The neuro-computational role of uncertainty in anxiety

Alexandra Kathryn Hopkins

A thesis presented for the degree of

PhD in Computational Psychiatry

Institute of Neurology

University College London

October 2021

Declaration

I, Alexandra Kathryn Hopkins, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Alexandra Kathryn Hopkins

"The only thing that makes life possible is permanent, intolerable uncertainty: not knowing what comes next." Ursula K. Le Guin, The Left Hand of Darkness

Abstract

Anxiety disorders are the most common mental health disorders and comprise a large number of years lost to disability. The work in this thesis is oriented towards understanding anxiety using a computational approach, focusing on uncertainty estimation as a key process. Chapter 1 introduces the role of uncertainty within anxiety and motivates the subsequent experimental chapters. Chapter 2 is a review of the computational role of the amygdala in humans, a key area for uncertainty computation. Chapter 3 is an experimental chapter which aimed to address gaps in the literature highlighted in the preceding chapters, namely the link between sensory uncertainty processing and anxiety and the role of the amygdala in this process. This chapter focuses on the development of a novel computational hierarchical Bayesian model to quantify sensory uncertainty and its application to neuroimaging data, with intolerance of uncertainty relating to greater neural activation in the insula but not amygdala. Chapter 4 targets the computational mechanisms underlying the negative self-bias observed in sub-clinical social anxiety. Again, this chapter focuses on the development of novel computational belief-update models which explicitly model uncertainty. Here, we see that a reduced trait self-positivity underpins this negative social evaluation process. The final experimental chapter presented in Chapter 5 investigates the link between different computational mechanisms, such as uncertainty, and a range of mood and anxiety symptomatology. This study revealed cognitive, social and somatic computational profiles that share a threat bias mechanism but have distinct negative-self bias and aversive learning signatures. Contrary to expectations, none of the uncertainty measures showed any associations with anxiety symptom subtypes. Finally, chapter 6 brings together the work in this thesis and alongside limitations of the work, discusses how these experiments contribute to our understanding of anxiety and the role of uncertainty across the anxiety spectrum.

Impact Statement

The primary focus of this thesis was on a unique exploration of the cognitive biases present in anxiety. The novel computational work in this thesis has implications for the study of anxiety, both within academia and to the development of clinical theory outside of academia. Given that anxiety disorders are so commonly diagnosed and create large amounts of suffering, it is especially important to understand the pathophysiological and computational pathways that lead to and maintain these disorders. The results of these studies provide important insight into the computational mechanisms that are important for understanding different forms of anxiety.

The work presented in Chapter 2 synthesises a large body of literature and provides directions for future research as well as hypothesis generating ideas for the development of new experimental studies.

The work presented in Chapter 3 uses a novel paradigm and includes the development of a novel hierarchical Bayesian model, which quantifies sensory uncertainty. This paradigm and model can be used by other researchers in this field and could be extended across other sensory modalities.

The work presented in Chapter 4 on sub-clinical social anxiety provides candidate computational mechanisms that may be useful for assessing outcomes of therapeutical interventions. This work has already been published in an open-access peer-reviewed journal (Hopkins et al., 2021), as well as code materials on Github. The models developed in this work will also be used as part of a mega-analysis developed by our co-author Kate Button at Bath University (see Open Science Framework https://osf.io/hx6p9/ for pre-registration) and will therefore contribute to the understanding of social evaluation across multiple studies.

The work presented in Chapter 5 makes use of a novel method for developing a computational phenotype of anxiety, which is to combine multiple tasks and computational models that assay anxiety related constructs within the same individuals. This work makes an important contribution to computational psychiatry research, as this method could be utilised by researchers interested in other mental health problems in order to understand transdiagnostic mechanisms and identify potential computational targets for intervention studies.

This thesis generally provides an important contribution to the field of computational psychiatry, with the development of novel computational models that explicitly quantify uncertainty. The models developed over the experimental chapters can be used in further research to understand learning and decision-making processes and could be applied to understanding other psychiatric disorders, such as depression. The threat perception model may also be useful for understanding sensory processing differences in people with autism or PTSD.

Furthermore, as this thesis was oriented towards understanding anxiety, this work may be important for the clinical world outside of academia. The key identified computational parameters that distinguish high from low anxiety may be important targets for therapeutical interventions. It also moves the field closer to precision psychiatry, by enabling better understanding of computational profiles, such as those identified in Chapter 5, which may require different treatment pathways.

Acknowledgements

The work in this thesis would not have been possible without the support of amazing friends, family and colleagues. It really does take a village to raise a child and my village was full of incredible people, for whom I am infinitely grateful. Firstly, in acknowledgement of the academic knowledge and environment that I have greatly benefitted from, I would like to thank my supervisors Michael Moutoussis, Peter Zeidman and Ray Dolan. This is also a good place to express my extreme gratitude to all of the support staff at the FIL and MPC, who literally make the world go around. To Megan, Clive, Kurt, David, and Al, thank you for making the countless hours spent in the basement cake-filled and interesting, for everything you taught me about MRI and for being so patient and supportive. Thank you to Monica, Kamlyn, Toyah and Maddie for being absolute angels who were always so helpful and kind.

It is not an exaggeration to say that I could not have done this without the worlds greatest best friend, (Dr) Rachel Bedder. It has been a wild ride with you, over the past 4+ years. Who would have thought that running around Zurich in the middle of a thunderstorm could bond two people so closely. If I had to choose the best thing about my PhD, I'd say you.

To my kind and supportive colleagues, especially my work wife, Geert-Jan, who made the office such a bright and fun place to be. Jochen, you always were the life and soul of the place and your encouragement helped me a lot. Tobias, thank you for always supporting me and for being a great role model, it has been special to watch you become such an amazing supervisor and leader. Ben, it was wonderful to talk about the universe with you during our long evenings spent at the scanner. Magda, you make being a scientist look easy and your hard work and motivation is inspiring and somehow contagious. Jolanda, Alisa and Yuki, you all made the best PhD group I could have wished for. Matilde, Andrea, Marion, Elisa and Nour, it was great to share an office with you over the years. Toby Wise, you have the

dual honor of being the best road trip buddy and best collaborator and I'm excited to watch you grow as a scientist, I know you'll do amazing things. I will also think fondly of DEATH reading group with Jess, Paul and Evan, where a lot of my thinking around anxiety was formed.

Spending almost two years of my PhD in the middle of a global pandemic was a strange and disorienting experience. I was lucky to live with 5 amazing and supportive flatmates who made lockdown bearable and sometimes even fun. Thank you for the amazing food we shared, days spent sunbathing in the garden and dancing to loud music together in an empty house. I also want to thank Karen for being such an amazing pandemic (and beyond) friend. As always, thank you to the wolfpack, En, Jo, Lucy, Sophie and Tom, you are truly wonderful friends and I am so incredibly lucky to have you in my life.

I would also like to thank my incredibly supportive family. It has been a difficult few years and a lot of time spent apart, but you believed in me all the way. Mum, I hope reading this makes you as proud of me as I am of you. AB, I know you are watching over me, smiling. Grandma and Uncle Steve, thank you for always being there to support me, and for sending me letters, it means the world.

Lastly, to Alice Liefgreen, my hero - all my love, always *(across all existences)*.

