# Controllability Priors Modulating Over- and Under-Segmentation of Latent Causes in Fear Conditioning

Ines Aitsahalia

April 22nd, 2022

Advisors: Yael Niv, Sashank Pisupati

*I pledge my honor that this thesis represents my own work in accordance with University regulations.*

*/s/ Ines Aitsahalia*

This Senior Thesis was submitted to Princeton University in partial fulfillment of the requirements for the Degree of Bachelor of Arts in Neuroscience & Certificate in Cognitive Science

**Table of Contents**

**Acknowledgements**

Foremost, I would like to acknowledge the considerate and conscientious mentorship of Professor Yael Niv and Dr. Sashank Pisupati, without whom this thesis would never have been possible. Yael—you inspire me as a teacher, scientist, activist, and person. Thank you for supporting every step of this project and giving me the tools to explore what is most interesting to me. Between class, lab, and the subcommittee on inclusive teaching, I could not imagine my time at Princeton without you. Sashank—your dedication to not only developing but patiently explaining neuroscience theory is the groundwork of this thesis. I am so thankful for your guidance, wisdom, and appreciation of my jokes. If you'll forgive one more terrible pun: you are a role *model*, both as a modeler and a mentor.

I would also like to thank all members of the Niv Lab for their continued support of my growth as a researcher and person, especially Sam Zorowitz, whose technical help made running the experimental parts of this thesis possible, and whose kindness and humor have kept me going since NEU 201, and Branson Byers, whose organization, knowledge of the IRB, and positive attitude as lab manager made PNI feel like home. I want to acknowledge the other members of the Screams Task team: Isabel, Jamie, Yongjing, Maddie, and Sebastian, whose work has been foundational to this thesis, and whom I hope to keep collaborating with for a long time.

This thesis would also not be possible without the generous funding granted by Princeton's Program in Cognitive Science as well as the Nancy J. Newman, MD '78 & Valerie Biousse, MD Senior Thesis Research Fund for Neuroscience. Thank you to Dr. Paryn Wallace and Anna Colasante for your help managing the logistics of these grants and believing in my work.

Thanks as well to all my professors at Princeton who gave me fundamental knowledge to apply to this thesis, either directly or indirectly, especially Nathaniel Daw and Cate Peña.

Thank you to EPSP, Empowering diversity and Promoting Scientific equity at PNI, which has been a constant support in my time here.

Of course, I also want to acknowledge the participants in this study: wherever in the world you are, thank you!

I must also extend my most sincere thanks to my friends both inside and out of science. Thank you for your love, reassurance, joy, and reminder that there exists life outside of jspsych and angry shapes. Thank you specifically to Meigan Clark, who co-produced our second thesis: *Lia.* Thank you for trusting me with so vulnerable and beautiful a play as well as for your constant friendship and support of my "big science."

Thank you to my partner, Adam, for the long FaceTime sessions while I debugged a single line of code, for your support of me pursuing neuroscience anywhere in the world, and for distilling happiness into a smile.

And finally, to my family, especially my parents: Sophie Meunier and Yacine Aitsahalia, for their unending love and support (including their tolerance of listening to loud screams playing in their house on repeat all weekend). Je vous aime.

Thank you!

**Abstract**

Despite the prevalence of fear-based psychiatric disorders, the computational mechanisms underlying persistent, generalized fear remain unknown. This thesis argues that under- and over-segmentation of latent causes best explains the different forms of extinction failure observed in fear conditioning, ranging from overgeneralization of fear and slow learning, to "over-accommodation" and rapid new associative learning without updating old beliefs. Through the latent cause model, these different learning regimes can be achieved through one parameter: the observation prior, which regulates the degree to which observations in the world are thought to be stochastic or determinstic. This parameter has an intuitive relationship with many common cognitive distortions, such as black-and-white thinking. I begin by describing the latent cause model and its application to fear conditioning by simulating existing behavioral results. Then, I present a new, online differential fear conditioning task to validate "virtual shock," replicate previous results capturing individual differences in fear extinction, and evaluate model predictions. I conclude by discussing ongoing work extending this model to ask questions about trauma's effect on fear learning, by considering trauma as an uncontrollable stress that can modulate observation priors.

## 1. Introduction

Nearly 20% of American adults are diagnosed with a fear-based disorder, such as obsessive compulsive disorder (OCD), panic disorder, or post-traumatic stress disorder (PTSD) (Harvard Medical School, 2005). These disorders can be debilitating and are often characterized by avoidance behavior (Mahoney et al., 2018), increased negative beliefs (Ramos-Cejudo & Salguero, 2017), hormonal changes (Thorsell, 2010), and persistence of fear, even after exposure to contrary evidence (Moutoussis et al., 2018). People who suffered early life stress—such as physical or sexual abuse, loss, and housing or food insecurity, among other stressors—are at a higher risk of later developing these disorders than the general population (Famularo et al., 1992; Fierman et al., 1993; Heim & Nemeroff, 2001). However, the computational mechanisms underlying persistent, generalized fear and why it may arise more in some individuals than others remain unknown.

Fear conditioning, a form of classical or Pavlovian conditioning, can be used as a laboratory proxy for acquired fear. Its undoing, termed "fear extinction" or, simply, "extinction," can be used to understand how learned fear diminishes in safe contexts, providing a useful model behind many exposure therapies, such as Cognitive Behavioral Therapy (CBT) (Nair et al., 2020). However, extinction (and therapy) can fail (Dunsmoor et al., 2015), and many patients report resurgence of fear after treatment (Moutoussis et al., 2018). One useful Bayesian learning model explaining extinction failure is the latent cause model, which posits that individual differences in the failure to extinguish fear emerge from differences in the inferred causal structure of the environment (Gershman & Niv, 2010, 2012).

The latent cause model assumes that, rather than directly associating cues with outcomes, individuals infer hidden (or "latent") causes underlying both stimuli. Differences in this inferred casual structure—such as single versus separate causes underlying acquisition and extinction – could explain the different observed fear learning outcomes (see Figure 1 for more details) (Gershman & Niv, 2010, 2012). The number of latent causes created in response to observations, which I refer to as segmentation, can be optimal (where the inferred structure perfectly matches the structure of the environment) or not. I posit that latent cause creations lie along a continuum surrounding optimal

segmentation, with under-segmentation referring to the creation of too few causes and over-segmentation to the spurious creation of too many causes.

Recently, failures in extinction have been shown to be more common in participants with PTSD, accompanied by an under-segmentation of latent causes (Norbury et al., 2021), suggesting a link between the disorder and over-generalization of fear. Among healthy adults, exposure to uncontrollable stress before fear conditioning and extinction has also been shown to lead to extinction failure (Hartley et al., 2014).

This thesis seeks to combine these results in a unified theoretical manner, examining the computational basis for individual differences in latent cause assignments in fear conditioning and relating this to a single parameter: the variance of the observation prior, which I then connect to beliefs in controllability. Through simulations, I show that extreme values of this observation prior parameter lead to under- or over-segmentation of latent causes, resulting in different learning regimes which have been observed in individuals, both with and without psychiatric illness (Gershman & Hartley, 2015; Norbury et al., 2021).

I begin by describing the latent cause model and an intuitive understanding of its application to fear conditioning by simulating existing behavioral results. Then, I present a new, online differential fear conditioning task to help validate "virtual shock" paradigms, replicate previous results capturing individual differences in fear extinction, and present preliminary results supporting this task's use in measuring people's observation priors. Next, I present current work on extending this model to ask questions about trauma's effect on fear learning, by considering trauma as uncontrollable stress that modulates the observation prior. I conclude by incorporating results, both theoretical and behavioral, into a holistic framework for understanding individual differences informed by life experience. This is not only crucial to clinicians treating patients with fear-based disorders, but it also contributes to an emerging literature on trauma-informed care, Disability studies (Rashed, 2019), and the newly coined "dignity neuroscience" (White & Gonsalves, 2021).

