, 24(2), 165–178.  
<https://doi.org/10.1093/arclin/acp010>
- McLaughlin, K. A., Koenen, K. C., Friedman, M. J., Ruscio, A. M., Karam, E. G., Shahly, V., Stein, D. J., Hill, E. D., Petukhova, M., Alonso, J., Andrade, L. H., Angermeyer, M. C., Borges, G., de Girolamo, G., de Graaf, R., Demyttenaere, K., Florescu, S. E., Mladenova, M., Posada-Villa, J., ... Kessler, R. C. (2015). Subthreshold Posttraumatic Stress Disorder in the World

- Health Organization World Mental Health Surveys. *Biological Psychiatry*, 77(4), 375–384.  
<https://doi.org/10.1016/j.biopsych.2014.03.028>
- Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2017). Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Computational Psychiatry*, 1–21. [https://doi.org/10.1162/CPSY\\_a\\_00014](https://doi.org/10.1162/CPSY_a_00014)
- Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *ELife*, 7. <https://doi.org/10.7554/eLife.34779>
- Nord, C. L., Prabhu, G., Nolte, T., Fonagy, P., Dolan, R., & Moutoussis, M. (2017). Vigour in active avoidance. *Scientific Reports*, 7(1), 60. <https://doi.org/10.1038/s41598-017-00127-6>
- Orederu, T., & Schiller, D. (2018). Fast and slow extinction pathways in defensive survival circuits. *Current Opinion in Behavioral Sciences*, 24, 96–103.  
<https://doi.org/10.1016/j.cobeha.2018.06.004>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Pittig, A., Wong, A. H. K., Glück, V. M., & Boschet, J. M. (2020). Avoidance and its bi-directional relationship with conditioned fear: Mechanisms, moderators, and clinical implications. *Behaviour Research and Therapy*, 126, 103550. <https://doi.org/10.1016/j.brat.2020.103550>
- Scott, J. C., Matt, G. E., Wrocklage, K. M., Crnich, C., Jordan, J., Southwick, S. M., Krystal, J. H., & Schweinsburg, B. C. (2015). A quantitative meta-analysis of neurocognitive functioning in posttraumatic stress disorder. *Psychological Bulletin*, 141(1), 105–140.  
<https://doi.org/10.1037/a0038039>
- Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, 59(Suppl 20), 22–33.
- Sherbourne, C. D., & Stewart, A. L. (1991). The MOS social support survey. *Social Science & Medicine*, 32(6), 705–714. [https://doi.org/10.1016/0277-9536\(91\)90150-B](https://doi.org/10.1016/0277-9536(91)90150-B)
- Stevens, J. S., & Jovanovic, T. (2019). Role of social cognition in post-traumatic stress disorder: A review and meta-analysis. *Genes, Brain and Behavior*, 18(1), e12518.  
<https://doi.org/10.1111/gbb.12518>
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.2307/2346178>

- Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D. G., Marx, B. P., & Keane, T. M. (2013). *Clinician-Administered PTSD Scale for DSM-5 (CAPS-5)*.  
<https://www.ptsd.va.gov/professional/assessment/adult-int/caps.asp>
- Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD Checklist for DSM-5 (PCL-5)*.  
<https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp>
- Williams, M. B., Karg, R. S., & Spitzer, R. L. (2015). *Structured Clinical Interview for DSM-5—Research Version (SCID-5 for DSM-5, Research Version; SCID-5-RV)*. American Psychiatric Association.
- Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, 11(1), 4179. <https://doi.org/10.1038/s41467-020-17977-w>
- Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of Machine Learning Research*, 13(Apr), 1059–1062.