Have patience with everything unresolved in your heart and to try to love the questions themselves as if they were locked rooms or books written in a very foreign language. Don't search for the answers, which could not be given to you now, because you would not be able to live them. And the point is to live everything. Live the questions now. Perhaps then, someday far in the future, you will gradually, without even noticing it, live your way into the answer. - Rainer Maria Rilke

Contents

List of Figures

List of Tables

Abbreviations

ACC	Anterior cingulate cortex
ADHD	Attention deficit hyperactivity disorder
AIC	Akaike information criterion
AMY	Amygdala
ANX	Anxiety
AL	Associative learning
BDI	Beck Depression Inventory
BIC	Bayesian information criterion
BOLD	Blood oxygen-level dependent
BU	Belief-update
DDM	Drift diffusion model
DSM	Diagnostic and Statistical Manual of Mental Disorders
EPI	Echo planar imaging
FA	Factor analysis
FLASH	Fast low angle shot
fMRI	Functional magnetic resonance imaging
FNE	Fear of negative evaluation
FWE	Family-wise error
FWHM	Full width at half maximum
GAD	Generalised Anxiety Disorder
GLM	General linear model
HA	High anxiety
HGF	Hierarchical Gaussian Filter
HRF	Haemodynamic response function
IUS	Intolerance of uncertainty
LA	Low anxiety
LOO	Leave-one-out cross validation
LR	Learning rate
MCMC	Markov chain monte carlo
MLE	Maximum likelihood estimation
MPRAGE	Magnetisation-prepared rapid-acqusition gradient echo
OCD	Obsessive compulsive disorder
OFC	Orbitofrontal cortex
PCA	Principal components analysis
PE	Prediction error
PH	Pearce-Hall
PTSD	Post-traumatic stress disorder
RDoC	Research domain criteria
RL	Reinforcement learning
ROI	Region of interest
RT	Reaction time
RW	Rescorla-Wagner
SA	Social anxiety
SD	Standard deviation
SDT	Signal detection theory
STAI	Spielberger State-Trait Anxiety Inventory
TE	Echo time
TOS	Threat of shock
TR	Repetition time
VDM	Value decision making
vmPFC	Ventromedial prefrontal cortex

"Anxiety may be compared with dizziness. He whose eye happens to look down the yawning abyss becomes dizzy." Kierkegaard

1 Chapter 1: General Introduction

1.1 Theoretical and Empirical Thesis Work

The work presented in this thesis is an examination of the role of uncertainty in anxiety. I use a combination of computational modelling, neuroimaging and large-scale population studies in order to investigate cognitive biases found in anxiety, paying special attention to the role of uncertainty as a key mechanism. Throughout, I use novel computational models that explicitly quantify uncertainty in order to understand the dynamics and integration of uncertainty in decision making. I first conduct a review of a key brain region thought to be heavily implicated in uncertainty processing, the amygdala. I focus this review on human computational studies in order to spotlight the complex computations the amygdala performs and examine its relation to psychopathology.

The first empirical study of my PhD builds upon the evidence that the amygdala plays an important role in uncertainty processing, as suggested by my review. This study aimed to understand how sensory uncertainty is related to subjective threat perception and whether individuals who find uncertainty aversive process sensory uncertainty differently. The second empirical study of my PhD compared models of uncertainty (belief-updating models) with models of associative learning for understanding the negative self-bias observed in people who are highly fearful of negative evaluation, a key symptom of social anxiety. This study aimed to determine the roles of these two important psychological frameworks in self-evaluation, and thus provide a more holistic understanding of how individuals learn from social feedback over time. The final empirical study of my PhD investigated the relationship between anxiety symptomatology and four different cognitive biases prominent in anxiety, as well as model derived measures of uncertainty. By assessing different computational

processes within the same individuals, it allowed the investigation of the interrelation between them, as well as the relationship between key parameters and subdimensions of anxiety.

Overview of Chapter 1 The first Chapter of my thesis aims to provide a background and motivation for the key questions investigated in this thesis. I firstly introduce the concept of uncertainty and how such a broad concept can be broken down into different but interrelated forms. I then introduce the key method used in my thesis, computational modelling and highlight how computational modelling can make uncertainty estimation explicit, allowing it to be measured. I discuss how individual differences impacts preferences and decisions made under uncertainty and spotlight the prominence of uncertainty aversion within anxiety. Finally I review the literature on neural representations of uncertainty, which leads onto the review of the amygdala in Chapter 2.

1.2 Uncertainty everywhere

'Our new Constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes.' — Benjamin Franklin, in a letter to Jean-Baptiste Leroy, 1789

Uncertainty is ubiquitous in nature and as the famous quote by Benjamin Franklin suggests, there are very few situations that are fully predictable. Indeed, even death carries a form of temporal uncertainty - we know it is coming, but we likely don't know when. This uncertainty may reduce as we age, depending on many factors such as our expectations about how long we think we will live, whether we are healthy or sick, or whether we are an optimist or pessimist. As uncertainty is the rule rather than the exception, it is likely that organisms have evolved to operate under conditions of uncertainty, integrating it into their learning and decision-making process in order to make effective decisions within the world.

Indeed, throughout the animal kingdom, evidence of uncertainty integration has been demonstrated in a myriad of ways. A striking example is found in honey bees who, in a visual discrimination task, were shown to opt out of trials with less information, thus showing a sensitivity to the amount of uncertainty within the environment and directly incorporating this sensitivity into their decision making process (Perry and Barron, 2013). Flexible adaptations to uncertainty have also been demonstrated within unicellular organisms, such as bacteria, who actually modify their phenotypic presentation in conjunction with their expectations about future changes to their environment (Arnoldini et al., 2012). The phenotypic expression of the bacteria, then, is not fixed, but is *probabilistic* and is in close dynamic relationship

with the environment, rendering uncertainty a crucial signal for which expression to manifest.

Given the potentially infinite amount of information that the organism is bombarded with, it is difficult to imagine how integrating all this information is achieved in such a seamless manner. Not only are there a slew of signals from the environment, there are also internal signals that the body is constantly relaying to the brain, a crucial part of homeostasis. Thus, a key challenge faced by an organism in an uncertain environment is how to balance the many different kinds of uncertainties that rarely exist in isolation (Bach and Dolan, 2012). I now turn to discuss the many different forms of uncertainty found in the environment, their interrelations and how this uncertainty can be formally quantified, using computational modelling.

1.2.1 Disambiguating uncertainty: The uncertainty hierarchy

Uncertainty is multi-faceted and theoretical work has sought to decompose this vast concept into a number of intersecting, yet distinct forms (Bach and Dolan, 2012; Hart et al., 2014). Figure 1.1 displays a graphical depiction of different levels of uncertainty that are combined to create our everyday conscious experience. Here, exteroception and interoception are two sources of uncertainty signal, either from the external world in the case of exteroception, or internally from the body in the case of interoception. These signals are combined by the brain, rendering our model of the self as an agent in the world. Different forms of uncertainty, can be related to both external and internal signals and different layers within the hierarchy can impact each other (see Box 1 for further explicit definitions).