## 2. Theoretical Background

This section discusses the latent cause model's applications to fear extinction in two parts: first, mathematical assumptions and second, behavioral simulations of other experimental work and predictions for the new task.

### 2.1 Latent Cause Model Definition and Assumptions[1]

In the standard view of extinction, the learned association is that the conditioned stimulus (shown in Figure 1 as a lightbulb) causes the unconditioned stimulus (shown in Figure 1 as a lightning bolt for the shock) and therefore the unconditioned response (such as freezing or tensing up). However, there are other possible structures. For instance, one latent cause could explain both the cue and the shock, or both could be caused by two unrelated latent causes.

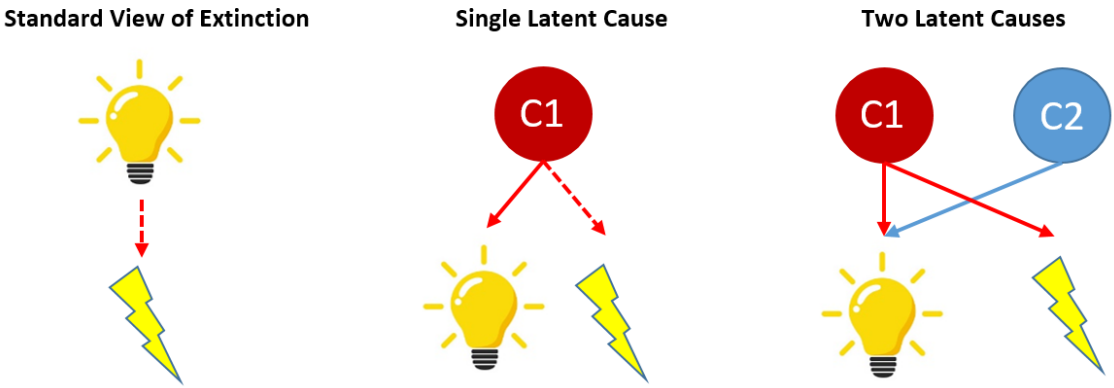


**Figure 1.** Visual representation of different possible structures of relationships between causes, a single cue, and a single outcome, based on the model of Gershman and Niv (2010).

While in the standard view and the single latent cause structure (shown in the two leftmost panels) repeated presentation of the light without shock would lead to a weakening of the association and extinction of fear, the two latent cause structure has separate underlying causes for the grouping of the light and shock and the light on its own. For an agent with this model structure, repeated presentation of the light without

[1] This section contains text that is based closely on, or identical to, text found in my junior paper.

shock would do nothing to target the fear association held in C1, which could lead to spontaneous recovery of fear when presented with the light again later.

People cluster experiences together, and these clusters dictate what associations are updated during learning (Gershman & Niv, 2010, 2012). Therefore, accurately targeting learned associations during extinction requires identifying the correct cluster. To do so, this model uses nonparametric Bayesian inference with an infinite capacity prior—also called a "Chinese Restaurant Process (CRP)." (Navarro & Perfors, n.d.) The CRP metaphor explains how in a crowded restaurant with infinite tables (possible latent causes), each incoming customer (current trial) will be seated, most often at the same tables as other customers. There can never be more non-empty tables than there are customers, and each customer can only sit at one table. For every customer, there is also a probability that they will be seated at a new, empty table, and this probability depends on how popular the currently non-empty tables are (Equation 1).

Translated to the latent cause inference model, the assumptions are as follows: first, each trial observation (of both cues and outcomes) can only be ascribed to a single latent cause; second, the total number of possible causes is only bounded by the total number of trials, but current trial outcomes are more likely to be assigned to heavily-assigned causes (Lloyd & Leslie, 2013).

$$P(C_{t+1}|C_{1\ldots t}) = \begin{cases} P(C_{t+1} = k) = \dfrac{N_k}{t+\alpha} \\ P(C_{t+1} = K+1) = \dfrac{\alpha}{t+\alpha} \end{cases}$$

***Equation 1.*** The probability that the next customer (trial) sits at (is caused by) the kth table (cause) is proportional to the number of other customers seated at it (number of trials explained by the cause). This is calculated as follows: $\sum_{k=1}^{k} N_k + \alpha = t + \alpha$ . The probability of being assigned to a new, K+1th table, completes the space of all possible options for an incoming customer.

This clustering is useful because it allows for the generalization of association learning. Consequently, the posterior probability of a particular cluster assignment is given by Bayes' rule, shown in Equation 2.

$$P(C_{1:t}|F_{1:t}) = \frac{P(F_{1:t}|C_{1:t})P(C_{1:t})}{P(F_{1:t})}$$

***Equation 2.*** Bayes' rule, in an applied form to the probability of a set of cluster assignments given all the features observed. This describes the probability of an event given another based on the prior conditions and context. The probability that the current cause *C* is the real cause given the current observation in feature *F* is equal to the probability of the observation given the latent cause multiplied by the independent probability of the cause, all over the independent probability of the observation. The solution to the applied form is intractable due to the marginalization over all possible cluster assignment, $C_{1:t}$, but it can be approximated using particle filtering.

This posterior distribution, $P(C_{1:t}|F_{1:t})$, is the agent's belief about the underlying latent cause structure of their experience after making an observation. Exact inference in this model is not tractable, since it requires marginalizing over all possible cause allocations, which grow exponentially with trials. Therefore, this model approximates the inference using particle filtering (for more, see Gershman & Niv, 2012; Pisupati, 2021). Finally, the likelihood term, $P(F_{1:t}|C_{1:t})$, which represents the likelihood of the observations if the person's cause attributions were true, requires an additional layer of Bayesian inference. Since the actual probability that a given observation is caused by a specific latent cause is unobservable, the agent must infer it. The model assumes that people may have non-uniform Beta priors over observation probabilities, and that binary observations are generated from a Bernoulli process dependent on probabilities phi, or $\phi_{i,k}$, as shown in Equation 3. The inferred probabilities resulting from this process contribute to the likelihood of cause assignments or "importance weights" during inference and importance sampling, and are updated after every trial yielding learned associations (Speekenbrink, 2016).

$$P(F_t|c_t - k) - P\left(f_t|_{C_t} - k\right) - \prod_P \left(f_{i,t}\Big|_{C_t} - k\right) - \prod \phi_{i,k}; P(\phi_{i,k}) = Beta(\alpha, \beta);\ \alpha, \beta \begin{cases} = 1 \\ < 1 \\ > 1 \end{cases}$$

***Equation 3.*** For each feature *f* in the feature vector F, random observations are generated through a Bernoulli process governed by probabilities $\phi_{i,k}$. These are formulated via a Beta prior.

The observation prior is a prior over phi in Equation 3, and it can be thought of as a parameter that defines the degree of an agent's belief in the randomness of the world. Starting out with a "deterministic" prior, i.e. a belief that probabilities are often all-or-

nothing, leads to surprising observations being assigned to new causes. Creating too many latent causes can give the impression of rapid learning but does not extinguish the original association, failing to generalize between acquisition and extinction trials. Such over-accommodating individuals could be more susceptible to spontaneous recovery of fear. On the other hand, a "stochastic" prior, i.e. a belief that probabilities are often close to chance, leads to attribution of disparate observations to a single cause and overgeneralization. The creation of too few latent causes leads to poor discrimination between stimuli and therefore poor extinction performance through the generalization of fear to safe stimuli.

This process detailed above considers every trial as unique, which is important to the interleaved nature of most fear conditioning experiments, however, it does not account for blocking and time effects.