## Supplementary material

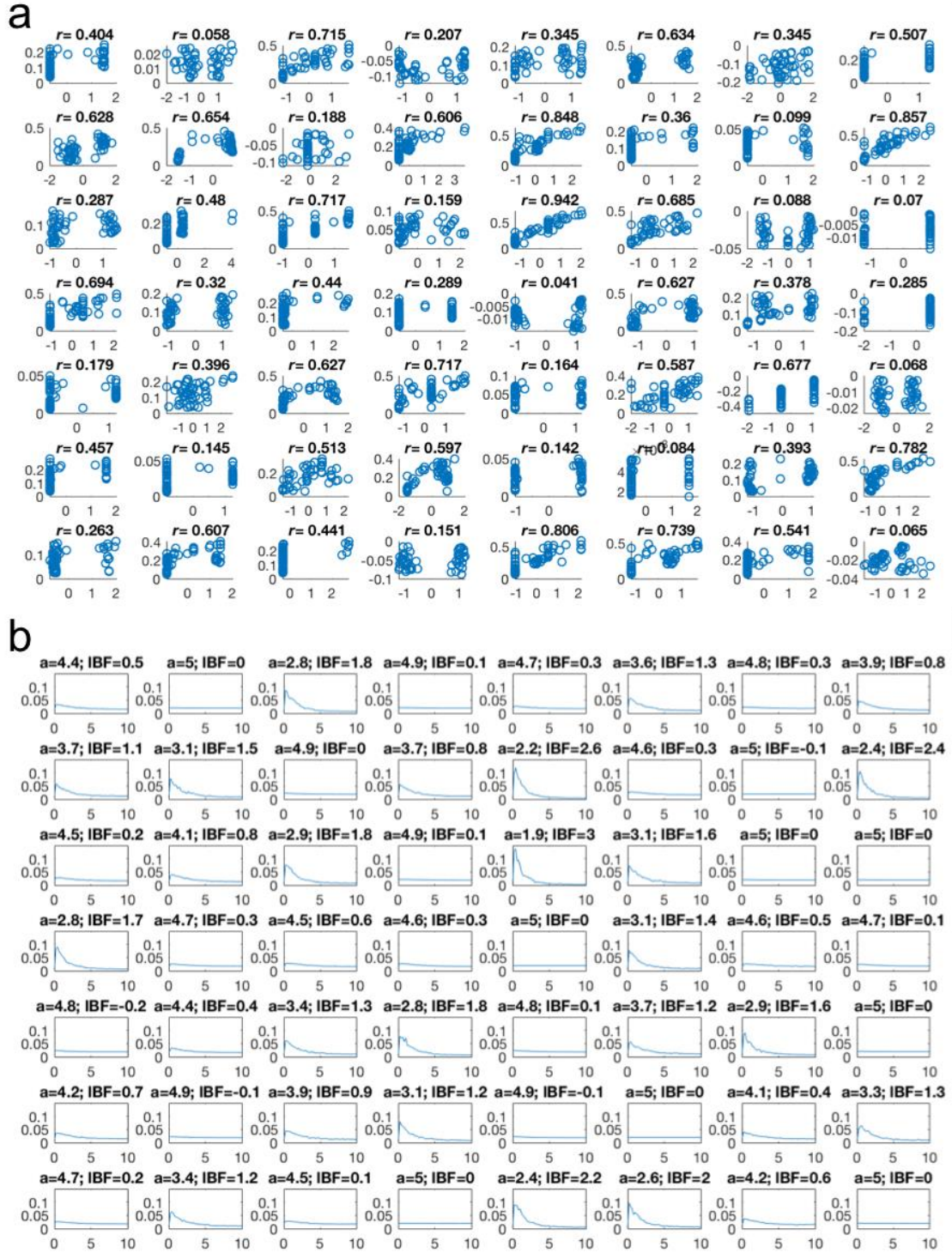


Clinical variable	Summary statistics	Histogram
PCL-5 re-experiencing symptoms	mean (SD): 7.9 (5.3) min < med < max: 0 < 9 < 18 IQR (CV) : 8.5 (0.7)	
PCL-5 avoidance symptoms	mean (SD): 4.1 (2.8) min < med < max: 0 < 4 < 8 IQR (CV): 4.2 (0.7)	
PCL-5 negative affect	mean (SD): 4.5 (3.6) min < med < max: 0 < 4.5 < 12 IQR (CV): 6.2 (0.8)	
PCL-5 anhedonia	mean (SD): 4.5 (3.8) min < med < max: 0 < 4.5 < 12 IQR (CV): 6.0 (0.9)	
PCL-5 externalizing behaviours	mean (SD): 2.2 (2.2) min < med < max: 0 < 2 < 9 IQR (CV): 3.0 (1.0)	
PCL-5 anxious arousal	mean (SD): 3.89(2.7) min < med < max: 0 < 4 < 8 IQR (CV): 4.2 (0.7)	