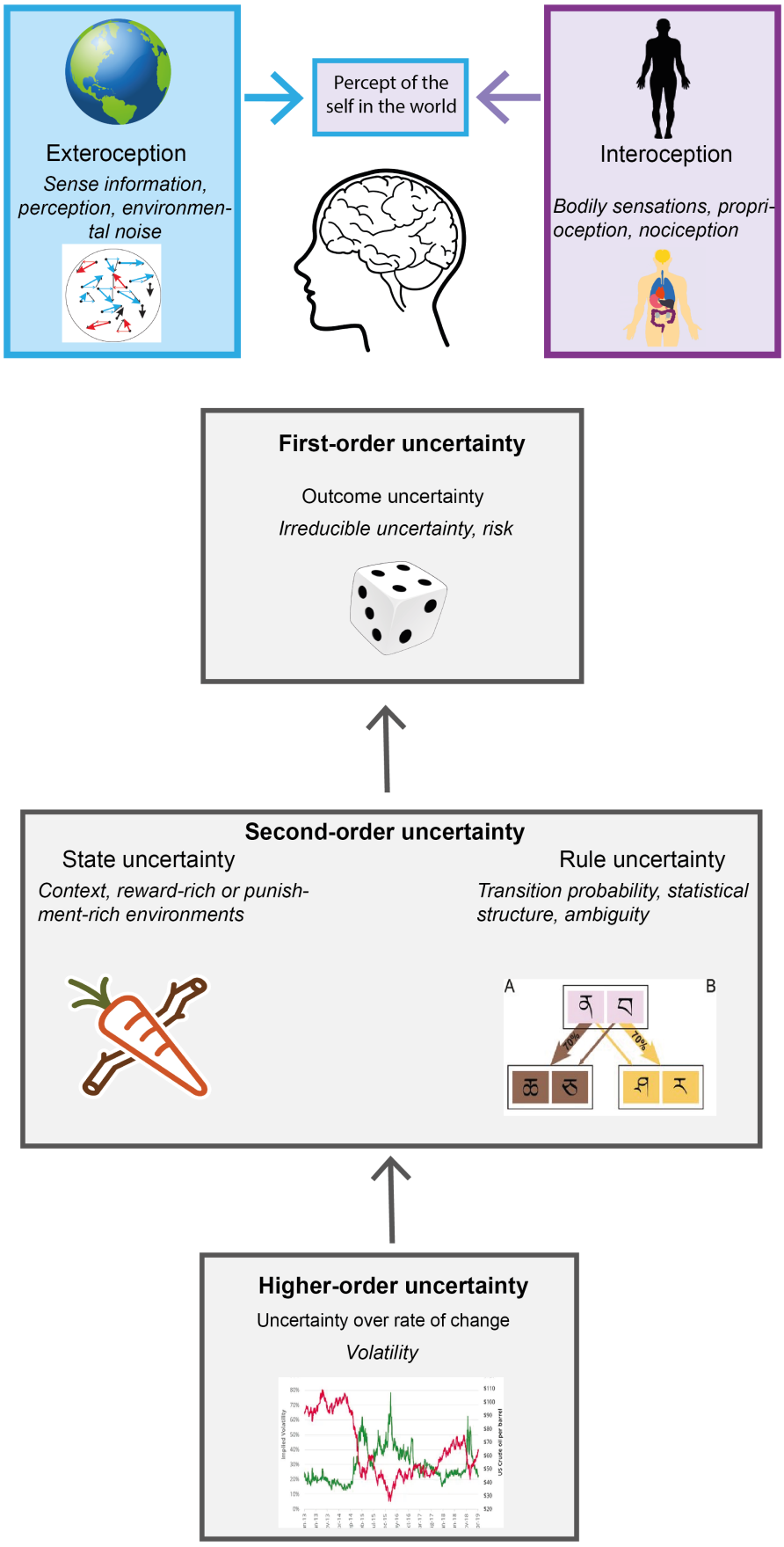

Figure 1.1: Graphical depiction of the uncertainty hierarchy, with forms of uncertainty, external and internal combining to form the percept of the self embedded in the world. In grey are uncertainty variables which external and internal uncertainty can take the form of.

As a conceptual example of this in everyday life, think about riding a bike down a busy road in London. This road is a notoriously dangerous road, with accidents happening at an eye-wateringly high rate due to a faulty traffic light that sometimes displays a green light instead of red. The apologetic maintenance crew can only give road users information about how often this occurs - 20% of the time. The cyclist has the information in advance, however there still exists the inherent uncertainty about which outcome will occur (the traffic light will be working or not). The cyclist has to determine whether the probability of the traffic light not working (and thus potentially leading to a bad outcome) is worth the risk by integrating this kind of first-order uncertainty, or *outcome uncertainty* (see Box 1), into their decision-making. Outcome uncertainty (or risk, in economics) refers to the uncertainty surrounding the probabilistic nature of action-outcome pairings. Even when probabilities are fully known, uncertainty still exists around which outcome will arise and this cannot be reduced by learning, due to the stochastic nature of our environment. This uncertainty varies as a function of the probability of the outcome, following an inverse u-shape curve, with more extreme probabilities providing a higher level of certainty than probabilities closer to 0.5 (Fiorillo et al., 2003). For example, the outcomes of an (unbiased) coin toss are always completely unpredictable and prediction accuracy will fluctuate closely around 50%. Therefore with each new flip of the coin, you will have the same confidence in the outcome as you did for every previous flip of the coin, which is a complete guess. Unless outcome contingencies are fully predictable i.e. P = 0 or P = 1, there will always be some degree of uncertainty associated with event outcomes. Therefore, when discussing probabilistic outcomes, it is necessarily a discussion about uncertainty.

Box 1: Levels of uncertainty

Sensory uncertainty Sensory/perceptual uncertainty is the uncertainty associated with the environment, either from internal noise in the estimation process itself or stimulus uncertainty (Bach and Dolan, 2012). It often studied in the visual and auditory domains using paradigms such as random dot motion tasks, which assess the speed and accuracy in which people can judge the direction of moving dots. Computationally, people integrate sensory uncertainty in a near optimal Bayesian manner (Zhou et al., 2020).

Outcome uncertainty Outcome uncertainty (risk) is a form of *irreducible* uncertainty, as it can never be overcome by learning. There will always be an element of uncertainty over the outcomes due to stochasticity (unless the probabilities are 0 or 1). People are generally risk averse and prefer higher probability options (more certain outcomes) even if they have lower expected value (Kahneman, 1979). Outcome uncertainty in the aversive domain has been linked to physiological markers of stress, such as pupil diameter and SCR (de Berker et al., 2016).

Estimation uncertainty Estimation uncertainty, or ambiguity, is the uncertainty surrounding the mapping between stimulus to outcome. It is *reducible*, as it can be fully learned, as long as the statistics are stable and there is information given about the probability mappings. As more stimulus-outcome pairings are observed, estimation uncertainty is reduced towards zero as the contingencies are fully learned (asymptote) (Payzan-LeNestour and Bossaerts, 2011). Ambiguity is commonly studied using probabilistic learning tasks or using 'urns' tasks, where people have to choose between an urn with known values and an urn with unknown values (ambiguous urn). People generally show a preference for non-ambiguous urns (Ellsberg, 1961). Ambiguity can also be studied in the context of sensory information from mixed categories, such as faces that have features of both anger and sadness (Kaminska et al., 2020).

Higher order uncertainty In hierarchical models of uncertainty, there is often assumed to be uncertainty over the estimation uncertainty, known as higher order uncertainty. This is observed when uncertainty changes over time, for example if the stimulus-outcome pairings change. Volatility encompasses this uncertainty of the change in uncertainty. Volatility can also be learned and estimated, and is therefore reducible, unless the rate of volatility changes.

Lets return again to our London cyclist. They have recently discovered a shortcut through a park that reduces their journey time to work. Upon visiting this route the next day however, they are met with a closed gate and have to turn around. The cyclist is unsure whether this is just a one-off event, or if this might happen more regularly and decides to revisit the

park in order to find out and reduce their uncertainty. Over many visits to this park, the cyclist learns that the gate is actually closed around 75% of the time, which incurs quite a time cost. This example of second-order uncertainty (or ambiguity, in economics, see Box 1) refers to uncertainty about the predictive relationships between variables. This kind of uncertainty can be reduced with learning, allowing the agent to make accurate predictions, if the probabilities are estimated correctly. The most prominent example of this can be seen in classical conditioning paradigms, where a stimulus, usually a tone, will be paired with a negative outcome, usually a shock, at a particular rate, say 25%. After repeated observations, the agent can learn that in roughly ¼ trials they will receive a shock. They can then use this information to make decisions about whether they want to choose this stimulus over other options with different shock probabilities (Andreatta and Pauli, 2015).

6 months later, after giving up on their shortcut, a cyclist friend starts bragging about a new route to work through a park. This sounds familiar and our cyclist goes to investigate, discovering the gate is indeed open again, but is uncertain about whether this is an uncommon but still possible outcome, or whether something in the environment has changed. Upon returning many times, the cyclist is surprised to see that the gate remains open and deduces that in Summer the park may open earlier due to the lighter mornings. Encouraged by this, the cyclist continues to use the shortcut until the days begin to get shorter. This kind of higher-order uncertainty, known as *volatility* (see Box 1), refers to the rate at which the statistics of the environment change. This again, can be learned, but necessitates the re-learning of the second-order stimulus-outcome pairings. In environments with low volatility, uncertainty around environmental statistics is low and stimulus-outcome pairings are stable. Within highly volatile environments, stimulus-outcome pairings may become unreliable and lose

their predictive quality completely, impacting decision-making (Browning et al., 2015).

