Psychological Medicine

cambridge.org/psm

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

Agnes Norbury[1], Hannah Brinkman[1], Mary Kowalchyk[1], Elisa Monti[1], Robert H. Pietrzak[2,3], Daniela Schiller[1,4] and Adriana Feder[1]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; [3]United States Department of Veterans Affairs, National Center for Posttraumatic Stress Disorder, Clinical Neurosciences Division, VA Connecticut Healthcare System, West Haven, CT, USA and [4]Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Original Article

Cite this article: Norbury A, Brinkman H, Kowalchyk M, Monti E, Pietrzak RH, Schiller D, Feder A (2022). Latent cause inference during extinction learning in trauma-exposed individuals with and without PTSD. *Psychological Medicine* **52**, 3834–3845. https://doi.org/10.1017/S0033291721000647

Received: 17 September 2020
Revised: 1 February 2021
Accepted: 9 February 2021
First published online: 8 March 2021

Key words:
Causal inference; extinction learning; PTSD; generalization; computational psychiatry

Author for correspondence:
Agnes Norbury,
E-mail: agnes.norbury@mssm.edu

Abstract

Background. 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 learning models have been unable to provide a convincing account of 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.

Methods. 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-characterised individuals with varying degrees of post-traumatic stress ($N = 56$). Model parameters describing individual differences in causal inference were related to multiple post-traumatic stress disorder (PTSD) and depression symptom dimensions via network analysis.

Results. Individuals with more severe PTSD were more likely to assign observations from conditioning and extinction stages to a single underlying cause. Specifically, greater re-experiencing symptom severity was associated with a lower likelihood of inferring that multiple causes were active in the environment.

Conclusions. We interpret these results as providing evidence of a primary deficit in discriminative learning in participants with more severe PTSD. Specifically, a tendency to attribute a greater diversity of stimulus configurations to the same underlying cause resulted in greater uncertainty about stimulus-outcome associations, impeding learning both that certain stimuli were safe, and that certain stimuli were no longer dangerous. In the future, better understanding of the role of causal inference in trauma-related psychopathology may help refine cognitive therapies for these disorders.

Introduction

Post-traumatic stress disorder (PTSD) can be thought of as 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 convincingly account for why such fear persists – particularly in the face of prolonged exposure (extinction) training, or when considering relapse (spontaneous return of fear) (Dunsmoor, Niv, Daw, & Phelps, 2015; Levy & Schiller, 2021). Recently, a novel computational account of extinction learning – latent cause modelling – has been proposed by Gershman, Niv, and colleagues (Gershman & Niv, 2010, 2012; Gershman, Blei, & Niv, 2010). 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, context, 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, Monfils, Norman, & Niv, 2017). According to this account, the inappropriate fear responses observed in post-traumatic stress syndromes

CAMBRIDGE UNIVERSITY PRESS

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, and (2) failure to successfully form an extinction memory in the first place.

Computationally, the first case can be formalised as a 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 et al., 2010; Gershman & Niv, 2012; Gershman, Norman, & Niv, 2015). Indeed, a 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 prerequisites 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, Hammoud, Klumpp, Simon, & Milad, 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 (Kaczkurkin et al., 2016; Lissek & van Meurs, 2015; Norbury, Robbins, & Seymour, 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 a 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 associated with that cause) (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-characterised 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, 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, Kindt, Fanselow, & Beckers, 2017; Pittig, Wong, Glück, & Boschet, 2020), and there is some evidence that avoidance-related traits predict poorer response to cognitive therapy for PTSD (Badour, Blonigen, Boden, Feldner, & Bonn-Miller, 2012; Békés, Beaulieu-Prévost, Guay, Belleville, & Marchand, 2019). Following recent theoretical developments that favour modelling psychological disorders including post-traumatic stress 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 PTSD and depression symptom dimensions concurrently in an exploratory network analysis (see Armour, Fried, Deserno, Tsai, & Pietrzak, 2017; de Haan *et al.*, 2020; Fritz, Fried, Goodyer, Wilkinson, & van Harmelen, 2018; Greene, Gelkopf, Epskamp, & Fried, 2018).

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, a 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, Shahar, Hauser, & Dolan, 2018).

Methods

Participants

Participants were World Trade Center (WTC) 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 both studies included individuals who currently met diagnostic criteria for full or subthreshold PTSD (for full inclusion/exclusion criteria see online Supplementary Material), with one study also including trauma-exposed individuals who were assessed as never having met criteria for PTSD. 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.

Clinical and sociodemographic measures

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 Structured Clinical Interview for DSM-5 and Clinician-Administered PTSD Scale for DSM-5 (Weathers et al., 2013a; Williams, Karg, & Spitzer, 2015), and for $N = 31$ participants this was the Mini-International Neuropsychiatric Interview for DSM-5 (Sheehan et al., 1998). Additionally, participants completed the PTSD checklist for DSM-5 (PCL-5) and The Beck Depression Inventory version II (BDI-II) self-report measures of PTSD and depression symptoms (Beck, Steer, & Brown, 1996; Weathers et al., 2013b). PTSD symptoms were parsed into seven dimensions (re-experiencing, avoidance, negative affect, externalizing behaviour, anxious arousal, and dysphoric arousal symptoms clusters) which have been previously been demonstrated to provide the best account of symptoms

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

data across multiple 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 previous longitudinal analysis of diverse samples of depressed individuals (Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015). Information about lifetime trauma history and perceived levels of social support was also available (see online Supplementary Material). 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).

Extinction learning task

Participants completed an 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 three phases: an initial aversive conditioning phase (in context A), extinction learning phase (in context A), and further extinction learning (in novel context B) (Fig. 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 extinction phase CS+ trials 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. In order to test the feasibility for future remote work, the extinction learning task was administered online (see online Supplementary Material).