PCL-5 dysphoric arousal	mean (SD): 3.6 (2.7) min < med < max: 0 < 4 < 8 IQR (CV): 6.0 (0.7)	
BDI-II cognitive symptoms	mean (SD): 3.2 (4.0) min < med < max: 0 < 1.5 < 18 IQR (CV): 5.2 (1.2)	
BDI-II physical/affective symptoms	mean (SD): 7.75(7.0) min < med < max: 0 < 6 < 24 IQR (CV): 14.0 (0.9)	

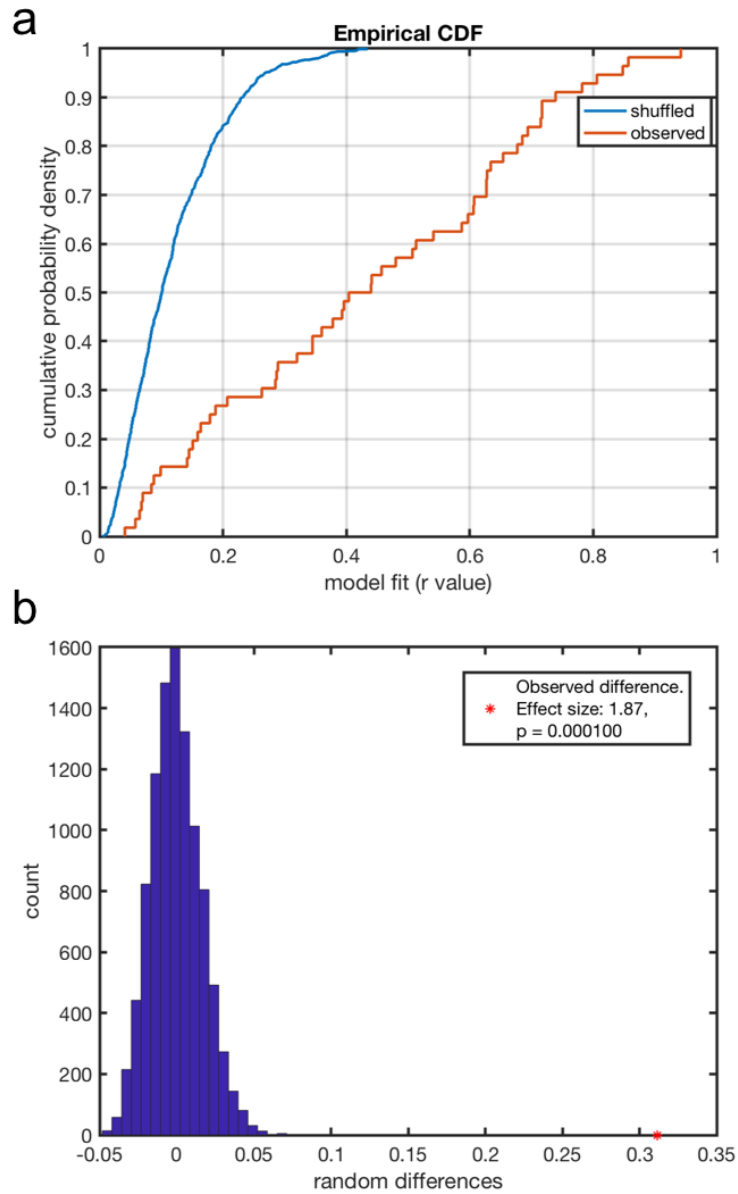
**Table S1. Summary statistics and distribution of scores for PTSD and depression symptom dimensions.** *PCL-5*, PTSD Checklist for DSM-5; *BDI-II*, Beck Depression Inventory version II; *IQR*, interquartile range; *CV*, coefficient of variation (relative standard deviation). Table generated using the R package summarytools (<https://cran.r-project.org/package=summarytools>).