The success of fear extinction can also be tested temporally, with spontaneous recovery and relearning measures, which this basic CRP model cannot capture. Experimental evidence shows that animals and people show the resurgence of a previously extinguished fear when presented with the same stimuli as in extinction after a certain amount of time. Agents are also faster to relearn an old fear association if it begins being reinforced again than to learn a new association (Myers & Davis, 2007). To account for the importance of the time interval between extinction and spontaneous recovery or relearning, an additional parameter must be added to the model: temporal persistence (Pisupati, 2021).

This can be modeled using an extension of the CRP for non-interchangeable data, called the Distance Dependent CRP (DDCRP) (Blei & Frazier, 2011). Rather than assigning incoming customers to tables, the DDCRP connects customers to each other, then determining table assignments, allowing for the preservation of identification or customers (Blei & Frazier, 2011). Translated to the model, this allows the time point of the observation to matter, while still functioning on the interleaved presentation of stimuli.

## 2.2 Model Predictions and Simulations

Previous research using this latent cause model framework for fear extinction has distinguished between individuals who attribute fear acquisition and extinction trials to a single cause and those who attribute these to different causes. Healthy participants in the latter group showed higher resurgence of fear than the former (Gershman and Hartley, 2015) while participants with PTSD were more likely to fall into the former category (Norbury et al. 2021) (Figure 2).

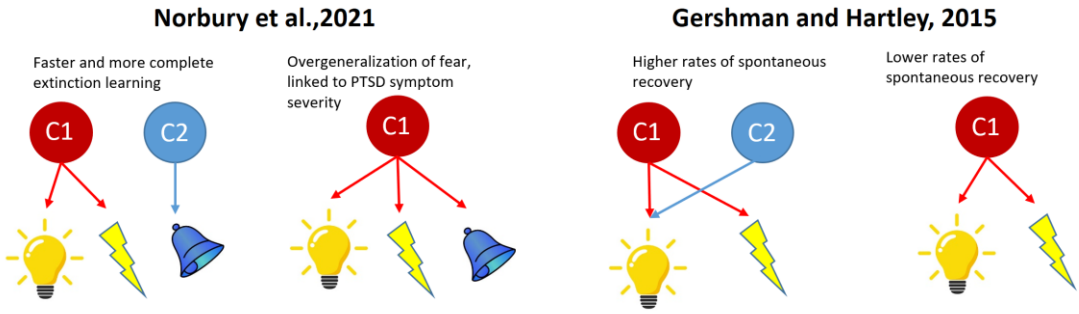


**Figure 2.** Latent cause regimes and their consequences as found by Norbury et al. 2021 and Gershman and Hartley 2015. Attribution of optimal learning to either one-cause or two-cause regimes in these papers differs.

However, this delineation does not capture the full range of possible regimes that could contribute to differences in learning and fear extinction, especially in extinction tasks with both safe (CS-) and dangerous (CS+) stimuli. Here, I propose that inferred causal structures lie on a spectrum between too few and too many latent cause assignments, corresponding to "discrimination" failures (failures in distinguishing between safe and dangerous stimuli) and "over-accommodation" failures (failures in updating beliefs about stimuli that are no longer dangerous) respectively, with "good" extinction falling in between. Using simulations, I show that these extremes can be reached through changes in the observation prior parameter, shown in Figure 3 below. My proposed delineation of latent cause regimes clarifies the mapping between erroneous inference of causal structure and different types of extinction failures. Moreover, I relate these regimes to a single parameter, the observation prior, that can reflect beliefs in the

controllability of outcomes, possibly modulated by an individual's past experiences (for more on the connection to past life experience, see section 4).[2]

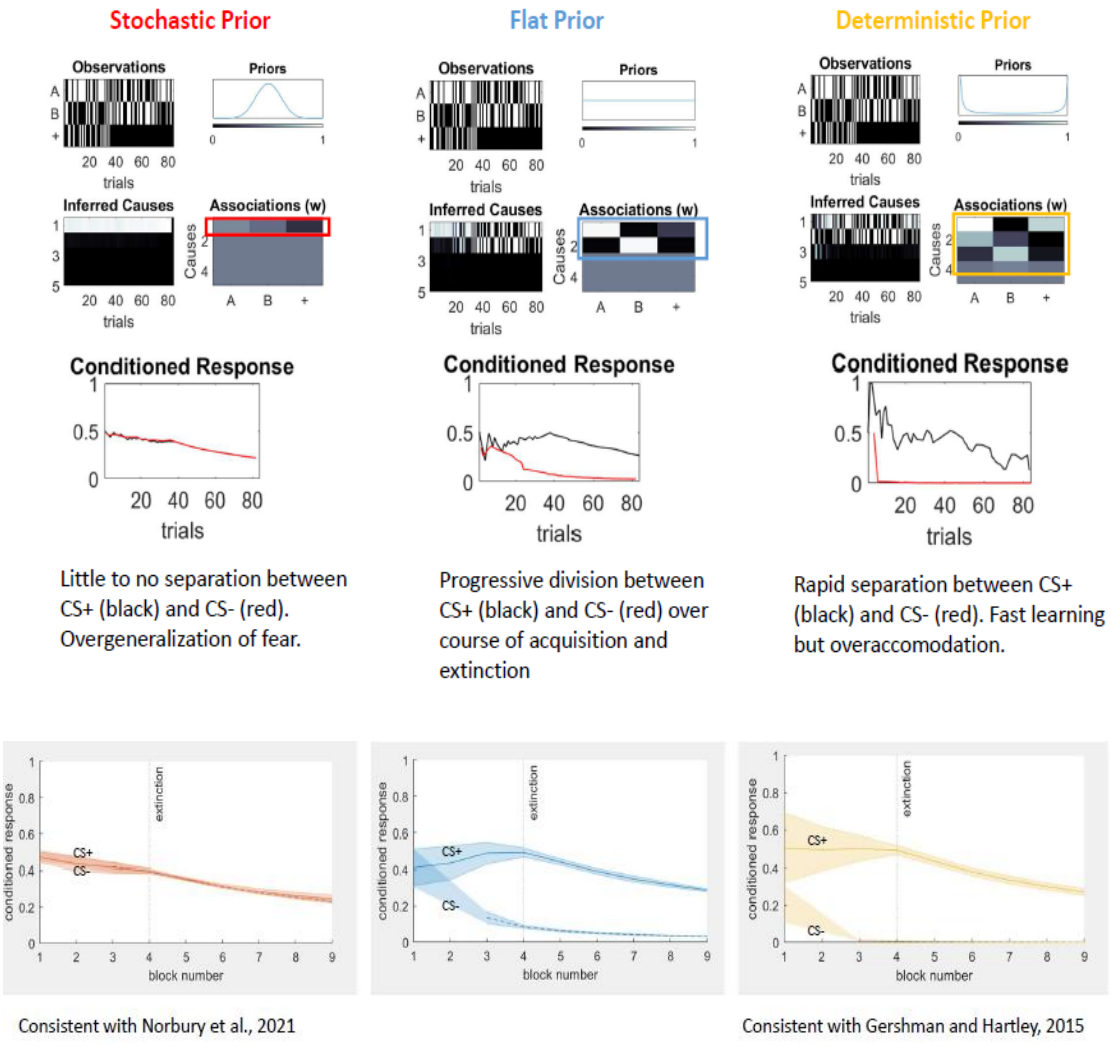


**Figure 3.** Simulation of Hartley et al. 2014 experimental set-up results show three clear learning regimes, consistent with extinction failures found by Norbury et al. and Gershman and Hartley as well as optimal extinction (wherein optimal is defined by similarity of the inferred latent structure to the true structure of the task). 100 "participants" were simulated in each observation prior category for 36 trials of acquisition and 48 trials of extinction (for the stimulus set-up, see Hartley et al. 2014).