Analysis

Statistical analyses were carried out in R version 3.6.1 (R Core Team, 2019) and MATLAB R2019a (MathWorks Inc., 2019). Analysis code and version information for R packages is available at https://github.com/agnesnorbury/latent-cause-PTSD.

Extinction task data

Effects of within-task manipulations (effects of CS type and task stage) on loss expectancy ratings and response input times were explored using repeated-measures analysis of variance (ANOVA; see online Supplementary Material). Extinction resistance was defined as mean loss expectancy rating for the aversively conditioned CS (CS+) 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 the difference in values between CS+ and non-aversively conditioned (CS−) stimuli, as – in contrast to other types of data such as global signal regression or BOLD signals – expectancy ratings have an absolute meaning. Further, experimental evidence suggests that individuals with PTSD may over-generalise negative information from conditioned to unconditioned stimuli (see Introduction) – 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). Since, across the group as a whole, (1) we observed maintained differential responding to CS+ and CS− stimuli at the end of initial extinction training (indicating incomplete learning), and (2) there were no obvious effects of the change-in-context manipulation on learning traces (Fig. 1b), we did not investigate potential 'recall' effects (such as spontaneous recovery) between extinction trials at the end of context A and start of context B. Such an analysis may also be of limited validity here, as, in contrast to previous investigations (e.g. Gershman and Hartley, 2015), there was no significant temporal delay between the two extinction training phases.

Latent cause modelling of extinction task data

Latent cause modelling of loss expectancy ratings data was carried out using code associated with Gershman and Niv (2012) (https://github.com/sjgershm/LCM). 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 also held out, so that model output would be unbiased by trials used to calculate extinction resistance both at the end of this stage (context A), and following further extinction learning (in context B).

Briefly, the model assumes that the participant learns to associate groups of stimuli they observe with different underlying states or causes. On each trial, participants compute the posterior probability that a given cause c generated the observed configuration of stimuli (here, a 3D binary vector representing presence/absence of the CS+, CS−, and US), using Bayes' rule:

$$P(\text{cause} = c \mid \text{stimuli}) \propto P(\text{stimuli} \mid \text{cause} = c) \times P(\text{cause} = c)$$

The inferred probability of an existing cause c being active on a given trial, given the observation of the trial stimuli, is proportional to the likelihood of that cause (consistency between current stimuli and prototypical stimulus configuration associated with cause c), multiplied by a prior term that indexes an individual's preference for simpler or more complex causal structures (Fig. 2a). 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 α (i.e., the distribution over states is modelled using a Chinese Restaurant Process with concentration parameter α – see Gershman & Niv, 2012; Gershman *et al.* 2015). Smaller values of α bias individuals towards simpler clusterings, where observations tend to be assigned to the same cause, and larger values towards more complex clusterings, where observations are assigned to different causes. As the learner has some uncertainty about the stimulus configuration associated with each cause, the output on each trial is a posterior probability distribution across potential underlying causes (each learner starts with an internal representation consisting of a single cause, and more causes are added as required by the model, up to a maximum limit). The model was fit to task data under a generative framework, by comparing how well models with a range of different α values could account for participants' task performance (see online Supplementary Material).

The key output submitted to further analysis was the likelihood (for each participant) of a model where α was allowed to be >0 (i.e., with multiple inferred causes), compared to a model where $\alpha = 0$ (single underlying cause), computed as a log Bayes factor (logBF). A $\text{logBF} \geqslant 1$ is generally interpreted as representing strong evidence in favour of the comparator hypothesis (here, in favour of a multi-cause model) (Kass & Raftery, 1995).