<b>THS item</b>	<b>TLEQ item</b>
1. Life-threatening illness or injury	7. Have you ever had a life threatening illness?
2. A really bad car, boat, train, or airplane accident	2. Were you ever involved in a serious motor vehicle accident which required medical attention or that badly injured or killed someone?
3. A really bad accident at work or home	3. Were you ever involved in any 'Other' accident where you or someone else was badly hurt? (e.g. plane crash, drowning/nearly drowning, electrical or machinery accident, explosion, home fire, chemical leak, overexposure to radiation or toxic chemicals)
4. A hurricane, flood, earthquake, tornado, or fire	1. Have you ever experienced a natural disaster? (been caught in a flood, hurricane, earthquake).
5. Hit or kicked hard enough to injure - as a child	12. While you were growing up, were you ever physically punished in a way that resulted in bruises, burns, cuts, or broken bones?
6. Hit or kicked hard enough to injure - as an adult	9. Have you ever been hit/ beaten up and badly hurt by a stranger or by someone you didn't know well? OR 14. Have you ever been slapped, punched, kicked, beaten up, or otherwise physically hurt by your spouse (or former spouse), a boyfriend/ girlfriend, or some other intimate partner?
7. Forced or made to have sexual contact - as a child	15. Before your 13th birthday, did anyone who was at least 5 years older than you ever touch/ fondle your body in a sexual way, or make you touch/ fondle them in a sexual way? OR 16. Before your 13th birthday, did anyone close to your age ever touch you in a sexual way or make you touch them in a sexual way against your will or without your consent? OR 17. After your 13th birthday and before your 18th birthday, did anyone touch the sexual parts of your body or make you touch the sexual parts of their body against your will or without your consent?
8. Forced or made to have sexual contact - as an adult	18. After your 18th birthday, did anyone touch the sexual parts of your body or make you touch the sexual parts of their body against your will or without your consent?
9. Attacked with a gun, knife, or weapon	11. Has anyone ever threatened to kill you or cause you serious physical harm?
10. During military service - saw something horrible or was badly scared	4. Have you lived, worked, or had military service in a war zone and been exposed to warfare or combat? (e.g. been in the vicinity of a rocket attack, people being fired upon, seeing someone get wounded or killed).
11. Sudden death of close family member or friend	5. Have you experienced the sudden, unexpected death of a close friend or loved one?
12. Seeing someone die suddenly or get badly hurt or killed	10. Have you ever seen someone else attacked and seriously injured or killed?
15. Some other sudden event that made you feel very scared, helpless, or horrified	23. Have you experienced (or seen) any other events that were life threatening, caused serious injury, or were highly disturbing or distressing? (e.g. lost in the wilderness, a serious animal bite, violent death of a pet, being kidnapped or held hostage, seeing a mutilated body or body parts).