I argue that over-and under-segmentations of latent causes best explains the different forms of extinction failure possible in fear conditioning. Furthermore, this

[2] This section contains text and figures based closely on, or identical to, my presented poster at SfN 2021 (Aitsahalia, 2021).

model provides a way to explain the intriguing stressor-controllability effects on fear extinction failure found by Hartley et al., which show an "Instrumental-to-Pavlovian" transfer that can be modeled by a change in the observation prior parameter, which could be modulated by perceived control (or lack thereof) (2014). Considering psychological control computationally, as explained by Huys and Dayan (2009), control can be formulated as the achievability of outcomes given actions. People may have different priors over this mapping, ranging from the belief that outcomes and actions have no relationships to the belief that particular actions always lead to particular outcomes. These priors can change due to reinforcement and the observation of outcomes in relation to actions, and may be different for different domains (Huys & Dayan, 2009). However, these beliefs may also influence learning in situations where no actions are taken. I propose that controllability affects the assignment of observations to latent causes in future Pavlovian learning by influencing observation priors. Beliefs that the world is uncontrollable, or that one's actions have no effect over the reward or punishment seen, can generalize to beliefs that unknown causes generate observations randomly (i.e. a stochastic observation prior), which would lead to an over-assignment of disparate observations to a single cause, or under-segmentation.

Based on model simulations, the conditioned response to the CS- starting in acquisition can be used as a proxy for the learning regime and therefore observation prior, with higher values of the CS- indicating a more stochastic prior and lower values denoting a more deterministic prior. Based on this model, the expectancy of the CS- predicts not only the CS+ in acquisition, but also extinction and spontaneous recovery. If people were behaving according to this model, those with higher expectancies for the CS- in acquisition would not have higher expectancy ratings overall, but rather lower separation between the two stimuli in both acquisition and extinction, indicating their overgeneralization. This model would also predict that people falling into this regime would assume more threat for novel stimuli, especially those that share some degree of similarity with the CS+. On the other hand, people with deterministic observation priors would be more likely to rapidly separate the CSs and not ascribe danger to novel stimuli. However, this rapid learning protects the original CS+ association with the US, meaning that we expect to see higher rates of spontaneous recovery of fear and faster relearning of

the association as well, since the association does not need to be retrained, rather reactivated (Pisupati, 2021).

The addition of persistence or distance-dependence (Figure 4) to the model adds key predictions for behavior in day 2 of the task: that too deterministic a prior would lead to higher rates of spontaneous recovery.

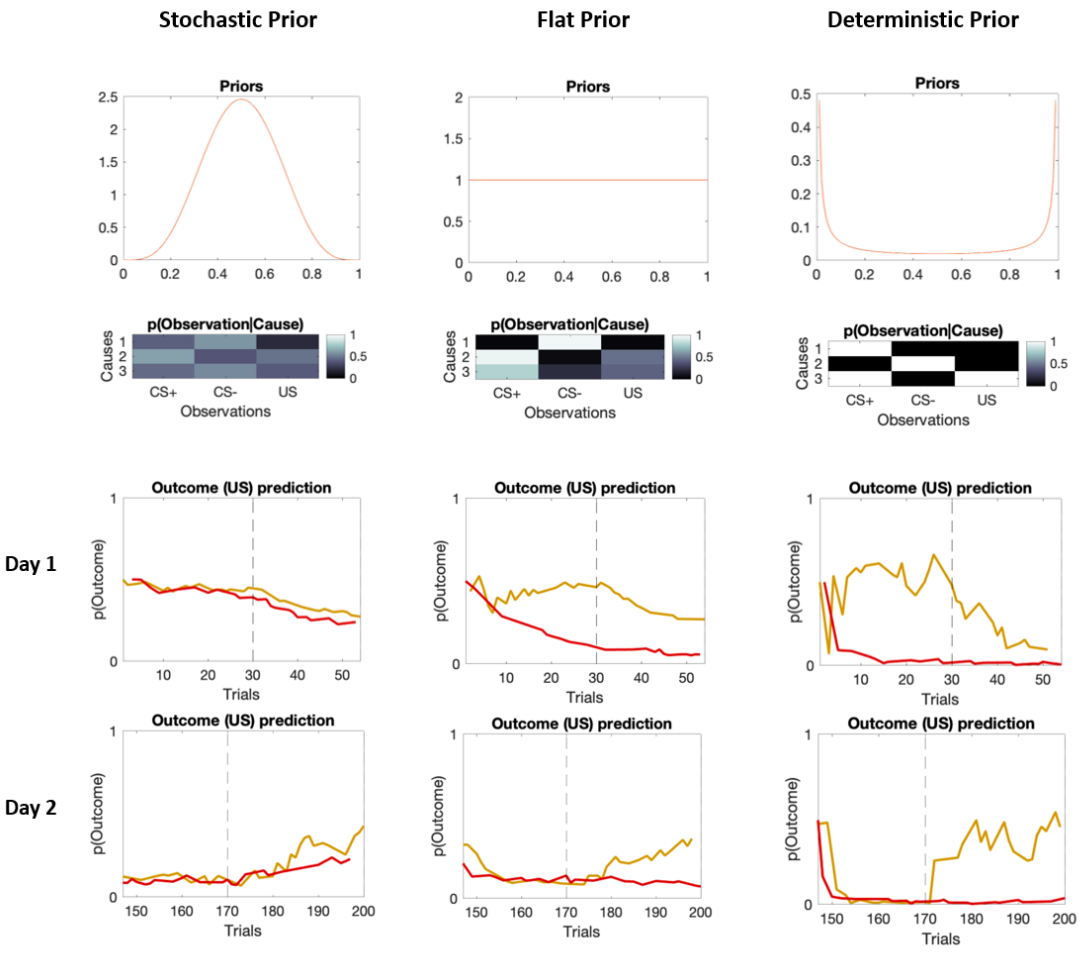


**Figure 4.** Simulation of performance on modified online task across all three regimes. Key predictions include that stochastic priors lead to slower extinction (Day 1), slower relearning (Day 2), and lower spontaneous recovery (Day 2) when compared to flat or deterministic priors.

## 3. Experiments

This section describes the central experiment presented in this thesis: a virtual differential fear conditioning task using a loud scream as an unconditioned stimulus to

test the model predictions described above.[3] This experiment allows for fear conditioning on humans to be conducted online and without the use of electric shock, and also captures the wide range of individual differences in fear extinction in the general population.

### 3.1 Experiment 1: Individual Differences in Differential Fear Conditioning

The described experiments were approved under IRB 11968 Behavioral Studies of Learning and Decision-making. All experimental code was written and served using NivTurk software, python, and jsPsych.

#### 3.1.1 Methods

Online participants completed two tasks over two days. The first task was a classic fear conditioning and extinction task, using a loud scream as the US with two different stimuli. Second, at least one day later, participants were recalled to test their spontaneous recovery of fear. Across both Pavlovian tasks, participants were asked to rate their expectancy of the scream as a behavioral readout of their predictions.

Participants were found and screened through Prolific. 53 participants participated in two parts of the experiment over two days (with 71 participating solely in day 1), not including the 8 participants excluded from the sample (6 due to data saving errors using Prolific, and 2 due to a failure to complete the audio check). All participants provided informed consent.