Even without having to make predictions about traffic light safety and shortcut availability, our cyclist has to integrate huge amounts of sensory information that is accompanied by varying amounts of uncertainty. For example, cycling through a cloud of thick fog reduces visibility and renders visual cues less reliable, as it is more difficult to differentiate between objects. If the shade of red presented on the traffic light was very close to the shade of orange, this would render discriminating between the two difficult, as with greater overlap in spectral colour there would be more uncertainty about true underlying colour, due to less specificity in visual activation patterns (Kalloniatis and Luu, 1995). This is on top of the sensory uncertainty that is always present due to internal noise in any estimation process (Bach and Dolan, 2012).

Whilst all this is going on, inside the body there is a huge array of internal bodily signals being relayed to our cyclist's brain. Information such as how hot or cold they are (Craig and Bushnell, 1994), how fast their heart rate is and how much pain they are experiencing in various parts of the body (LaMotte et al., 1982). These signals may become more or less important, depending on the external situation. On a very hot sunny day, temperature may be an especially important internal variable and the uncertainty around future bodily temperature changes may inform decisions about how much water to drink or how many breaks to take. When cycling in busy traffic, muscle fatigue may become an important variable, as weaving in and out of traffic requires precise muscle control and tiredness may reduce the cyclist's ability to avoid cars.

As we can see, something as seemingly simple as riding a bicycle involves an enormous amount of information to be processed at incredibly fast speeds. And yet, this consolidation of information is performed seamlessly, not breaking into our subjective experience. A caveat to this might be the feeling of being indecisive, when the mechanisms that act to reduce uncertainty get stuck, or become delayed. In extreme cases, such as in OCD, the thresholds that boundary decision-making are higher, rendering decision-making slower, requiring more evidence to reach termination point (Hauser et al., 2017). The question of precisely how the brain integrates these multiple layers of uncertainty is still largely unknown, however the interdependent nature is thought to be reflected in a hierarchical neurobiological representation (Bach and Dolan, 2012; Bach et al., 2011). This neurobiological search has recently been accompanied by a substantial amount of studies employing computational modelling, which, through precise mathematical formulation can directly quantify these concepts.

1.3 Computational Modelling

The proliferation of cognitive computational models has made strides in understanding the critical mechanistic processes that drive the observed outcome variables of experiments. When using traditional frequentist analysis methods that often use a summary statistic approach, for example the comparison of means when using an ANOVA, key information contained in the trial-by-trial dynamics is lost. By formulating a generative model of the assumed underlying process, computational modelling allows the rigorous characterisation of these dynamics. Precise hypotheses about psychological phenomenon are tested through the use of latent variables that are parameters of the model (Farrell and Lewandowsky, 2010). Important variables that change over time, such as value and uncertainty can therefore be

quantified and assessed in relation to experimental manipulations. The burgeoning use of computational modelling has necessitated the development of gold standard computational modelling procedures, such as those outlined in Palminteri et al. (2017) and depicted in Figure 1.2. The work in this thesis aimed to adhere as closely as possible to these guiding principles. Computational modelling usually requires work to be done before data collection actually begins (Ex Ante), beginning with the mathematical formulation of the model and therefore the computational processes of interest. This allows hypotheses to be tested out on synthetic data using simulations, which are crucial for not only understanding the dynamics of the process of interest, but for ensuring that the hypothesis can actually be tested using the model. An important part of this process is verifying that the model can be estimated and that parameters can be accurately recovered. This work also informs the experimental design used to collect data and can be used to optimise designs, for example by simulating how different reward schedules interact with the process of interest. After data collection (Ex Post), the models are fit to the data during model estimation. Multiple models providing different computational interpretation are fit to the data and then formally compared during model comparison. Common model comparison techniques involve comparing the Akaike Information Criterion (AIC) (Akaike, 1998) or Bayesian Information Criterion (BIC) (Schwarz, 1978), however standardised model comparison techniques, such as computing the pseudo-r^2 value Daw (2011) can also be useful to allow comparison across studies. This thesis therefore reports standardised fit values alongside AIC and BIC estimates. Palminteri et al. (2017) also stressed the importance of the model being able to recapitulate the original data (generative performance), as models that can provide a statistically sound fit to data might still mis-specify important parts of the process of interest. Failing the generative performance test is taken to be an absolute rejection criterion.

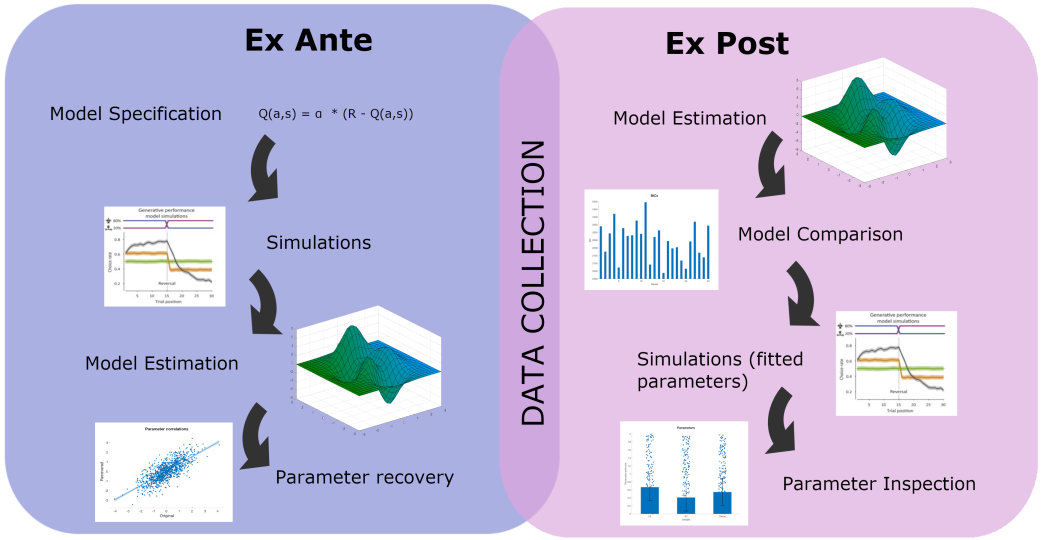

Figure 1.2: Graphical depiction of gold standard computational modelling procedure outlined in Palminteri et al. (2017). Ex Ante (pre-data collection), the initial model is specified and simulations are used to generate synthetic data in order to determine whether the model discriminates between hypotheses to describe the process of interest. This process likely iterates until a good candidate model is found. The model is then fit to the synthetic data to ensure parameter recovery is possible. Ex Post (post-data collection), the candidate models are fit to the real data and model comparison determines which model statistically describes the data best. Simulations are performed using the fitted parameters from each individual to determine whether the best-fitting model actually captures the process of interest. Individual parameters are inspected for hypothesis testing.

Associative learning (AL) models, such as those derived from Rescorla-Wagner (RW) equations (Rescorla and Wagner, 1972) are popular and useful computational models that provide a mathematical framework for evaluating how the *value* of stimuli changes over time, as more outcomes are observed. They have been applied across a wide range of fields (Roesch et al., 2012a; Hampton et al., 2008; Zhang and Gläscher, 2020; Lockwood et al., 2016; Lindström and Tobler, 2018) and have substantively increased our knowledge of the processes involved in both behavioural and neural correlates of learning. More detail, including the equations underpinning these models is discussed throughout the subsequent chapters and is therefore omitted here. Briefly, these models describe how associations are learned between stimulus outcome pairings through the process of reinforcement (Sutton et al., 1992). The

value of a stimulus, is updated according to the prediction error (PE), the difference between the actual and expected outcome. Larger PEs necessitate a larger value update, as more unpredictable outcomes signal that more learning is required for the associations to become known.

One disadvantage of these models however, is that without modifications, they do not explicitly provide estimates of trial-by-trial variations in uncertainty, as value is quantified using point estimates. Other AL models, such as Pearce-Hall models (Pearce and Hall, 1980) incorporate the magnitude of the PE into the value update, using an associability parameter (discussed in more detail in **Chapter 2**), which accounts for greater uncertainty at the beginning of the learning process and adjusts the size of the value update. However, uncertainty here only directly informs the speed of learning, it does not provide a mechanism for uncertainty to feed into action selection or allow different forms of uncertainty to interact.