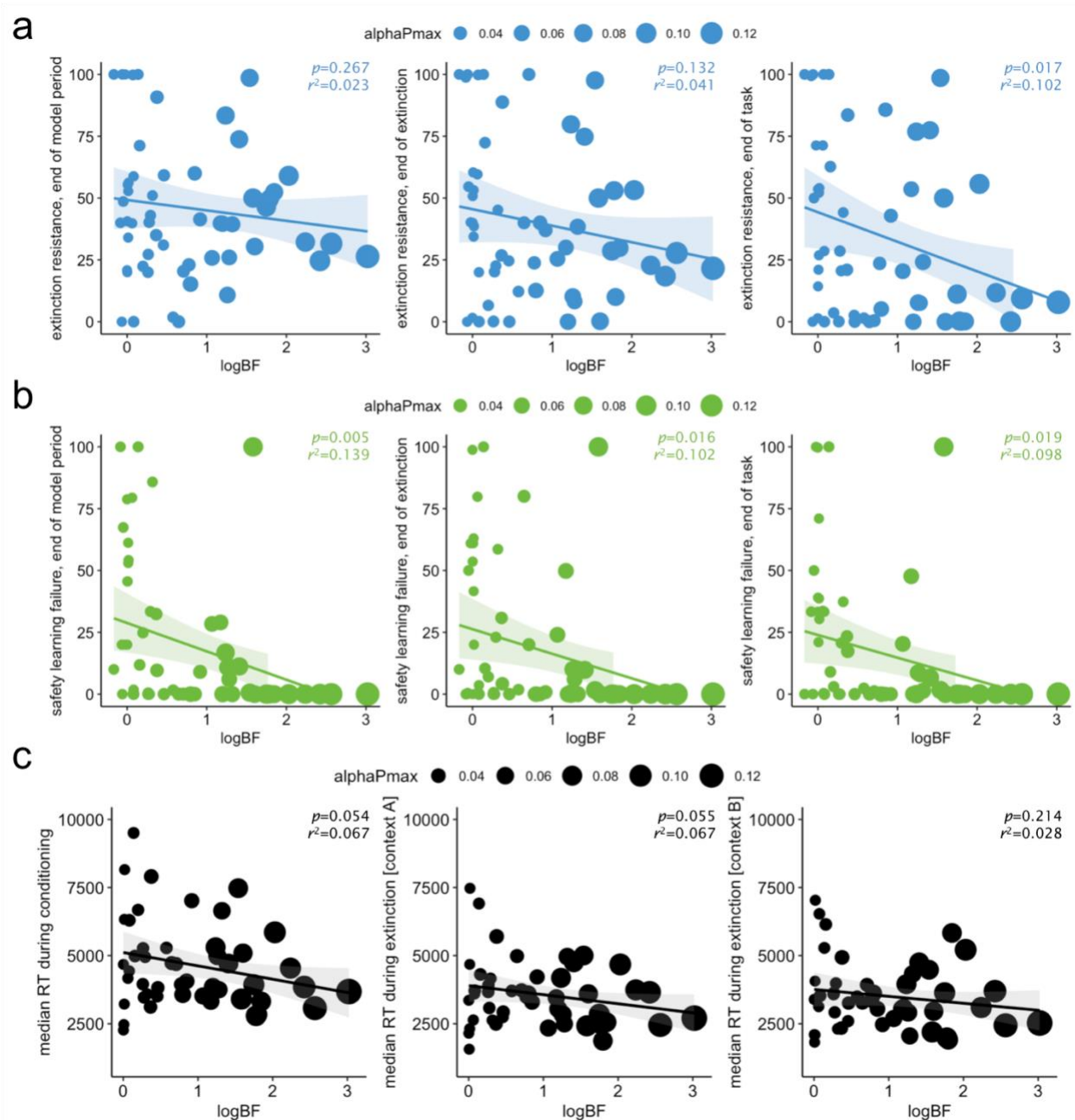
**Table S2. Derivation of lifetime trauma history scores from items probing common trauma types across the Trauma History Screen (THS) and Traumatic Life Events Questionnaire (TLEQ) measures.** For each of the 13 items (12 kinds of common trauma and one item probing experience other highly distressing events), participants scored 1 if they endorsed a particular event over their lifetime, and 0 otherwise (frequency-weighted response items from the TLEQ were collapsed to scores of 0 [never] or 1 [more than once]). Where multiple TLEQ items were deemed to be equivalent to a single THS item (THS items 6 and 7), endorsing any of the equivalent TLEQ items resulted in a score of 1 for that trauma type. Items from either scale that asked about events not probed in the other scale (e.g., sudden abandonment by partner or family from the THS, experience of miscarriage from the TLEQ) were not included – therefore this score may not represent complete trauma load for some participants. Across all participants, the mean score on this measure was 4.9 (SD 2.5, range 0-11, maximum possible score 13).