Day 1: Fear Conditioning, Generalization, and Extinction

This differential fear conditioning task was modified from Hartley et al. (2014) to a shortened online format and different abstract colored stimuli. Participants were instructed to complete the experiment with headphones and were allowed to play the scream (US) to set their audio volume to a level that was considered “unpleasant but not unbearable”. They were then instructed to maintain their computer volume at this level throughout the task. At the start of each trial, participants were shown a fixation cross, which was sometimes shown along an audio instruction to press a particular letter key as

[3] Scream .mp3 and other experimental stimuli are available upon request.

an audio check. Participants were asked to rate the likelihood that the scream would play on a scale of 1-9 after the presentation of each stimulus.

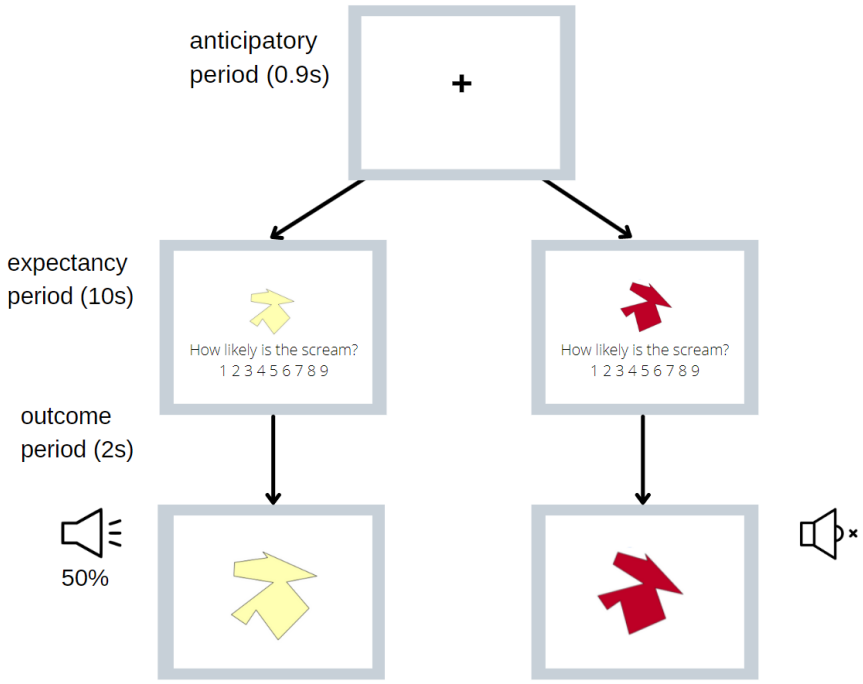


**Figure 5.** Trial structure for fear conditioning. Following a fixation cross, participants saw one of two stimuli in within blocks randomized order and were asked to rate the likelihood of a scream. Then, they were presented with the shape on its own and the accompanying reinforcement, either silence or the scream. Trial structure remains the same during extinction, but without any reinforcement.

Participants completed 30 trials of acquisition, comprised of 12 CS- trials and 18 CS+ trials reinforced at 50% with the loud scream. Participants were then asked to fill out the Anxiety Control Questionnaire (ACQ), a validated self-report measure about perceived control over internal events and threats in the world (Rapee et al., 1996) (See Appendix for questionnaire with attention checks). This was followed by the generalization phase, which consisted of 10 trials: 2 presentations of each of 5 stimuli: the CS+, CS-, and 3 intermediate gradated stimuli, none of which were reinforced. To prevent extinction from occurring during generalization, participants were told that the experiment audio was muted during this phase, and asked to rate how likely it was that the scream was playing while they could not hear it. In the final round of day 1, participants were presented with 24 trials of extinction, half of which were the CS+ and the other half were the CS-.

Day 2: Spontaneous Recovery, Generalization, and Relearning

At least one day after completing Day 1, participants were invited to complete Day 2, which also consisted of 3 rounds. Following a new audio check, participants were shown 24 trials of both CSs without scream, to test spontaneous recovery of fear. Next, they were shown 30 trials of relearning, which were identical to the acquisition trials. The final round was generalization, where they were also instructed that the audio was turned off and shown the same 5 generalization stimuli twice each to assess differences in generalization from day 1. Finally, participants were asked affective questions about their feelings in response to the two main stimuli as well as whether or not they noticed the scream happening more often following one of the shapes than the other.

### 3.1.2 Results

Expectancy ratings, reaction time, audio checks, survey attention checks, and survey responses were recorded from all participants. 2 participants were excluded from the sample due to failing the audio attention check.

First, the average expectancy rating to all presentations of the two differential stimuli were averaged across all participants (Figure 6A, middle). To determine learning regimes, the average expectancy rating for the CS- throughout 12 trials of acquisition was calculated and compared.[4] The 20 participants with the lowest ratings for CS- were considered "deterministic" (shown in Figure 6 on the right) while those 20 with the highest ratings were considered to have "stochastic" priors (shown on the left). As predicted, those participants with a higher CS- rating in acquisition did not have simply higher ratings overall, but rather showed less separation between conditioned stimuli (Figure 6A, left). These stochastic prior participants also showed higher ratings for intermediate novel stimuli in generalization, including stimuli that closely resembled the CS- (Figure 6B, left), showing a higher generalization of fear than participants with deterministic priors.

The two groups, separated based solely on CS- expectancy rating in acquisition, had statistically significant differences in CS+ expectancy in extinction (mean CS+ lower

[4] All analyses were performed using MatLab R2019b. Code is available upon request.

in extinction for deterministic group, $p<0.01$), generalization on novel intermediate stimuli (mean rating for all intermediate stimuli lower for deterministic group, $p<0.01$), and spontaneous recovery (mean starting value for CS+ higher for deterministic group, $p<0.05$) and relearning (mean CS+ value higher for deterministic group, $p<0.05$) in day 2.