1.3.1 Making uncertainty explicit

Explicitly quantifying uncertainty allows for a similar exploration of the dynamics of learning and value updating as AL models, but further allow levels of uncertainty to feed into each other, elucidating the dynamics of uncertainty integration. There are numerous formulations of computational models that provide explicit uncertainty estimates, many of which are based on the principles of Bayesian updating, developed by the statistician and philosopher Thomas Bayes. Again, these models provides a way to understand and quantify changes in value through the integration of prior beliefs and sensory evidence (likelihood) to form a posterior belief distribution. The Hierarchical Gaussian Filter (HGF) (Mathys et al., 2014, 2011), for example, is a popular Bayesian computational model framework that quantifies

different levels of uncertainty that act on a three-level hierarchy. Uncertainty at the top level (here, volatility) feeds down onto the parameter estimates of the levels below, thus providing a way to capture the interconnected between variables. In a study by de Berker et al. (2016), the authors applied this model to an aversive learning task, and were able to pinpoint physiological markers of stress, such as pupil dilation directly to changes in *outcome uncertainty*, suggesting that stress tracks forms of uncertainty that cannot be reduced through learning. This distinction is important and could not have been understood through models that don't make explicit separations between uncertainty levels. However, the HGF may not be appropriate for simple experimental setups that do not directly manipulate volatility, or for understanding how models of the self and other are constructed and maintained (explored in **Chapter 4**). Moreover, models may need to be specifically formulated to study the task being performed and therefore the use of such a general model may not always be the best approach (Bröker et al., 2018).

A simple computational model, based on approximate Bayesian updating that uses a formulation analogous to AL models is that of a belief-updating (BU) model, parameterised by a beta distribution. The parameters of the BU model, α and β form a belief distribution over the probability of outcomes and are directly updated upon receipt of each observation by increasing the count of both or either parameter, depending on the formulation of the model. BU models provide an explicit quantification of uncertainty, as an entire probability distribution with both mean and variance parameters are updated and estimated over time. These mean and variance quantities can then be used to inform the decision-making process within the model itself or can simply be used to explore parameter-symptom correlations, as can be seen from work by Wise and Dolan (2020), who observed a relationship between

higher levels of anxiety and overestimation of uncertainty, a quantity which was directly derived using their uncertainty-based computational model. Empirical work that has utilised these models alongside AL models have found model evidence in favour of uncertainty based models over AL models (Wise et al., 2019; Wise and Dolan, 2020), however they still remain largely unused in comparison to models of AL. A key question therefore, is whether these models that quantify uncertainty directly, are better able to capture the mechanisms that underpin learning and choice behaviour than models that don't. In pursuit of this aim, the models presented in the experimental chapters always include AL models in formal model comparison. Moreover, it may indeed prove useful to understand the correspondence between these two frameworks as this can aid in conceptual interpretations of the process of interest. The cross-model correspondence between AL and BU models for understanding social evaluation is directly investigated in **Chapter 4**.

1.4 Uncertainty preference and aversion

The aforementioned work has highlighted why measuring uncertainty is important for understanding learning and decision-making and indeed there is a myriad of research that has demonstrated that people are influenced by experimental uncertainty manipulations. In general it appears that people are uncertainty averse, such that they prefer situations with less uncertainty. This is reflected both in attitudes and also decision making. People are, in general, risk averse (Holt and Laury, 2002; Kahneman, 1979) and prefer to choose options that have greater certainty over options that have equal or even higher expected value but that carry more uncertainty over the outcomes. People are also generally ambiguity averse, with a preference for known risks over unknown risks (Ellsberg, 1961).

There is further evidence for uncertainty aversion observed in decision-making tasks that manipulate other measures of uncertainty, such as the the distribution of reward, not just the probability. Using a classical conditional task, Burke and Tobler (2011) showed that when participants were given choices between stimuli that represented reward distributions that had low-variance and stimuli that represented reward distributions with high-variance, the low-variance stimuli were chosen more often, suggestive of what the authors term 'variance aversion'. Neurally, there was a greater response to high-variance stimuli which was localised to the anterior cingulate cortex, an area often cited as a seat for uncertainty monitoring (Stolyarova et al., 2019). This 'variance aversion' can also be thought of as a 'certainty preference' and has been shown to have influence on other variables aside from choice, even processes outside of conscious deliberation, such as reaction times (RTs). A recent study by Zajkowski et al. (2019) gave participants a series of choices between reward cues that had differing degrees of certainty (either 100%, 80% or 20%). Higher reward certainty was associated with faster RTs that coincided with more frequent choice decisions.

1.4.1 Reward vs threat

The response to uncertainty may be influenced by whether the context is a rewarding or threatening environment. Aversive outcomes that are distal or unpredictable elicit a more sustained, anxiety-like response consisting of heightened vigilance, apprehension and risk assessment that is maintained until the uncertainty can be resolved (Davis et al., 2010). This resolution can be difficult, as potential threats can be far into the future or dependent on variables outside of the organisms control (Grupe et al., 2013). Similarly to rewards, aversive outcomes can vary greatly in their magnitude and cost, although aversive outcomes are not symmetrical

to rewards of similar magnitude and cost (Kahneman, 1979). Losses loom larger than gains and people weight losses more heavily than gains in their decision making (Charpentier et al., 2016a). In exclusively rewarding environments, where the delivery of reward is guaranteed, studies have shown that in non-human animals, there is evidence of a variance *preference* for reward delay, but variance *aversion* for reward magnitude (Buchkremer and Reinhold, 2010). That is, animals prefer reward delivery schedules that have a variable time course, but prefer reward distributions in which the size of reward is more predictable. The relationship is slightly unclear for humans however, as some studies have found either no preference (Weiner, 1966; McKerchar and Mazur, 2016) or preference for a fixed reward schedule over a variable reward schedule (Kohn et al., 1992), whereas others have found preference for a variable reward schedule (Locey et al., 2009; Lagorio and Hackenberg, 2010). The key differences between these studies were the types of stimuli used, with rewards that involving earning points inducing no or fixed preference, whereas rewards involving video clips, which were more similar to stimuli used in the animal studies, were associated with variable reward preferences. This may suggest that reinforcers that are more naturalistic, such as video clips, interact differently with uncertainty than more abstract, secondary reinforcers, such as points, but further research is needed to understand this distinction. Many of these studies also suffer from small sample sizes, which makes it difficult to draw firm conclusions.

A more recent study with a larger sample by Tsetsos et al. (2012) showed participants streams of value information drawn from either broad (high variance) or narrow (low variance) Gaussian distributions. When asked to evaluate which distribution had higher value overall, participants showed a *pro-variance* effect, such that they were more accurate for the broad distribution. Strikingly, this effect was reversed when the decision framing was

changed, such that when participants had to reject the worst option, they rejected the broad distribution more. Multiple studies have since found the same effects (Tsetsos et al., 2016; Cavanagh et al., 2020). This increased preference for sequences with higher variance (risk-seeking for gains) appears to contradict the well-known risk aversion effect, however here, the risk distribution is *learned*, such that participants are directly experiencing the samples of the probability distributions and updating their estimates of the value of the options based on their direct experience. This direct experiencing of the likely outcomes may be crucial, as previous studies have shown there is a difference between hypothetical and experienced outcomes (Ludvig and Spetch, 2011; Steele et al., 2019; Aronsson et al., 2014). Crucially, this difference reverses the typical gain risk-aversion, loss risk-seeking pattern, such that, in the gain domain, people are risk averse for hypothetical outcomes but risk seeking for experienced outcomes and vice versa for losses (Ludvig and Spetch, 2011).