**Figure S1. Latent cause model output.** **a** Actual *vs* predicted loss expectancy ratings for each participant, as generated using the latent cause model of contextual extinction task data. *x axis*, actual ratings data (z-scored within-participants); *y axis*, predicted rating on each trial generated by the latent cause model. *r* values represent Pearson correlation coefficients between actual and predicted values for each participant. **b** Posterior distribution over alpha values for each participant. *x axis*, alpha value, *y axis*, posterior probability density (the prior was a uniform probability distribution over the interval [0,10]). *a*, posterior alpha estimate; *IBF*, logBF, or likelihood of a model where alpha>0 compared one where alpha=0.

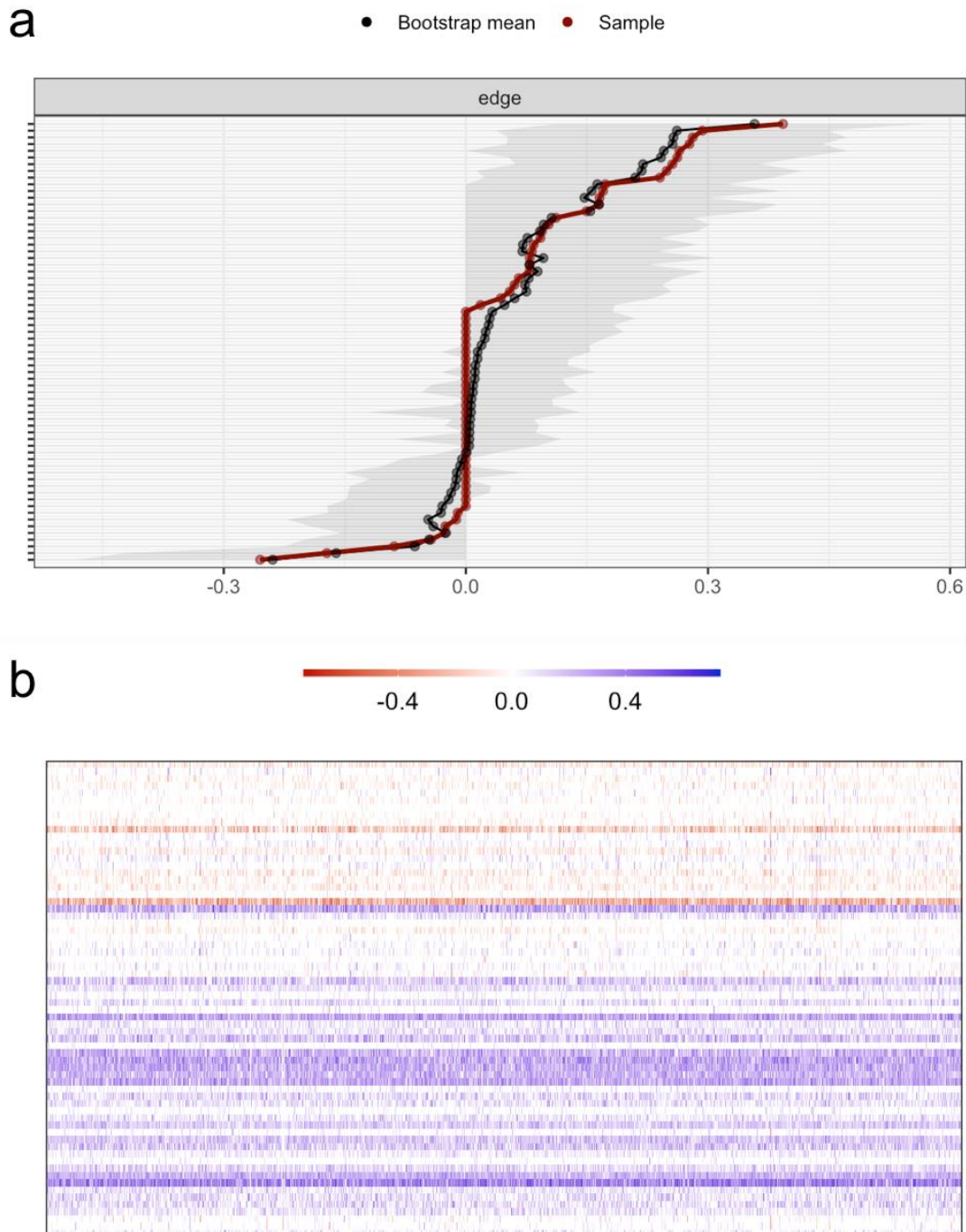


**Figure S2. Comparison of observed model fit values to those generated from randomly shuffled ratings data.** **a** In order to compare model fit ( $r$ ) values from observed data to randomly generated responses,  $N=1000$  dummy datasets were generated by random permutation (shuffling) of observed ratings data. These ratings data were then submitted to latent cause modelling and the model output compared to actual values in the same way as for the real data. **b** A permutation difference test was then used to formally compare the means of two distributions.