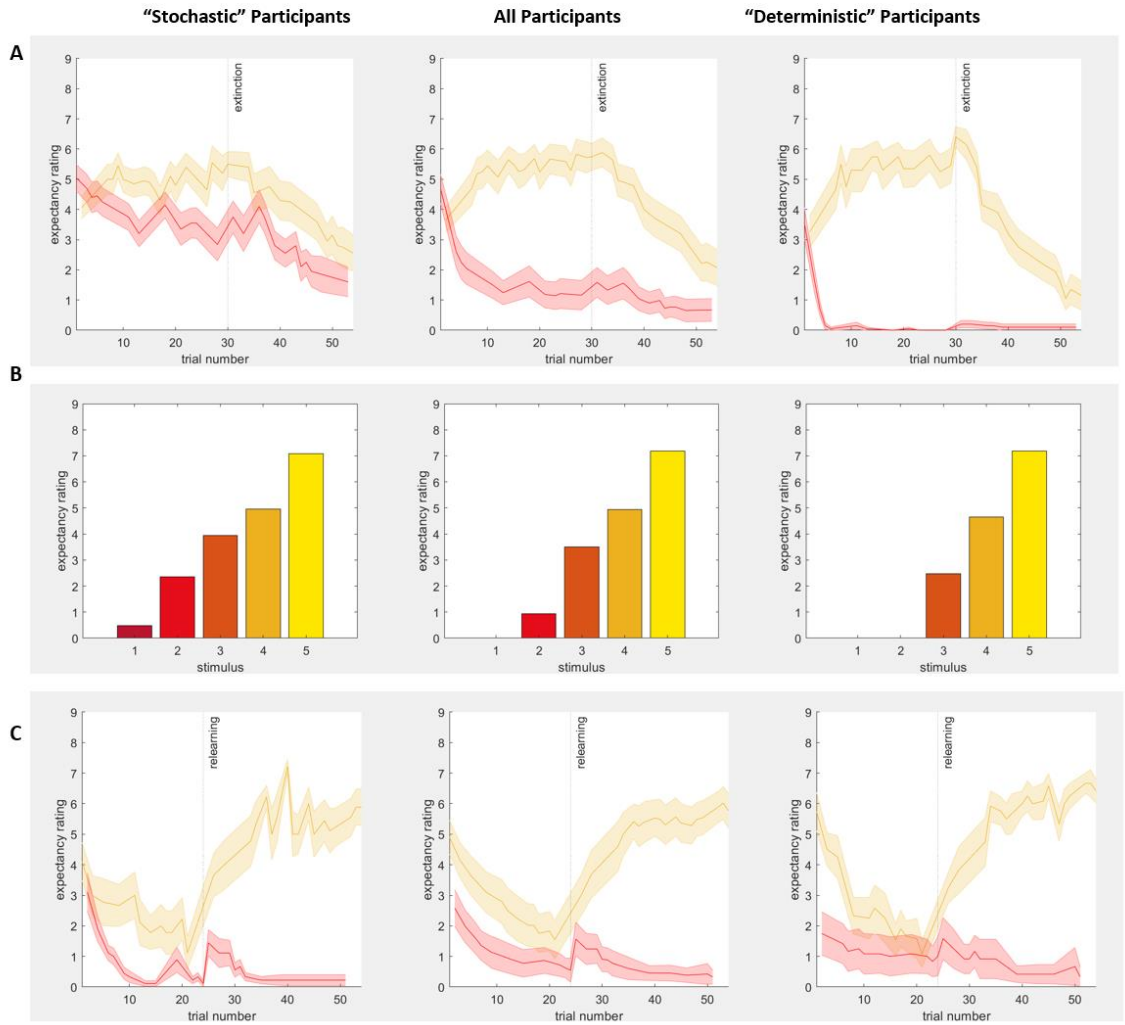


**Figure 6.** Expectancy Ratings over learning in day 1, generalization, and day 2 across regimes, with stochastic priors shown on the left (N=20), the average of all participants in the middle (N=71 for day 1, N=53 for day 2), and deterministic priors (N=20) shown on the right. The yellow lines correspond to the CS+ (the yellow stimulus) while the red lines correspond to the CS-. 95% confidence interval of SEM is shown shaded around mean learning curves. Stimuli 1-5 in panel B are color coded to represent the 5 generalization stimuli.

These results align well with the predicted model regimes, suggesting strong evidence that this task can be used to assess model parameters. The key features,

including separation between CSs and rate of extinction, spontaneous recovery, and relearning are shown as hypothesized in the behavioral data.

Trying to separate regimes based on CS- value (rather than number of participants), we find that more participants endorse an average CS- value of below 2 (N=40) than anything above 4 (N=7), suggesting that more people in this sample fit into the deterministic rather than stochastic prior regime, which is in line with other researchers' findings that people tend to have more deterministic priors over casual structures in their everyday life (Griffiths & Tenenbaum, 2009; Schulz & Sommerville, 2006).

However, the cause of the individual differences in priors is unclear. This experiment was tested in the general population, and it is possible that participants who either have fear-based disorders or endorse subclinical traits relating to them might be more likely to fall into the stochastic category, since feelings of a lack of control or helplessness are commonly reported by patients with psychiatric disorders (Abramson et al., 1989; Ross et al., 1999).

To test this, correlations between responses on the ACQ and learning regime were calculated. Statistical tests showed no significant difference between the mean ACQ score for participants in the stochastic versus deterministic category (however, it seems there is a trend towards conflating CSs among the participants who endorsed the least control, see Appendix). However, this lack of difference could be explained by a lack of specificity in the domain of control. While the ACQ focuses on control over internal events like emotions as well as some external negative events (Rapee et al., 1996), screaming shapes are unlikely to be encountered in everyday life. It is possible that the confines of the experiment are too narrow for a prior over general controllability to apply, with the observation prior captured in this experiment to be one of control over behavioral experiments, screams, video games, or a different specific domain.

### 3.2 Experiment 2: Controllability Manipulation

To see if controllability over a specific outcome can influence observation priors over that same outcome, we are currently piloting a controllability manipulation. Hartley

et al. found that exposure to inescapable stress (via electric shock) in a maze task prior to fear conditioning led to worse extinction of fear (2014), which can be explained in this model by the effect of controllability of shocks on observation priors over the randomness of shocks, influencing future learning involving shocks.

### 3.2.1 Methods

Day 1: Instrumental Controllability, Fear Conditioning

After performing audio tests and calibrating volume for the scream, participants are instructed that they will be travelling through a haunted house and exploring its rooms. They may choose to turn one of three directions: left, back, or right at the start of each trial, and that their choice will lead them into a room, which will either be safe or scary.

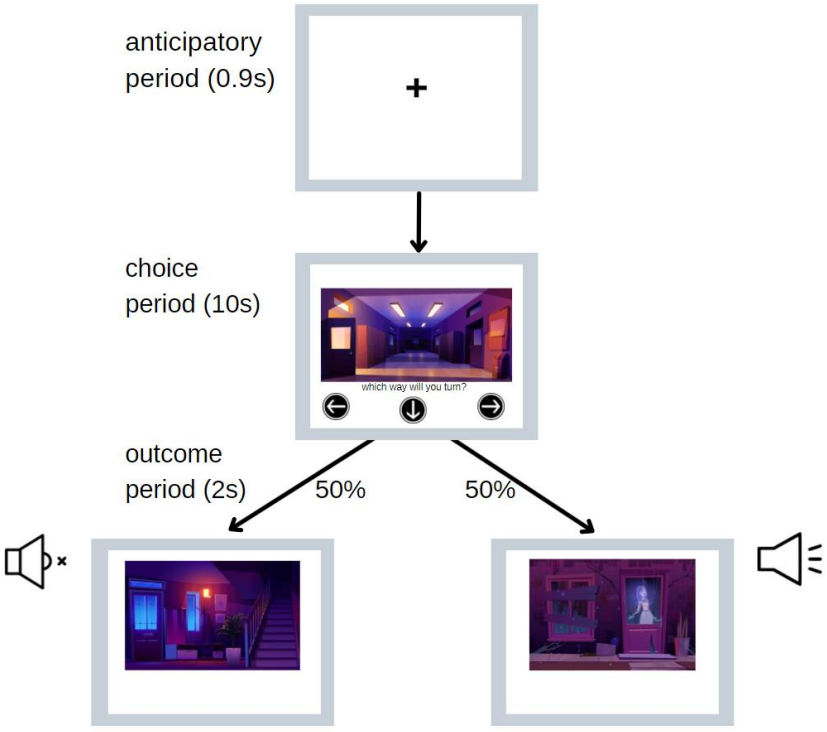


**Figure 7.** Uncontrollable stress manipulation trial structure. After a fixation cross, participants were shown an image of a hallway and asked which way they would like to turn. Three options were presented: left, back, and right. While participants were told they would be shown the room they had turned to, the actual structure of reinforcement was purely probabilistic, with half of trials resulting in the presentation of a calm, empty room and no scream, and the other half resulting in the presentation of a scary room and the loud scream.

We intend to then create a deterministic version of the haunted house manipulation, wherein participants do have instrumental control over the presentation of the scream, to later yoke more participants in the uncontrollable condition to, so that we can ensure any ensuing differences are not a result of differences in number of screams heard but rather control over them. Crucially, the only commonality between the controllability manipulation and the later fear conditioning (aside from the context of being in an online experiment) is the shared US, the scream. This should force participants to develop or adapt a controllability prior over that specific outcome, which will then be measured in the fear conditioning task.

Participants in both conditions will then undergo the fear conditioning and extinction detailed above (section 3.1.1).

Day 2: Spontaneous Recovery, Generalization, and Relearning

At least one day later, participants would return to complete the same second half of the task (detailed above under Day 2 in section 3.1.1). Behavioral data would then exist from three conditions: escapable stress, inescapable stress, and no stress, replicating Hartley et al.'s set-up (2014).