The aforementioned paradigms largely rely on secondary reinforcers, such as monetary outcomes or points. Paradigms that utilise primary reinforcers, such as electrical stimulation may provide more powerful ways to investigate uncertainty within threat contexts, as unlike secondary reinforcers, they elicit unconditioned responses. Threat of shock (TOS) paradigms are an effective uncertainty induction, as the participant does not know both how likely it is they will receive a shock (estimation uncertainty) or precisely when the shock is going to occur (outcome uncertainty). TOS studies have been consistently shown to induce state anxiety (Grillon et al., 2008; Robinson et al., 2011; Davis et al., 2010) alongside changes to behaviour including increased avoidance (Mkrtchian et al., 2017), perceptual processes (Robinson et al., 2013), and have shown to negatively affect working-memory task performance (Lavric et al., 2003; Shackman et al., 2006; Vytal et al., 2013; Patel et al., 2016; Balderston et al., 2017). However, there is evidence that executive functioning processes such as the framing effect,

which captures the change in decisions as a function of the decision being framed as a gain or loss (Robinson et al., 2015) are not impacted. Temporal discounting of future reward was also not impacted by TOS (Robinson et al., 2015), however when the outcome that is being discounted is aversive itself, a different pattern emerges, with people strongly preferring to receive a shock sooner rather than later (Story et al., 2013). Thus, uncertainty about precisely when an outcome is due to occur is impacted by the aversive nature of the outcome, such that the uncertainty period before the outcome is administered is itself aversive. Taken together, these studies suggest that people are generally uncertainty averse, with uncertainty impacting cognition and behaviour especially within threat contexts with aversive outcomes, which also induces negative affect. However, the type of reinforcement used, as well as the context framing interacts with uncertainty, rendering uncertain threats different to uncertain rewards.

1.4.2 Individual differences in uncertainty preferences

The universality of uncertainty may suggest that organisms all respond to uncertainty in the same way. However, there are important individual differences in the response to and subjective experience of uncertainty. For example, when asked to make decisions about how much money to give to another person, people who have lower tolerance towards ambiguity engage less in prosocial behaviours (Vives and FeldmanHall, 2018). The perception of uncertainty can be measured using the Intolerance of Uncertainty Scale (IUS) (Carleton et al., 2012), which captures negative beliefs about uncertainty, thought to be stable across different time points (Buhr and Dugas, 2002). Individuals who are highly intolerant of uncertainty display differences in behavioural, neural and physiological processes. Numerous studies (Chen and Lovibond, 2016; Nelson and Shankman, 2011) have shown that the startle reflex is higher for

high IUS individuals, but only during conditions of high uncertainty (even if threat levels are lower), for example when the likelihood of receiving a shock is 50% rather than 75% (Chen and Lovibond, 2016). Similarly, high IUS has been related to other physiological metrics such as lower heart-rate variabilty (HRV) (Deschênes et al., 2016) and an increased skin conductance response (SCR) to fear conditioned stimuli (Morriss et al., 2015, 2016).

These results highlight that despite the universal nature of uncertainty, individual attitudes to uncertainty can shape the response to it, for better or for worse. Importantly, attitudes towards uncertainty appear to be important for psychopathology, with higher levels of IUS reported across many, mainly internalising, psychiatric disorders [1] such as Generalised Anxiety Disorder (GAD) (Dugas et al., 1997) (Freeston et al., 1994; Dugas et al., 1998) depression (McEvoy and Mahoney, 2012) (McEvoy and Mahoney, 2012), panic disorder (Carleton et al., 2014), Social Anxiety Disorder (SAD) (Carleton et al., 2010) and Obsessive-Compulsive Disorder (OCD) (Tolin et al., 2003; McEvoy and Mahoney, 2012) and has also shown to be predictive of Post Traumatic Stress Disorder (PTSD) symptoms following a traumatic event (Oglesby et al., 2016) and of stress in students (Bardeen et al., 2017). IUS could therefore be a candidate transdiagnostic factor that emerges as an important modulator for responses to uncertainty (Tanovic et al., 2018), and appears especially relevant for anxiety disorders, which comprise the majority of studies (Carleton et al., 2014, 2010; Boswell et al., 2013; Dugas et al., 1997). I now turn to focus on anxiety in particular and discuss the prominent role of uncertainty within anxiety, which has been conceptualised as a disorder of uncertainty (Grupe and Nitschke, 2013), delving into the relationships between anxiety and uncertainty at different levels of the uncertainty hierarchy.

[1] The use of the word disorders is used to be consistent with conventional language, but isn't fully endorsed due to the sometimes implied notion that disorders are irrational responses to an ordered world, rather than rational responses to a disordered world. I prefer the term anxious distress.

1.5 The anxious journey: From unease to terror

'It is the feeling of having in the middle of my body a ball of wool that quickly winds itself up, its innumerable threads pulling from the surface of my body to itself' – Franz Kafka.

To feel anxious is a state that everyone will have experienced to some degree during their lifetime. Kafka's description of anxiety is likely recognisable and captures the essence of the phenomena that, in his case, became pathological and plagued him throughout his life. Definitions of anxiety generally converge on it being an aversive state which manifests in reaction to distal, unpredictable or sustained threats (Robinson et al., 2013). Accompanying the subjective negative affect associated with anxiety, there are also physiological, cognitive and behavioural changes (Grillon et al., 2008; Davis et al., 2010; Robinson et al., 2013). Evolutionary accounts of anxiety purport that non-pathological anxiety is adaptive (Marks and Nesse, 1994; Meacham and Bergstrom, 2016), as it guards the organism against danger and motivates preparation to avoid potential threat in the environment (Bach, 2015). This is useful for survival, however if anxiety continues for long periods of time, occurs at inappropriate times or becomes uncontrollable, it may become maladaptive. [2]

There are a myriad of reasons why someone might feel anxious, ranging from the practical 'what if I can't find a nice place to live' to more existential forms of anxiety that have, for centuries, provoked the interest of philosophers such as Kierkegaard, who believed that anxiety stemmed from the dizzying, overwhelming freedom of decisions one can make. With such a rich array of provocations, it is little surprise that manifestations of anxiety can differ

[2]The word maladaptive in this thesis is used for convention, but refers to processes that appear suboptimal given the specific task goals, or processes that impede ones goals generally. It does not refer to the underlying reasons for why someone might have high levels of anxiety, which may have originally been an adaptive process, e.g. in response to environmental stress.

both within individuals across time and between individuals, rendering it a difficult concept to study in a systematic way. A prominent method of studying pathological anxiety has been to use the anxiety disorder categories outlined in the DSM-5, developed by the American Psychiatric Association (2013). Anxiety disorders as a group consist of GAD, panic disorder, specific phobias, agoraphobia, SAD and separation anxiety disorder. Together, they are the most prevalent class of mental health disorders (Ritchie and Roser, 2018), affecting around 7.3% of the population globally (Baxter et al., 2013) and are significant contributors to disability (Hendriks et al., 2014). The categories largely differ in the focus of the anxiety, for example, individuals with SAD have anxiety around social situations, are fearful of being evaluated negatively and might worry about being embarrassed, whereas individuals with separation anxiety disorder have anxiety that revolves around being apart from a close other and might worry excessively about their safety. However, there is significant symptom overlap across categories (Enoch et al., 2008; Lara et al., 2006; Zhou et al., 2008). Behavioural markers of anxiety, which can be found in many anxiety disorder categories are excessive avoidance, behavioral inhibition and restlessness. Cognitively, anxious individuals struggle with chronic worry, indecisiveness and intolerance and aversion to uncertainty, as well as trouble concentrating. Physiological symptoms such as increased heart-rate, sweating, muscle tension, insomnia and headaches are also common (MacLeod et al., 1986).

1.5.1 Dimensional approaches to mental health

Clinicians and researchers have battled with the problem of heterogeneity within disorder categories and the significant symptom overlap between disorders for decades. At the same time, quantifying psychiatric disorders in a dimensional manner is beginning to become

more popular amongst researchers. Early attempts to formulate all psychiatric disorders in a dimensional manner (Foulds and Bedford, 1975) helped develop the continuum models of psychosis (Claridge and Beech, 1995; Chapman et al., 1995) and scales to assess symptoms and behaviour without relying on diagnosis (Fergusson and Horwood, 1995). This theoretical and clinical work has paved the way for initiatives such as the Research Domain Criteria (RDoC) project, introduced by the National Institute of Mental Health (NIMH) (Insel et al., 2010) and The Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2018). The main initiative behind these projects is to pivot research away from relying on categorical approaches such as the classification system of the DSM-5 and towards a dimensional approach that span multiple domains, linking the brain with behaviour and advancing transdiagnostic mechanisms that cut across disorder categories (Cuthbert and Insel, 2013).