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

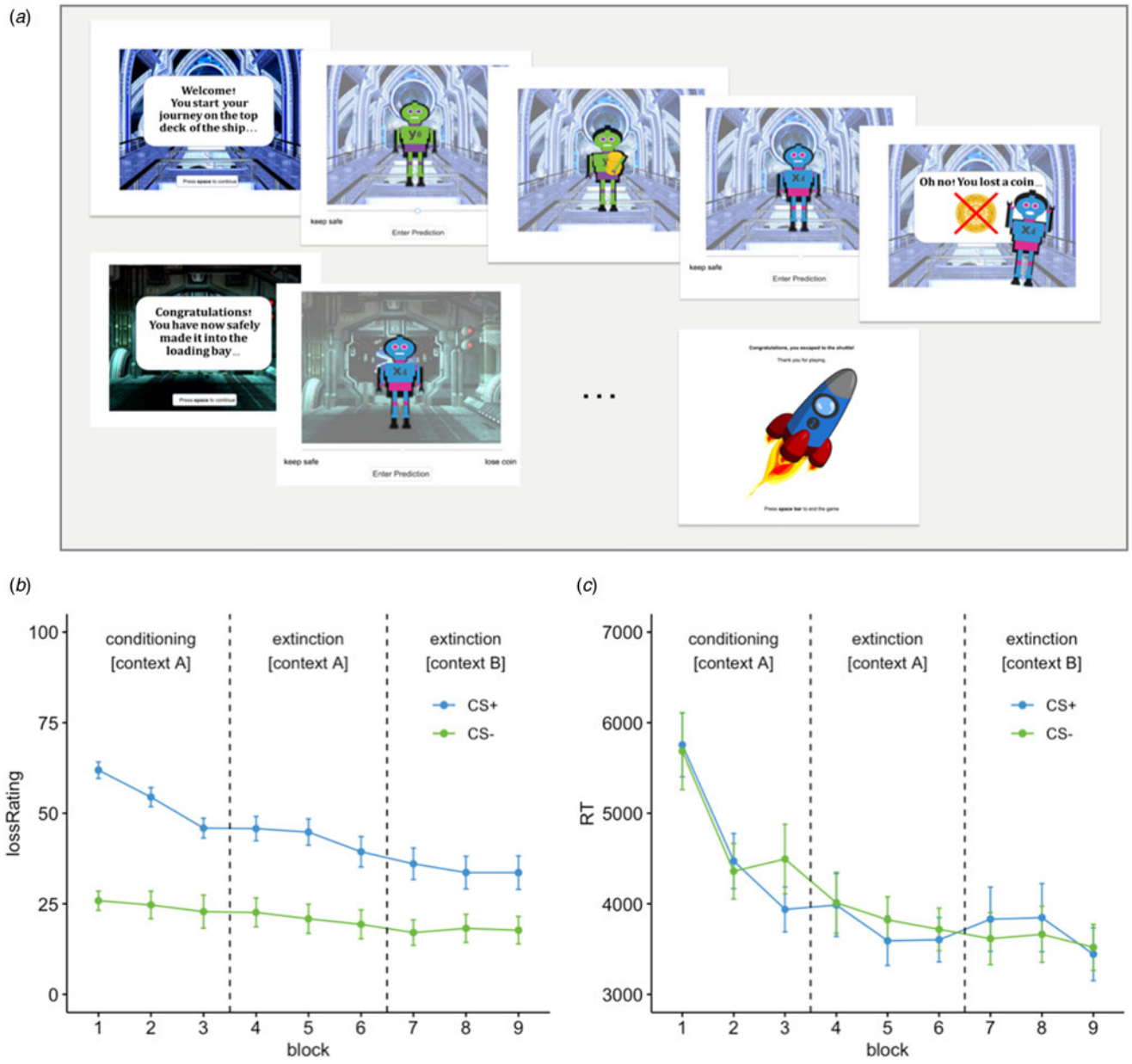

Fig. 1. Data from the online extinction learning task demonstrated that participants learned to discriminate between conditioned and unconditioned stimuli, and to decrease loss expectancy ratings for conditioned stimuli following the transition to extinction. (a) Depiction of trials from the online extinction 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 thought 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*(lose a coin|CS+) was 1/3 during initial conditioning and reduced to 0 during extinction training stages, [*P*(lose a coin|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 the 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 (each three blocks with ten trials per block). (c) Median RTs to input ratings, by CS type and task stage. Error bars represent the standard error of the mean. CS+, aversively conditioned (loss-associated) stimulus; CS−, non-loss-associated stimulus.

In order to assess goodness of fit, actual *v*. 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 the goodness of fit (r) values derived from the actual data to $N = 1000$ randomly shuffled dummy datasets. The ability of the model to reliably recover α and logBF estimates from task data was assessed using simulation and recovery analysis (see online Supplementary Material). Trial-by-trial associations between response times (RTs) and internal model quantities were examined using linear mixed models, using the R package lmerTest.