### 3.2.2 Predicted Results

As this task is currently being piloted, there is not a large enough sample size to report any significant results, but the presented model makes strong, falsifiable claims about performance in both days of the experiment. We predict, based on the model and previous results, that repeated lack of control over the presentation of the scream will push participants towards a belief in the stochasticity of the scream, and therefore worsen extinction (Hartley et al., 2014). On the other hand, participants exposed to controllable stress would be biased towards a more deterministic prior and their learning would align with that regime. Those results may not be particularly different from the non-manipulated control group, since we already see a bias towards deterministic priors in the general population (Figure 6A, middle).

One aspect of Hartley et al.'s results not captured in this simple prediction is that participants in the inescapable stress condition showed higher rates of spontaneous recovery (Hartley et al., 2014), which is inconsistent with the simulations that show lower rates of spontaneous recovery with stochastic priors. However, with stochastic enough a prior, it may take far more than the allotted number of extinction trials to fully extinguish the association, so the spontaneous recovery observed may be a direct continuation of the conditioned response. Further research and modifications to the model would need to be added to account for participants who behave in the stochastic regime in day 1 but seem to shift in day 2. It is possible that lack of control also changes learning rates asymmetrically for negative events, or that another unknown time effect is taking place.

## 4. Discussion

This thesis presents a new formalization of individual differences in fear learning as related to observation and controllability priors over outcomes. It draws from evidence in psychology, neuroscience, statistics, control theory, and new preliminary experimental results.

Considering control beliefs computationally, as a prior over the randomness of outcomes (Huys & Dayan, 2009), this research contributes to the field's understanding of fear generalization and the mechanisms by which persistent, generalized fear may emerge and be maintained in a single, theoretically unified model.

The experiment presented does have limitations, including the reliance on self-report both for control beliefs and expectancy. Being explicitly asked for expectancy ratings may shape learning, as has been shown in reversal learning tasks (Atlas et al., 2022). Future work could use electro-dermal or neuroimaging measures to compare with explicitly reported prediction, as well as use affective ratings to determine conditioned response. Despite being used in many human studies of fear conditioning, skin conductance response (SCR) does not seem to be directly linked to expectancy, with many researchers using differential SCR measures without clear rationale for what this readout corresponds to (Gershman & Hartley, 2015). Future work using this task could combine expectancy ratings with SCR to elucidate its relationship to model parameters.

The presented model could also be further developed with a more normative account for time effects, developing a clearer hierarchical Bayesian structure. The learning regimes described in this thesis were achieved via only one parameter, the observation prior; however, adding parameters will likely lead to a better fitting model for individual participants.

This work prompts many clinically relevant and basic science questions. For example, how far does fear generalize and how does this vary across people? In this proposed controllability manipulation, the specific US is the only constant between the instrumental and Pavlovian tasks. Would fear generalize using a different but similar US, for example, a different scream? What about an unrelated noise, or a shock? While previous results show the negative impacts of lack of control in a Pavlovian setting (Hartley et al., 2014), it would be important to test these effects in instrumental conditioning settings. PTSD has been shown to push people towards a stochastic regime (Norbury et al., 2021), but it is possible that other psychiatric disorders, such as OCD with magical thinking (Einstein & Menzies, 2004), may push people towards too deterministic a regime. It is also possible that lack of control in one domain prompts attempts to gain control in other areas, acting as a form of "Instrumental-to-Instrumental" transfer. Eating disorders, for example, have been described by patients as ways to control their body in an uncontrollable world (Froreich et al., 2016). Future work using this latent cause segmentation framework could model eating disorder patients' controllability priors and compare them with other clinical and neurotypical groups.

Understanding the role of external control over outcomes in psychiatric disorders can also help both neuroscientists and clinicians reduce stigma around mental illness.[5] Trauma-informed care seeks to ground patients' behavior in their life context, (Purkey et al., 2018) and this thesis contributes to that framework. By focusing on generalized fear as an adaptive response to an uncontrollable situation being carried into situations where the behavior is no longer adaptive, this research allows for the understanding of neurodivergent cognition as rational rather than defective. Most current theories in psychiatry and cognitive neuroscience describe the thoughts and behaviors in anxiety

[5] This section contains text that is based closely on, or is identical to, text found in my Junior Paper.

disorders as entirely biological (Friedman, 2007; Garcia De Miguel et al., 2012; Thorsell, 2010) or, when including cognition, as simply maladaptive (Calvete et al., 2013; Mahoney et al., 2018). However, much of this research does not account for or address the life history of patients. Not only does this lead to patients feeling ignored or invalidated (Rashed, 2019; Szasz, 1994), it also overlooks a potentially central mechanism by which these disorders develop. This computational and theoretical thesis aims to shift the focus from purely biological dysregulation to a more nuanced view that factors in the generalization of environmentally-appropriate stress responses to exaggerated fear. Drawing from the field of Disability studies, there is already language for understanding psychiatric disorders in a trauma-informed and humanistic way, focusing on how the behaviors arise and the agency of the patient (Amundson & Tresky, 2007; Oliver, 1990; Rashed, 2019).

## References


Abramson, L. Y., Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness depression: A theory-based subtype of depression. *Psychological Review*, *96*(2), 358–372. https://doi.org/10.1037/0033-295X.96.2.358

Aitsahalia, I. (2021, November 9). *Over- and Under-Segmentation of Latent Causes Lead to Different Types of Extinction Failures* [Poster]. Society for Neuroscience.

Amundson, R., & Tresky, S. (2007). On a bioethical challenge to disability rights. *The Journal of Medicine and Philosophy*, *32*(6), 541–561. https://doi.org/10.1080/03605310701680924

Atlas, L. Y., Sandman, C. F., & Phelps, E. A. (2022). Rating expectations can slow aversive reversal learning. *Psychophysiology*, *59*(3), e13979. https://doi.org/10.1111/psyp.13979

Blei, D. M., & Frazier, P. I. (2011). Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research*, *12*(74), 2461–2488.

Calvete, E., Orue, I., & Hankin, B. L. (2013). Early maladaptive schemas and social anxiety in adolescents: The mediating role of anxious automatic thoughts. *Journal of Anxiety Disorders*, *27*(3), 278–288. https://doi.org/10.1016/j.janxdis.2013.02.011

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

Einstein, D. A., & Menzies, R. G. (2004). Role of magical thinking in obsessive-compulsive symptoms in an undergraduate sample. *Depression and Anxiety*, *19*(3), 174–179. https://doi.org/10.1002/da.20005

Famularo, R., Kinscherff, R., & Fenton, T. (1992). Psychiatric Diagnoses of Maltreated Children: Preliminary Findings. *Journal of the American Academy of Child & Adolescent Psychiatry*, *31*(5), 863–867. https://doi.org/10.1097/00004583-199209000-00013

Fierman, E. J., Hunt, M. F., Pratt, L. A., Warshaw, M. G., Yonkers, K. A., Peterson, L. G., Epstein-Kaye, T. M., & Norton, H. S. (1993). Trauma and posttraumatic stress disorder in subjects with anxiety disorders. *American Journal of Psychiatry*, *150*(12), 1872–1874. Scopus. https://doi.org/10.1176/ajp.150.12.1872

Friedman, B. H. (2007). An autonomic flexibility–neurovisceral integration model of anxiety and cardiac vagal tone. *Biological Psychology*, *74*(2), 185–199. https://doi.org/10.1016/j.biopsycho.2005.08.009