This dimensional formula for research promotes the development of novel methods to investigate psychiatric symptomatology, such as using factor analyses on questionnaires that span a range of mental health disorders to understand transdiagnostic symptom-behaviour relationships (Gillan et al., 2016; Rouault et al., 2018; Wise and Dolan, 2020; Patzelt et al., 2019). Furthermore, RDoC aims to bridge the gap between different units of analysis, from genes, cells and molecules all the way up to behaviour and self-reports. This is notoriously difficult and research so far has demonstrated only weak evidence for correspondence between units of analysis within the social processing domain (Clarkson et al., 2020). However, there have only been a limited number of cross-domain research in the decade since the creation of RDoC and therefore only time will tell whether these advances prove fruitful.

My work closely follows from these frameworks, and therefore takes a dimensional approach that is fast becoming the bedrock of Computational Psychiatry research (Huys et al., 2016). It has been proposed that pathological anxiety sits at the more extreme end of a continuum of adaptive anxiety (Dillon et al., 2014; Lebeau et al., 2012). Similar ideas have been suggested for other mental health problems, such as OCD, which lies at the extreme end of a compulsivity spectrum (Hauser et al., 2017). Spectrum approaches compliment the idea that the right amount of anxiety can be adaptive, improving the ability to detect and defend against threat (Robinson et al., 2012). In a computational investigation that used a stress induction (in a similar manner to state anxiety inductions), de Berker et al. (2016) found that people who perceived *more* subjective stress had a *greater* percentage of correct choices, thus stress here was adaptive and facilitated performance. My work builds upon this idea, and examines the computational mechanisms important for anxiety by taking a dimensional approach, making use of the spectrum of anxiety symptoms in non-clinical samples. This does not replace the need for studies that use clinical samples using traditional diagnostic measures. Indeed, these studies are crucially important for understanding symptom interactions at the highest end of symptom severity. The search for a mechanistic understanding of anxiety is aided by decades of psychological, theoretical and neuroscientific research that has identified important cognitive and behavioural markers of anxiety and has paved the way for computational modelling to explain these biases in computational terms.

1.5.2 Computational biases in anxiety

In the past few years alone, there have been a number of key studies within the field of Computational Psychiatry investigating alterations to learning and behaviour in anxiety.

Figure 1.3 highlights key computational studies identified by Raymond et al. (2017) at the time of publication in 2017 and proposes an interplay between aetiological factors that give way to a trait vulnerability towards anxiety. Here, biological factors, such as alterations to the serotonergic system (Dayan and Huys, 2008) are thought to give way to differences in learning from aversive outcomes (Browning et al., 2015), which relates to differences in observed behaviour, such as increased behavioural inhibition (Bach, 2015) and heightened threat processing (White et al., 2010a,b, 2016). The downstream effect of the interplay between these factors is the subjective experience of anxiety. The backwards mechanism from subjective anxious states towards learning and behavioural concepts is only briefly emphasised, however more studies are beginning to show how affective experience shapes decision making (Charpentier et al., 2016a; de Berker et al., 2016). The relationships between the majority of these core concepts have not yet been empirically tested, and motivates the aim of the work presented in **Chapter 5**, in which I begin to investigate the interplay between aversive learning and threat processing, alongside other factors introduced below.

Since the publication of Raymond et al. (2017), there have been further computational studies, displayed in Table 1.1, which summarises key existing computational studies relating to anxiety and their main results as assessed at the time of this thesis completion (2021). Only a few of these studies use computational models that explicitly quantify uncertainty, and are discussed more in detail in the subsequent section on uncertainty and anxiety. Briefly, studies that investigated learning using volatility manipulations highlight differences in learning rates (discussed in detail in subsequent chapters) for individuals high in trait anxiety. Utilising a reversal learning paradigm, which features blocks of stable and changing (volatile) action-outcome pairings, Huang et al. (2017) found higher base learning rates and Browning et al. (2015) found a reduction in the ability to adjust learning rates from stable to volatile

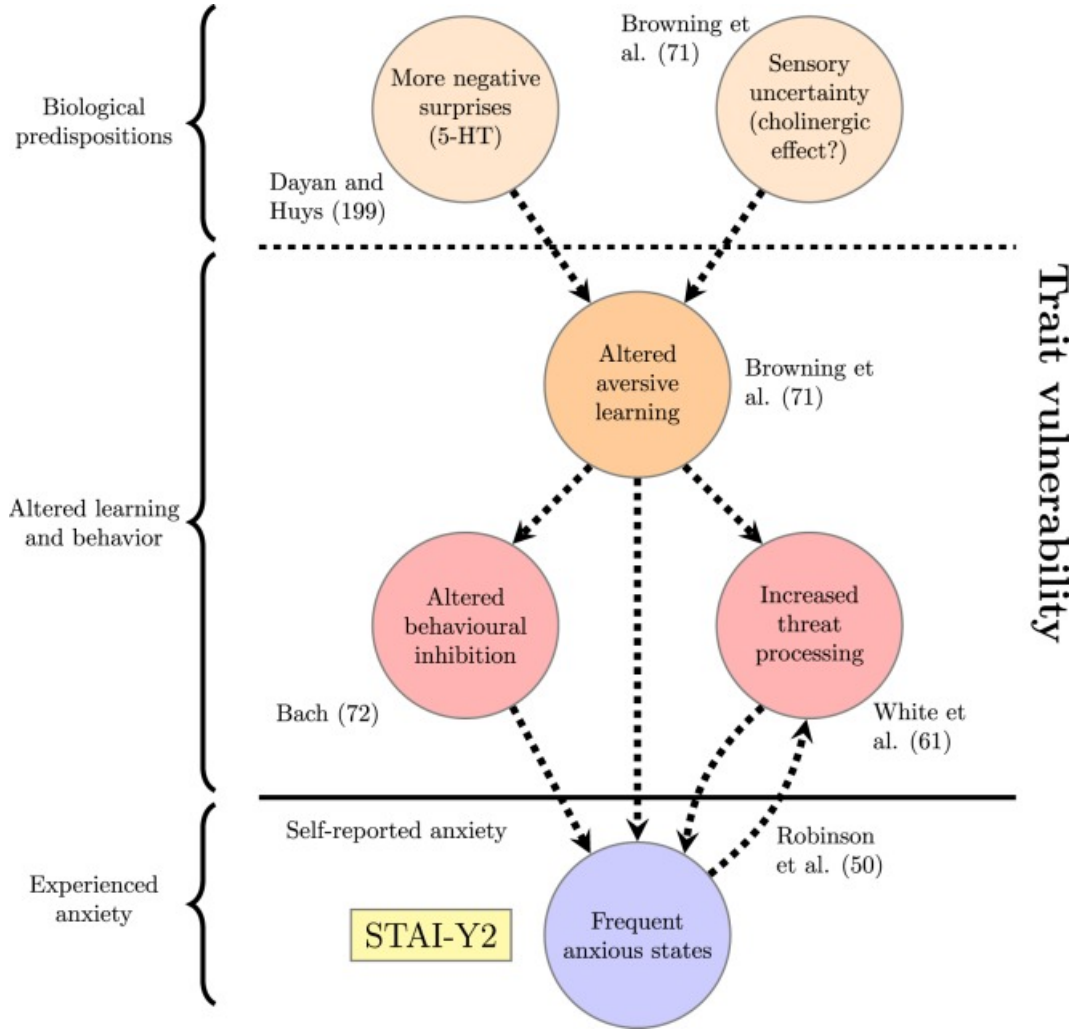

Figure 1.3: Reproduced with journal permission from Raymond et al. (2017). Factors underlying trait vulnerability to anxiety based on existing computational studies. Dashed lines represent connections that do not yet have experimental support.

conditions. This was replicated in a similar paradigm using monetary outcomes (Gagne et al., 2020) in a study which broadened understanding of these effects to depression as well as anxiety. These studies suggest that people with anxiety have difficulty in calibrating to environments where higher-order uncertainty is changing.