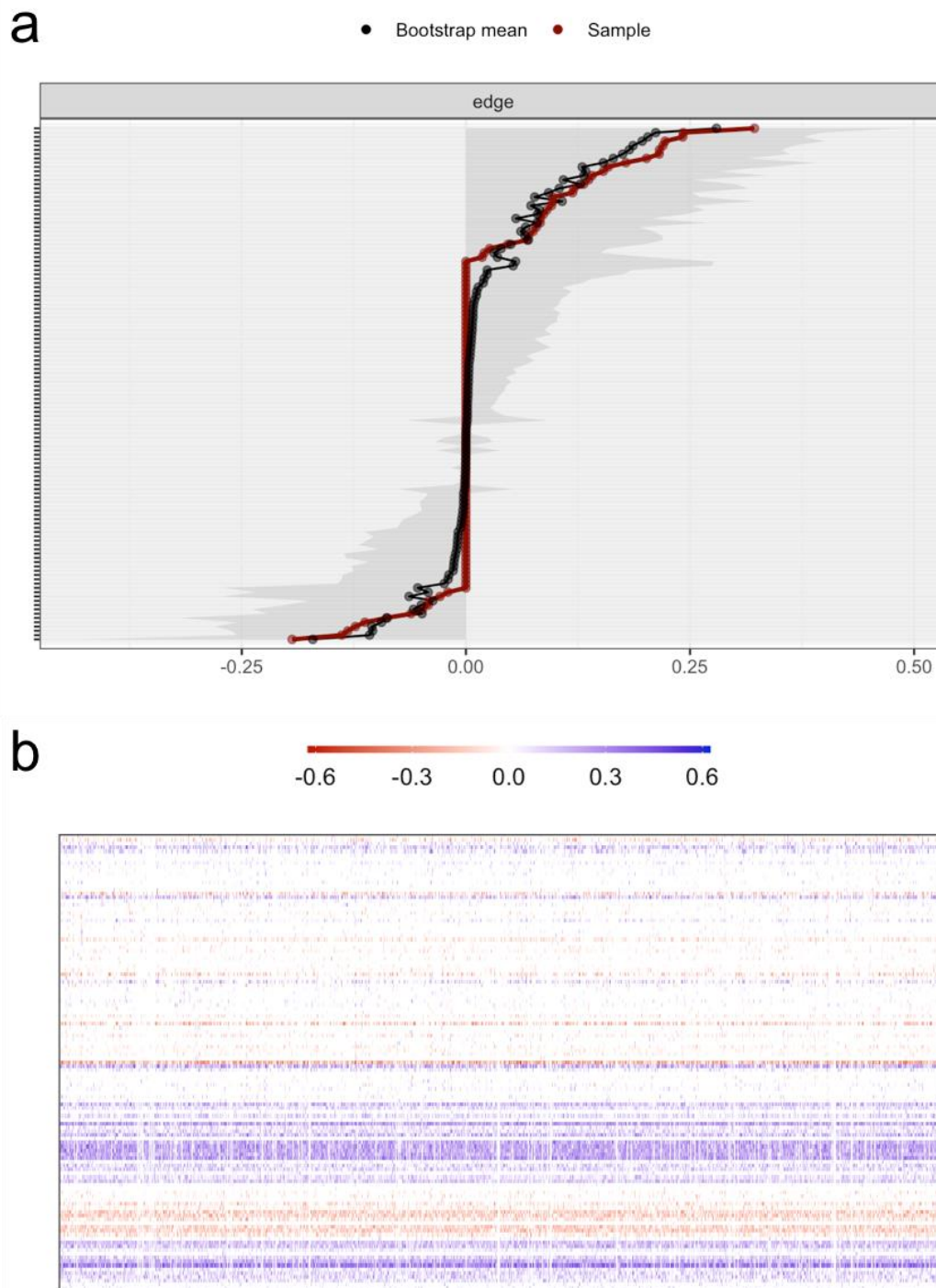


**Figure S3. Relationships between individual differences in latent cause inference and extinction resistance, safety learning, and response times on the contextual extinction task.** **a** Relationship between logBF values (probability of a model where  $\alpha > 0$ , compared to a single cause or  $\alpha = 0$  model), estimated from initial conditioning and first two blocks of extinction, and resistance to extinction index (mean residual CS+ loss expectancy rating) at the end of the modelled period (second block of extinction, block 5), end of extinction (block 6), and end of the task (block 9). **b** Relationship between logBF values and over-generalization or safety learning failure (CS- loss expectancy ratings, in the absence of any association between the CS- and the loss outcome) at the end of the modelled period (second block of extinction, block 5), end of extinction (block 6), and end of the task (block 9). **c** Relationship between logBF values and median response times (RTs) in ms to enter ratings during conditioning, initial extinction [in context A], and further extinction [in novel context B] stages. Regression lines represent linear model fit, weighted by posterior certainty in alpha parameter estimate ( $\alpha P_{max}$ ; higher certainty=larger dot size).





**Figure S4. Results of bootstrap analysis of edge strength accuracy for the network incorporating extinction task metrics with PTSD and depression symptom dimensions.