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 (α), logBF

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

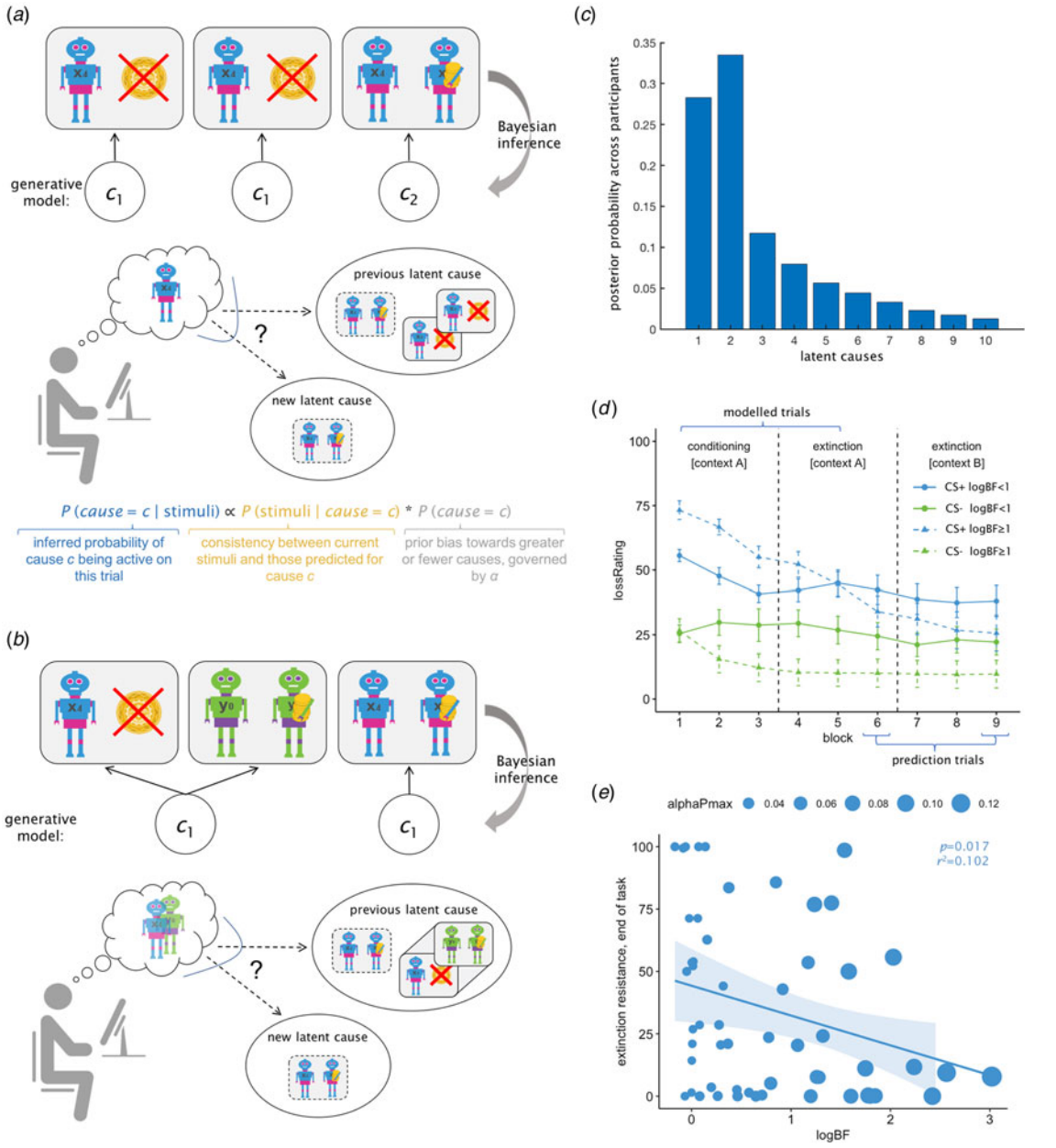

Fig. 2. Latent cause modelling of extinction task data revealed that trauma-exposed individuals whose extinction task data were better explained by a single cause model discriminated less between conditioned and unconditioned stimuli during initial learning, and showed greater resistance to extinction. (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 the 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, α). 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 overgeneralisation of aversive information, or hampering of safety learning by hyperarousal). Disparate configurations of stimuli and outcomes (CS+, CS−, US, and US omission) may be clustered together in the inferred causal structure of the environment, leading to greater uncertainty about the pattern of stimuli and outcomes associated with a given cause, and therefore slower learning of expected 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 a single cause model (where $\alpha = 0$). For illustration, behavioural data are displayed separately for individuals for whom model comparison favoured a model with more than one cause (logBF $\geqslant 1$, 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 the 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. The regression line and *p*-value represent linear model fit, weighted by posterior certainty in α parameter estimates (αPmax; higher certainty = larger dot size). Panels (*a*) and (*b*) are adapted from Gershman et al. (2015).

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

estimates were bivariately related to behavioural and clinical measures using weighted least squares regression, with regression weights set equal to the peak of the posterior probability distribution over α values. As probability distributions sum to 1, this peak value is inversely proportional to the width of the 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 was allocated to estimates from participants for whom this quantity was more confidently derived (online Supplementary Fig. S1b).

Network analysis of latent cause inference, clinical, and sociodemographic data

Network analysis was used to explore relationships between individual differences in latent cause inference and severity of multiple PTSD and depression symptom dimensions, whilst taking into account other relevant clinical and demographic factors. Specifically, networks were constructed using regularised Gaussian graphical estimation implemented in the R package qgraph, version 1.6.5 (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). Under this approach, nodes represent observed variables, and connections between nodes (edges) represent unique pairwise associations (partial correlation coefficients), after conditioning on all other variables in the dataset. The application of regularisation during estimation is intended to remove spurious connections from the network, such that any retained edges can be thought of as contributing meaningfully to the overall variance (according to simulation studies, at low ratios of a number of observations to a number of potential connections, 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; Williams, Rhemtulla, Wysocki, & Rast, 2019). The stability of parameter estimation and power-related properties of the network analysis were assessed via non-parametric bootstrap and simulation analyses, using functions from bootnet, version 1.4.3 (Epskamp, Borsboom, & Fried, 2018). For full methodology (per Burger et al., 2020), see online Supplementary Material. Extinction resistance (mean CS+ loss expectancy rating at the task) and safety learning failure (mean loss expectancy for the CS− at the end of the task), were included in the networks as well as logBF values in order to ascertain if the latent cause model parameter was more closely related to symptoms than these simple behavioural performance indices (i.e., had greater explanatory power than behavioural differences alone). As per Armour et al. (2017), two nested networks were estimated: one consisting of PTSD/depression symptoms scores and behavioural task variables alone, and one with additional clinically-relevant covariates (age, education level, perceived level of social support, and additional lifetime trauma history).

Results

Participants

Demographic and clinical data for study participants are summarised in Table 1. $N = 42$ (75%) participants currently met DSM-5 criteria for full or subthreshold WTC-related PTSD (mean PCL-5 total score 40.4 ± 11.0), and $N = 14$ (25%) participants were 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 ± 1.7). Participants with full or subthreshold PTSD also reported moderate levels of depression symptoms (PTSD group, mean BDI-II total score 14.5 ± 9.6; resilient group, mean BDI-II total score 0.43 ± 1.1). However, PTSD is known to be highly heterogeneous (Armour et al., 2015; Contractor, Roley-Roberts, Lagdon, & Armour, 2017), and previous analyses in WTC responders and other populations have revealed reliable differences in patterns of covariance across symptom clusters (Horn et al., 2016; Pietrzak et al., 2014). In our sample, PTSD subscores exhibited continuous variation across participants (online

Table 1. Summary of demographic and clinical variables for study participants ($N = 56$)

Age	53 (6.9)
Gender (*N* female)	19 (34%)
Race (*N*)	
Black or African American	6 (11%)
Asian	4 (7%)
Native American	1 (2%)
White or Caucasian	37 (66%)
Other	2 (4%)
Ethnicity (*N*)	
Hispanic/Latinx	12 (21%)
Education level (*N*)	
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 (*N*)	
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 (*N*)	
SSRI/SNRI (stable dose)	3 (6%)
NDRI (stable dose)	3 (6%)
sedative (night-time use only)	3 (6%)
Additional lifetime trauma history	
N trauma categories endorsed (0–13)	4.9 (2.5)
Childhood physical abuse (*N*)	16 (29%)
Childhood sexual abuse (*N*)	14 (25%)
Adulthood sexual trauma (*N*)	7 (13%)

Values represent mean (s.d.) 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 online Supplementary Table S1. All study participants had DSM-5 category A trauma exposure to the WTC disaster in 2001. For details about additional lifetime trauma categories, and how these were defined, see online Supplementary Table S2.