Froreich, F. V., Vartanian, L. R., Grisham, J. R., & Touyz, S. W. (2016). Dimensions of control and their relation to disordered eating behaviours and obsessive-compulsive symptoms. *Journal of Eating Disorders*, *4*, 14. https://doi.org/10.1186/s40337-016-0104-4

Garcia De Miguel, B., Nutt, D. J., Hood, S. D., & Davies, S. J. C. (2012). Elucidation of neurobiology of anxiety disorders in children through pharmacological challenge tests and cortisol measurements: A systematic review. *Journal of Psychopharmacology (Oxford)*, *26*(4), 431–442. https://doi.org/10.1177/0269881110372818

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., & 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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. https://doi.org/10.1037/a0017201

Hartley, C. A., Gorun, A., Reddan, M. C., Ramirez, F., & Phelps, E. A. (2014). Stressor controllability modulates fear extinction in humans. *Neurobiology of Learning and Memory*, *113*, 149–156. https://doi.org/10.1016/j.nlm.2013.12.003

Harvard Medical School. (2005). *National Comorbidity Survey*. NCS. https://www.hcp.med.harvard.edu/ncs/index.php

Heim, C., & Nemeroff, C. B. (2001). The role of childhood trauma in the neurobiology of mood and anxiety disorders: Preclinical and clinical studies. *Biological Psychiatry*, *49*(12), 1023–1039. https://doi.org/10.1016/S0006-3223(01)01157-X

Huys, Q. J. M., & Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, *113*(3), 314–328. https://doi.org/10.1016/j.cognition.2009.01.008

Lloyd, K., & Leslie, D. S. (2013). Context-dependent decision-making: A simple Bayesian model. *Journal of The Royal Society Interface*, *10*(82), 20130069. https://doi.org/10.1098/rsif.2013.0069

Mahoney, A. E. J., Hobbs, M. J., Newby, J. M., Williams, A. D., & Andrews, G. (2018). Maladaptive Behaviours Associated with Generalized Anxiety Disorder: An Item

Response Theory Analysis. *Behavioural and Cognitive Psychotherapy*, *46*(4), 479–496. https://doi.org/10.1017/S1352465818000127

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 (Cambridge, Mass.)*, *2*, 50–73. https://doi.org/10.1162/CPSY_a_00014

Myers, K. M., & Davis, M. (2007). Mechanisms of fear extinction. *Molecular Psychiatry*, *12*(2), 120–150. https://doi.org/10.1038/sj.mp.4001939

Nair, A., Rutledge, R. B., & Mason, L. (2020). Under the Hood: Using Computational Psychiatry to Make Psychological Therapies More Mechanism-Focused. *Frontiers in Psychiatry*, *11*. https://www.frontiersin.org/article/10.3389/fpsyt.2020.00140

Navarro, D., & Perfors, A. (n.d.). *The Chinese restaurant process*. University of Adelaide.

Norbury, A., Brinkman, H., Kowalchyk, M., Monti, E., Pietrzak, R. H., Schiller, D., & Feder, A. (2021). Latent cause inference during extinction learning in trauma-exposed individuals with and without PTSD. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S0033291721000647

Oliver, M. (1990). *Politics Of Disablement*. Macmillan International Higher Education.

Pisupati, S. (2021, April 30). *Two Factors Underlying Maladaptive Inference of Causal Structure Can Drive Resistance to Extinction in Anxiety*. Society of Biological Psychiatry's 2021 Annual Meeting.

Purkey, E., Patel, R., & Phillips, S. P. (2018). Trauma-informed care. *Canadian Family Physician*, *64*(3), 170–172.

Ramos-Cejudo, J., & Salguero, J. M. (2017). Negative metacognitive beliefs moderate the influence of perceived stress and anxiety in long-term anxiety. *Psychiatry Research*, *250*(Journal Article), 25–29. https://doi.org/10.1016/j.psychres.2017.01.056

Rapee, R. M., Craske, M. G., Brown, T. A., & Barlow, D. H. (1996). Measurement of perceived control over anxiety-related events. *Behavior Therapy*, *27*(2), 279–293. https://doi.org/10.1016/S0005-7894(96)80018-9

Rashed, M. A. (2019). In Defense of Madness: The Problem of Disability. *The Journal of Medicine and Philosophy*, *44*(2), 150–174. https://doi.org/10.1093/jmp/jhy016

Ross, C. E., Sastry, J., Aneshensel, C. S., & Phelan, J. C. (1999). *Handbook of the sociology of mental health*.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427–442. https://doi.org/10.1111/j.1467-8624.2006.00880.x

Speekenbrink, M. (2016). A tutorial on particle filters. *Journal of Mathematical Psychology*, *73*, 140–152. https://doi.org/10.1016/j.jmp.2016.05.006

Szasz, T. (1994). Psychiatric diagnosis, psychiatric power and psychiatric abuse. *Journal of Medical Ethics*, *20*(3), 135–138. https://doi.org/10.1136/jme.20.3.135

Thorsell, A. (2010). Brain neuropeptide Y and corticotropin-releasing hormone in mediating stress and anxiety. *Experimental Biology and Medicine*, *235*(10), 1163–1167. https://doi.org/10.1258/ebm.2010.009331

White, T. L., & Gonsalves, M. A. (2021). Dignity neuroscience: Universal rights are rooted in human brain science. *Annals of the New York Academy of Sciences*, *1505*(1), 40–54. https://doi.org/10.1111/nyas.14670

**Appendix**

**ACQ** (Rapee et al., 1996)**:**

Listed below are a number of statements describing a set of beliefs. Please read each statement carefully and, on the 0-5 scale given, indicate how much you think each statement is typical of you.

| 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Strongly Disagree | Moderately Disagree | Slightly Disagree | Slightly Agree | Moderately Agree | Strongly Agree |

1. I am usually able to avoid threat quite easily.

2. How well I cope with difficult situations depends on whether I have outside help.

3. When I am put under stress, I am likely to lose control.

4. I can usually stop my anxiety from showing.

5. When I am frightened by something, there is generally nothing I can do.

6. My emotions seem to have a life of their own.

7. There is little I can do to influence people's judgements of me.

8. Whether I can successfully escape a frightening situation is always a matter of chance with me.

9. I often shake uncontrollably.

10. I can usually put worrisome thoughts out of my mind easily.

11. When I am in a stressful situation, I am able to stop myself from breathing too hard.

12. I can usually influence the degree to which a situation is potentially threatening to me.

13. I am able to control my level of anxiety.

14. There is little I can do to change frightening events.

15. The extent to which a difficult situation resolves itself has nothing to do with my actions.

16. If something is going to hurt me, it will happen no matter what I do.

17. I can usually relax when I want.

18. When I am under stress, I am not always sure how I will react.

19. I can usually make sure people like me if I work at it.

20. Most events that make me anxious are outside my control.

21. I always know exactly how I will react to difficult situations.

22. I am unconcerned if I become anxious in a difficult situation, because I am confident in my ability to cope with my symptoms.

23. What people think of me is largely outside my control.

24. I usually find it hard to deal with difficult problems.

25. When I hear that someone has a serious illness, I worry that I am next.

26. When I am anxious, I find it difficult to focus on anything other than my anxiety.

27. I am able to cope as effectively with unexpected anxiety as I am with anxiety that I expect to occur.

28. I sometimes think, "Why even bother to try to cope with my anxiety when nothing I do seems to affect how frequently or intensely I experience it?".

29. I often have the ability to get along with "difficult" people.

30. I will avoid conflict due to my inability to successfully resolve it.

Attention checks (presented in randomized locations throughout the questionnaire):

1. I swim across the Atlantic every day to go to work. (Expected answer: 0)
2. I can hold my breath for at least one second. (Expected answer: 5)

## Supplementary Figure 1: Lowest ACQ Scorers Learning