** **a** Results of network stability analysis via nonparametric bootstrapping of edge weights ( $N=2000$  bootstraps). Each horizontal line represents one edge in the network, ordered from highest edge-weight to lowest edge-weight. The red line represents edge weight values in the same, the black line the mean value across bootstraps, and the gray area 95% bootstrapped CIs for each edge. **b** Multiverse plot representation of the same data as in **a**. Every row indicates an edge, and every column represents a bootstrap, with colour and intensity proportional to edge strength. Figures generated using the R package bootnet (<https://cran.r-project.org/package=bootnet>).



**Figure S5. Results of bootstrap analysis of edge strength accuracy for the network incorporating extinction task metrics with PTSD and depression symptom dimensions, and other sociodemographic information.** **a** Results of network stability analysis via nonparametric bootstrapping of edge weights ( $N=2000$  bootstraps). Each horizontal line represents one edge in the network, ordered from highest edge-weight to lowest edge-weight. The red line represents edge weight values in the same, the black line the mean value across bootstraps, and the gray area 95% bootstrapped CIs for each edge. **b** Multiverse plot representation of the same data as in **a**. Every row indicates an edge, and every column represents a bootstrap, with colour and intensity proportional to edge strength. Figures generated using the R package bootnet (<https://cran.r-project.org/package=bootnet>).