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

Supplementary Table S1), with only moderate correlations observed between subscores (mean $r = 0.67 \pm 0.16$; online Supplementary Fig. S1) – justifying the use of a multidimensional approach to PTSD symptomatology in our analysis.

Extinction learning task

Manipulation check

In order to check if participants performed as expected on the task, loss expectancy ratings and RTs were analysed by repeated-measures ANOVA. 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) (online Supplementary Results; Fig. 1b). Participants responded more quickly at later stages in the task, but median RTs remained >3000 ms, indicating preservation of relatively considered responding (Fig. 1c).

Latent cause modelling of extinction task data

Model fit and validation

In order to assess how well the model accounted for our data, observed loss expectancy ratings on each trial were plotted against model-predicted output for each participant (online Supplementary Fig. S2). Across participants, the mean correlation between actual and predicted loss expectancy ratings was 0.459 (S.D. 0.25). Permutation difference testing revealed that the mean r value for actual *v.* predicted loss ratings in our sample was significantly greater than that generated by fitting the same model to randomly shuffled data (mean r for shuffled data = 0.118; difference = 0.312, $p < 0.001$; online Supplementary Fig. S3). 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).

Simulation and recovery analysis revealed good parameter estimate stability and identifiability for task data (correlation between simulated and recovered α values = 0.838, $p < 0.001$; correlation with recovered β values = 0.053, $p > 0.4$) (online Supplementary Fig. S4a). Comparison of recovered logBF estimates for datasets simulated with $\alpha = 0$ *v.* $\alpha > 0$ revealed significantly different likelihood estimates [$t_{273} = -111$, 95% confidence interval (CI) −3.70 to −3.84, $p < 0.001$; Welch's two-sample t test]; with datasets simulated with $\alpha = 0$ favouring a single cause model (mean logBF = -1.03 ± 0.01), and datasets simulated with $\alpha > 0$ favouring a multi-cause model (mean logBF = 2.74 ± 0.17) (online Supplementary Fig. S4b).

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 (Fig. 2c; across participants, the marginal probability of a third cause was 0.117). In order to visualise differences in behaviour associated with model output, participants were divided into two groups, defined according to whether their behaviour provided strong evidence in favour of a multi- (*v.* single) cause model (logBF $\geqslant 1$ *v.* logBF < 1). Similar to Gershman and Hartley (2015), individuals whose responses provided no strong evidence in support of a multi-cause account appeared to learn more slowly across both conditioning and extinction stages (shallower curves for participants with logBF < 1, Fig. 2d). Formal comparison by fitting a simple linear slope to CS+ loss expectancy ratings over the course of the modelled period revealed significantly shallower gradients in the lower logBF group (logBF < 1, mean gradient = -2.67 ± 8.5; logBF $\geqslant 1$ mean gradient = -7.21 ± 7.1; $p = 0.034$, Wilcoxon signed-rank test).

Notably, the lower logBF 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 (logBF < 1, mean difference in rating = 18.2 ± 24.7; logBF $\geqslant 1$ mean difference in rating = 43.4 ± 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 lower logBF group (logBF < 1: mean CS+ rating = 46.2 ± 14.3, mean CS− rating = 28.0 ± 23.6; logBF $\geqslant 1$: mean CS+ rating = 58.3 ± 13.0, mean CS− rating = 15.0 ± 21.5; $p = 0.002$, $p = 0.035$, respectively; Wilcoxon signed-rank tests). This suggests that the slower extinction learning in individuals with low evidence of a multi-cause model might be a result of more similar observation representations across different trial types [CS+ (reinforced), CS+ (unreinforced), and CS− trials) in these individuals (Fig. 2b).

Importantly, simulated datasets where α was constrained to be close to or >0 were able to replicate this behavioural pattern: with the $\alpha \sim 0$ group (favouring a single latent cause model) showing both shallower gradients in CS+ loss expectancy ratings and the lower difference in ratings between CS+ and CS− stimuli than the $\alpha > 0$ group (favouring a multi-cause model), over the same task period (both $p < 0.001$, Wilcoxon signed-rank tests) (online Supplementary Fig. S5). This pattern was also robust to the choice of logBF threshold used to define groups (online Supplementary Material; Fig. S6).

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 α estimate), but were significantly related to extinction resistance at the end of the task (block 9, $\beta = -11.9$, $p = 0.017$; Fig. 2e, online Supplementary Fig. S7a). This suggests that latent cause inference during initial learning might predict future resistance to extinction training – with a 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 blocks 5, 6, and 9, respectively; all $p < 0.020$), although these associations appear substantially non-linear (online Supplementary Fig. S7b). This suggests that latent cause inference may also relate to either heightened generalisation 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

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

neurocognitive test battery that were available for a subset of participants. Although this test is likely underpowered ($N = 24$), we found no evidence of a relationship between Cogstate composite scores and logBF estimates ($\beta = 0.010$, $p = 0.338$). Values of the scaling parameter β, which may reflect individual differences in the use of the response scale, were unrelated to extinction resistance or failure of safety learning at the end of the modelled period, end of the initial extinction learning stage, or end of the task ($p > 0.11$, Spearman's rank correlation tests).

Relationship to RTs

In order to further test whether lower logBF scores might reflect a tendency towards a stimulus-independent or inattentive response style, we also examined whether median RTs during each stage of the task were related to logBF estimates. Interestingly, there was marginal evidence of a *negative* relationship between logBF and median RTs during conditioning and initial extinction learning stages ($\beta = -552$, $p = 0.054$; $\beta = -469$, $p = 0.055$) – with longer median RTs associated with lower logBF scores (during further extinction in novel context B: $\beta = -341$, $p = 0.214$; online Supplementary Fig. S7c). This may indicate greater uncertainty about predicted values in lower logBF individuals during initial learning and extinction training (Hyman, 1953).

To examine this relationship more precisely, we analysed trial-by-trial variance in RT as a function of uncertainty over underlying latent causes and logBF estimates, in linear mixed models controlling for time-on-task (trial number) and expected value (predicted probability of loss) (cf. Brown et al., 2018). Uncertainty about active causes was approximated by taking the maximum over the posterior probability distribution across causes on each trial, which is directly proportional to the certainty that the most likely cause was responsible for trial observations. We found that RTs were slower for individuals with lower logBF estimates [$\beta = -1395$ (S.E. 370) ms, $p < 0.001$], and on trials with a greater likelihood of a single particular cause [$\beta = 9457$ (S.E. 456) ms, $p < 0.001$]. Results were unchanged if the time-on-task effect was modelled as an exponential decay function, as suggested by Fig. 1c (see online Supplementary Material). This finding is consistent with slower RTs reflecting greater uncertainty about the configuration of observations associated with a *particular cause* (greater variance in prototypical stimulus and outcome vectors associated with that cause), rather than greater uncertainty about *which* cause was active on a given trial – in particular for individuals with greater tendency to group all observations as being the results of a single cause.

Relationship between latent cause inference and PTSD symptoms

Relationship with avoidance symptoms

In individual linear regression models weighted by posterior certainty in α 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 ($\beta = -5.6$, $p = 0.069$), Fig. 3a. Specifically, individuals with lower logBF values, indicating a 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 ($\beta = -1.9$, $p = 0.234$). β scaling parameter values were not related to PCL-5 avoidance symptoms, PCL-5 total score, or BDI-II total score ($p > 0.3$, Spearman's rank correlation tests).

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

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 (regularised edge weight −0.089, bootstrapped 95% CI for edge value = −0.224–0, Fig. 3b). As re-experiencing symptoms were positively connected to avoidance symptoms (regularised 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; regularised edge weight −0.026, bootstrapped 95% CI −0.153 to 0; for sample weights and bootstrapped 95% CIs for all network edges see online Supplementary Fig. S8). 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 [interquartile range (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 (regularised edge weight −0.041, bootstrapped 95% CI for edge value −0.175 to 0, Fig. 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 (regularised edge weights −0.029, −0.020; bootstrapped 95% CIs −0.272 to 0, −0.234 to 0; respectively; for sample weights and bootstrapped 95% CIs for all edges see online Supplementary Fig. S9). The overall structure of the network between PTSD/depression symptoms and extinction task variables was robust to the inclusion of the additional covariates, as the correlation between the edge weights derived from covariate controlled and non-covariate-controlled networks was high (Spearman's $\rho = 0.955$). 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.

Discussion

Here, we provide preliminary evidence that individual differences in latent cause inference, as measured during a simple behavioural extinction learning paradigm, may be related to the 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

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

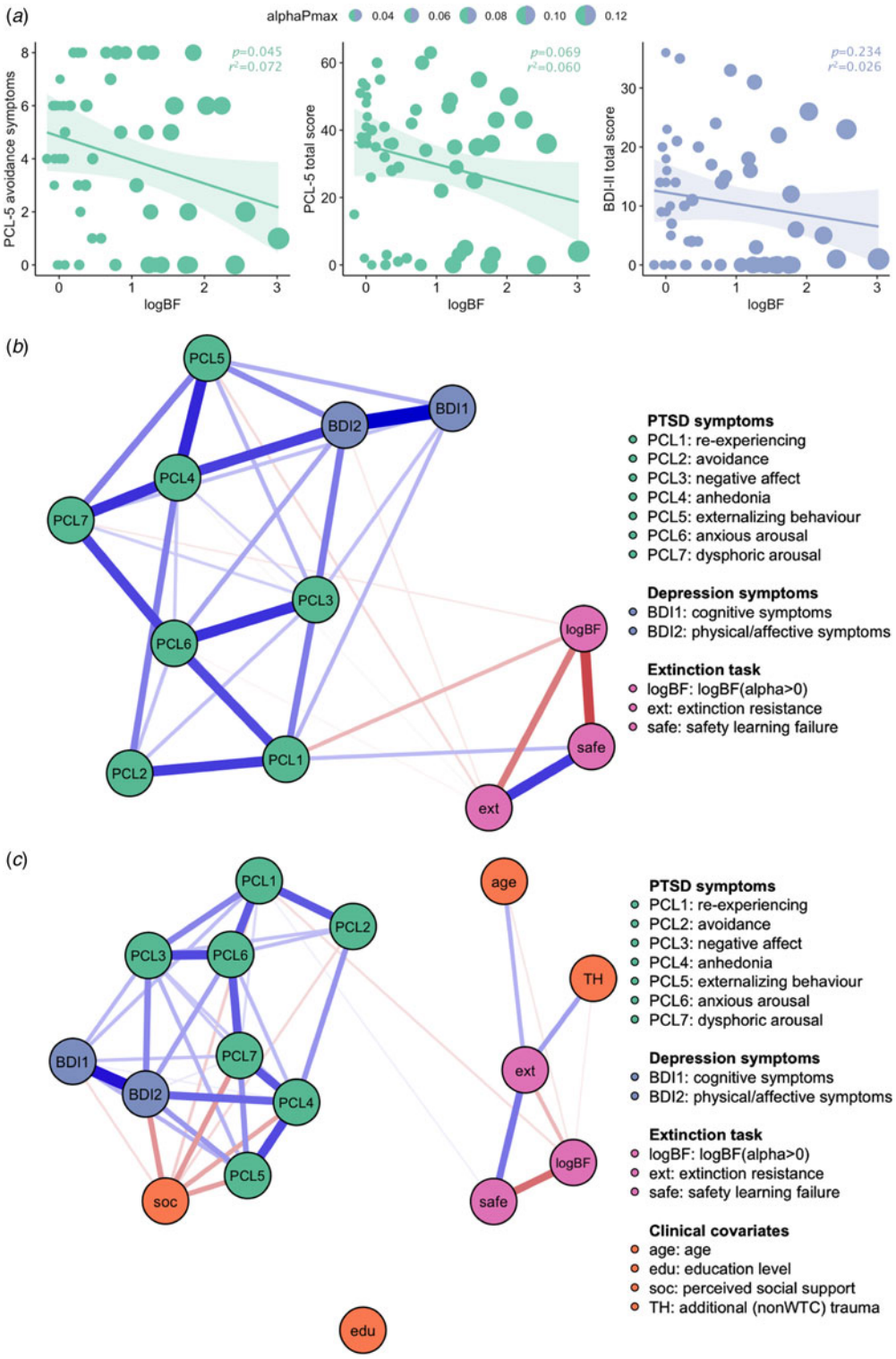

Fig. 3. Trauma-exposed individuals with greater tendency to infer that a single cause was responsible for observations across conditioning and extinction stages reported more severe PTSD, but not depression, symptoms. (a) Bivariate relationships between latent cause inference and avoidance, total PTSD, 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 α (αPmax; higher certainty = larger dot size). (b) Regularised 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). (c) Regularised network model incorporating PTSD and depression symptoms, extinction task performance measures, and clinically relevant covariates: specifically age, self-reported education level, perceived social support (Medical Outcomes Study Social Support Survey total score), and additional lifetime trauma history. For both networks, connections between nodes (edges) represent partial correlation coefficients retained following least absolute shrinkage and selection operator (LASSO) regularisation, a conservative approach that 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 (max value = 0.4).

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

model with a single underlying cause exhibited greater resistance to extinction training in future trials, poorer safety learning, and higher levels of avoidance symptoms. In line with previous observations that, during the same measurement occasion, within-subjects deviations in internal avoidance symptoms are significantly associated with within-subjects deviations in the occurrence of flashbacks (Greene et al., 2018; Hoffart, Langkaas, Øktedalen, & Johnson, 2019), our exploratory cross-sectional network analysis incorporating multiple PTSD and depression symptom clusters indicated that the bivariate association between latent cause inference and avoidance may be mediated via greater severity of PTSD re-experiencing symptoms (intrusive thoughts, nightmares, flashbacks, and emotional/physiological reactivity to reminders). Importantly, this multivariate analysis also controlled for individual differences in task performance (extinction resistance and safety learning failure), indicating that the 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 WTC disaster in 2001). Participants reported a range of current PTSD symptom levels, from minimal symptoms (resilient) to severe cases (mean PCL-5 total score 31 ± 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 symptoms and task network model revealed satisfactory specificity for a discovery analysis (0.77) – minimising the chance of identifying false positive connections (estimated sensitivity of our analysis, or rate of discovery of true positives was slightly lower at 0.69: therefore, some true connections may be missing from the identified network structure).

Considering model parsimony (only two free parameters) and the continuous nature of the response variable, the latent cause model generally provided a good account of participants' loss expectancy data – however the extent to which this was truly varied across subjects (online Supplementary Fig. S2). 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 α values). In weighted models, the likelihood of a multi-cause model [$\log BF(\alpha > 0)$] was negatively associated with extinction resistance (residual CS+ loss expectancy ratings) at the end of the task (Fig. 2). LogBF values were also strikingly negatively related to the 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 (online Supplementary Fig. S7). Individuals with lower logBF values were also slower to enter ratings across the modelled period – which may indicate greater uncertainty about expected values (Hyman, 1953; McDougle & Collins, 2021).

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 'overgeneralisation'), rather than a 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 multiple different potential combinations of observations (CS+, CS−, US and US omission) as being produced by the same underlying cause (Fig. 2b). This results in greater uncertainty about the likelihood of specific stimulus-outcome associations *within the overarching causal structure*: hampering both initial learning of correct CS–US associations (successful discriminative learning during conditioning), and subsequent learning that these associations have changed (during extinction). This explanation is consistent with the behavioural pattern observed in individuals whose behaviour supported lower likelihood of a multi-cause model: who showed both slower learning and less differentiated responses to CS+ *v.* CS− stimuli (Fig. 2d).

Intriguingly, two recent computational studies have identified greater weighting of previous error signals during value updating for aversively conditioned stimuli in PTSD, associated with greater volatility in stimulus value estimates (Brown et al., 2018; Homan et al., 2019). This over-correction in the face of errors in prediction might be expected if individuals with PTSD have less confidence or certainty in their internal model of the environment. Here, we propose that this results not from greater uncertainty about which causes are active in their environment, but from greater uncertainty about specific stimulus-value associations *within their internal representation of that cause,* exacerbated by – or reflected in – a tendency to group all observations as resulting from a single underlying cause (a greater diversity of observations attributed to the same latent cause results in the greater estimated variance of the observation prototype associated with that cause, and therefore greater likelihood that further disparate observations will be assigned to that cause; Gershman and Niv, 2012).

An alternative explanation for our findings is that logBF values and current symptoms levels may both be 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, these data were 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 probing beliefs about stimulus value and task structure.

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 and PTSD symptoms (e.g. Brown *et al.*, 2018; Norbury *et al.*, 2018; Wise & Dolan,

https://doi.org/10.1017/S0033291721000647 Published online by Cambridge University Press

2020), more evidence is needed that the outcome employed here is engaging the kind of cognitive processes relevant to the processing of traumatic experience. Future studies using this framework will therefore explicitly probe the aversiveness of the loss outcome to study participants, and further, attempt to increase emotional engagement with the task by using more immersive graphics and taking a more gamified approach to task presentation (see Nord et al., 2017; Wise & Dolan, 2020, for successful examples of this approach). Another important difference between the data presented here and that from some previous investigations is that there was no significant temporal delay between the two extinction training sessions, such that we are unlikely to be probing 'recall' of associations from longer-term memory. It is therefore possible the effects described here are specific to mechanisms subserving rapid extinction learning, which may be at least partially distinct from those supporting longer-term learning (Orederu & Schiller, 2018).

Finally, effect sizes reported here are modest – with our measure of latent cause inference explaining around 7–8% of the variance in self-reported PTSD symptoms. However, these associations persisted under a conservative (regularised) analysis approach, which tends to shrink connection 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 for PTSD (Feder et al., 2016; Karam et al., 2014). It will be necessary to test if these relationships persist in 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., 2018).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291721000647.

Financial support. This work was supported by CDC-NIOSH U01 awards to AF (grant nos. OH011473 and OH010729), and a NARSAD Young Investigator award from the Brain and Behavior Research Foundation to AN (grant no. 28604).

Conflict of Interest. AF is named co-inventor on a patent application in the USA, and several issued patents outside the USA filed by the Icahn School of Medicine at Mount Sinai related to the use of ketamine for the treatment of PTSD. This intellectual property has not been licensed. All other authors have no relevant conflicts of interest to declare.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Data availability statement. De-identified raw data for the contextual extinction task are 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 but are available from the authors on request.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th edn). Washington, DC: Author. 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

Badour, C. L., Blonigen, D. M., Boden, M. T., Feldner, M. T., & Bonn-Miller, M. O. (2012). A longitudinal test of the bi-directional relations between avoidance coping and PTSD severity during and after PTSD treatment. *Behaviour Research and Therapy*, *50*(10), 610–616. https://doi.org/10.1016/j.brat.2012.06.006

Beck, A. T., Steer, R. A., & Brown, G. K.. (1996). *Manual for the Beck depression inventory-II*. San Antonio, TX: Psychological Corporation.

Békés, V., Beaulieu-Prévost, D., Guay, S., Belleville, G., & Marchand, A. (2019). Trauma-related negative cognitions mediate the relationship between avoidant personality beliefs and impeded response to psychotherapy for PTSD. *Journal of Aggression, Maltreatment & Trauma*, *28*(3), 297–312. https://doi.org/10.1080/10926771.2018.1500504

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

Brown, V. M., Zhu, L., Wang, J. M., Frueh, B. C., King-Casas, B., & Chiu, P. H. (2018). Associability-modulated loss learning is increased in posttraumatic stress disorder. *ELife*, *7*, e30150. https://doi.org/10.7554/eLife.30150

Burger, J., Isvoranu, A.-M., Lunansky, G., Haslbeck, J., Epskamp, S., Hoekstra, R. H. A., … Blanken, T. (2020). *Reporting standards for psychological network analyses in cross-sectional data*. *PsyArXiv*. https://doi.org/10.31234/osf.io/4y9nz

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., … … 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

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., … 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*, *48*(4), 1–18. 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

Feder, A., Mota, N., Salim, R., Rodriguez, J., Singh, R., Schaffer, J., … … 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

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., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209. https://doi.org/10.1037/a0017808

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

Hoffart, A., Langkaas, T. F., Øktedalen, T., & Johnson, S. U. (2019). The temporal dynamics of symptoms during exposure therapies of PTSD: A network approach. *European Journal of Psychotraumatology*, *10*(1), 1618134. https://doi.org/10.1080/20008198.2019.1618134

Holt, D. J., Boeke, E. A., Wolthusen, 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

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., … Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature Neuroscience*, *22*(3), 470. https://doi.org/10.1038/s41593-018-0315-x

Horn, S. R., Pietrzak, R. H., Schechter, C., Bromet, E. J., Katz, C. L., Reissman, D. B., … Feder, A. (2016). Latent typologies of posttraumatic stress disorder in World Trade Center responders. *Journal of Psychiatric Research*, *83*, 151–159. https://doi.org/10.1016/j.jpsychires.2016.08.018

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188–196. https://doi.org/10.1037/h0056940

Kaczkurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., … 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., … … 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

Levy, I., & Schiller, D. (2021). Neural computations of threat. *Trends in Cognitive Sciences*, *25*(2), 151–171. https://doi.org/10.1016/j.tics.2020.11.007

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

McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–30. https://doi.org/10.3758/s13423-020-01774-z.

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2018). Computation in psychotherapy, or How computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, *2*, 50–73. https://doi.org/10.1162/CPSY_a_00014.

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *ELife*, *7*, e34779. 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

Pietrzak, R. H., el-Gabalawy, R., Tsai, J., Sareen, J., Neumeister, A., & Southwick, S. M. (2014). Typologies of posttraumatic stress disorder in the U.S. adult population. *Journal of Affective Disorders*, *162*, 102–106. https://doi.org/10.1016/j.jad.2014.03.024

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., … 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., … 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.

Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D. G., Marx, B. P., & Keane, T. M. (2013a). *Clinician-administered PTSD scale for DSM-5 (CAPS-5)*. Washington, DC: U.S. Department of Veterans Affairs. 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. (2013b). *The PTSD Checklist for DSM-5 (PCL-5)*. Washington, DC: U.S. Department of Veterans Affairs. 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)*. Arlington, VA: American Psychiatric Association.

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*(5), 719–750. https://doi.org/10.1080/00273171.2019.1575716

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