In aversive learning paradigms without volatility manipulations, there were some counter-intuitive results with anxiety reported in Wise et al. (2019); Wise and Dolan (2020), with greater updates for *positive* outcomes in state and somatic anxiety, but greater threat updating in cognitive anxiety. The reason for this discrepancy is unclear, but warrants further investigation, given that anxiety is typically thought to be a response to negative, not positive outcomes. However these results may highlight the importance of safety signalling or relief of negative outcomes as motivating factors in behaviour (Eldar et al., 2016). There also appears to be a relationship between the learning processes, and avoidance. A study by Norbury et al. (2018) found greater generalisation to aversive but not neutral feedback within an RL context, resulting in greater avoidance. Greater avoidance has also been observed in a pavlovian conditioning context, (Mkrtchian et al., 2017), with a state anxiety induction increasing reliance on a pavlovian avoidance parameter, suggesting a prepotent bias towards avoidance under conditions of state anxiety. In studies investigating model-based learning in particular, there is a mixture of results, with Gillan et al. (2016) observing no relationship of 'anxious-depression' with differences in model-based learning, however Sharp et al. (2020) saw evidence for impaired learning of state transitions. It is important to note that these two studies use different paradigms, with Sharp et al. (2020) specifically designing a paradigm that specifically enabled the delineation of state transition learning from other processes,

which is not the case in Gillan et al. (2016).

In non-learning contexts, increased threat processing, or a high threat bias is one of the most consistent results found. Using drift-diffusion models (DDM), White et al. (2010b, 2016) were able to localise a greater bias towards threatening vs neutral words to differences in the evidence accumulation process, observing larger threat drifts rates, lower neutral drift rates and a greater starting bias towards threat in high anxiety (HA). Similarly, Aylward et al. (2019) found a negative bias to ambiguous stimuli, characterised by a lower drift rate for positive responses to ambiguity. Within value based decision making (VDM), people with HA show greater risk aversion (Charpentier et al., 2017), preferring safe to probabilistic outcomes (this is discussed further in relation to uncertainty aversion below). Rouault et al. (2018) also reported an interesting relationship of anxiety and metacognition, with an anxious-depression dimension similar to the one in (Gillan et al., 2016) relating to lower confidence in perceptual decisions, but *greater* metacognitive efficiency, highlighting a dissociation between subjective feelings and objective performance.

The study of social processes within anxiety has not yet been the focus of many studies, however recent research into the computational mechanisms of self-esteem identified an interpersonal vulnerability factor consisting of depression, social anxiety and trait and state anxiety which was associated with greater weights on social PEs meaning their self-esteem was less stable (Will et al., 2017). Within the domain of social anxiety in particular, there is further evidence for enhanced learning from negative information observed in social evaluative situations (Koban et al., 2017), but no relationship to higher-order volatility manipulations (Beltzer et al., 2019), unlike in trait anxiety (Browning et al., 2015; Gagne

et al., 2020). These studies suggest differences in learning in social contexts within anxiety, however it is not yet known how this relates to uncertainty. This brief summary introduced some of the key computational mechanisms related to anxiety and I now examine more closely the uncertainty related processes specifically.

Study	Model	Paradigm	Stimuli	Subjects	Results
Charpentier et al. (2017)	Prospect	Probabilistic gambling	Monetary gains/losses	n = 25 GAD n = 23 HC	HA greater risk aversion
White et al. (2010b)	DDM	Lexical DM	Neutral/threat words	LA n = 21 HA n = 21	HA higher threat drift rate
White et al. (2016)	DDM	Lexical DM	Neutral/threat words	LA n = 39, HA n = 37	HA higher threat drift rate and starting point
Aylward et al. (2019)	DDM	Associative learning	Aversive (shock)	LA n = 47 HA n = 30	HA slower drift rate to positive ambiguous
Gillan et al. (2016)	RL	Two-step	Monetary reward	Unselected sample. N = 548	No MB relationship
Rouault et al. (2018)	SDT	Metacognition	Perceptual DM	Unselected n = 955	Lower confidence, greater metacognitive efficiency
Sharp et al. (2020)	RL	One-step re-evaluation	Monetary reward	Unselected online sample n = 174	Impaired learning of state transitions
Huang et al. (2017)	RL	Change Point Detection	Monetary reward (points)	LA n = 45, HA n = 77	HA higher base LR
Browning et al. (2015)	RL/bayes	Reversal learning	Aversive (shock)	n = 31	HA reduced stable - volatile change in LR
Gagne et al. (2020)	Bayes	Reversal learning	Pain / monetary gain / monetary loss	n = 88. 20 MDD, 12 GAD, 26 HC.	Reduced learning to +PE
Mkrtchian et al. (2017)	RL	Go/no-go	Aversive (shock)	HC n = 58 HA n = 43	HA increased avoidance parameter
Norbury et al. (2018)	Bayes	Instrumental avoidance task	Pain / monetary losses	n = 26, n = 482	HA greater generalisaton from aversive feedback
Wise et al. (2019)	Leaky beta	Aversive learning	Aversive (shock)	Unselected n = 65	Higher updating for no-shock outcomes in state anxiety. No relationship to trait anxiety
Wise and Dolan (2020)	Leaky beta	Aversive learning	Points	Unselected n = 400	Greater safety update state and somatic anxiety. Greater learning from threat, depression and social anxiety
Koban et al. (2017)	RL	Speech and social feedback	Performance feedback	21 SAD, 35 HC	SA higher learning rate for self-negative information
Beltzer et al. (2019)	RL	Volatile Social Learning (Cyberball)	Social reward and exclusion	Unselected n = 222	No relationship to volatility

Table 1.1: Anxiety and mood focused empirical computational modelling studies and key results. Studies that directly quantify uncertainty using a computational model are highlighted in bold.

1.5.3 Uncertainty and anxiety

The aforementioned computational work suggests that many of the biases observed in anxiety relate directly to experimental manipulations of uncertainty. Supporting this idea, important theoretical work by Pulcu and Browning (2019) has proposed that the estimation, or *mis*estimation, of uncertainty is especially important for the aetiology and maintenance of affective disorders such as anxiety. For example, the aforementioned learning rate differences between low and high anxiety individuals in Browning et al. (2015) arise from the change in task demands and speed of learning required by highly unpredictable (volatile) vs predictable (stable) conditions. Thus, as Pulcu and Browning (2019) suggest, to be uncertain to the right degree is crucially important in order to set one's learning rates correctly. The results of de Berker et al. (2016) provides computational support for this idea, as individuals who more accurately tracked the different levels of uncertainty had better performance overall. Thus the veridical coupling between task uncertainty and ones estimate of that uncertainty appears important for behaviour. Wise and Dolan (2020) has shown experimental evidence for the overestimation of uncertainty in aversive learning, at least in cognitive anxiety, but more computational studies are needed that explicitly quantify uncertainty in order to investigate this in different contexts. Additional importance of the role of uncertainty in anxiety arises from the subjective perception of uncertainty. Whilst being intolerant of uncertainty by itself is not necessarily indicative of mental health problems, it is a core feature of pathological anxiety (Boswell et al., 2013) and a significant risk factor for different anxiety disorders (Gu et al., 2020), suggesting its transdiagnostic importance. Thus it is crucial to understand how uncertainty impacts both behaviour and subjective perception, which is explored in *Chapter 3*. I now turn to look at how anxiety is related to different forms of uncertainty in more detail, making reference to computational studies wherever possible.

1.5.4 Anxiety and the uncertainty hierarchy

Sensory uncertainty

A key feature of anxiety is a hyper exaggerated vigilance response, which manifests in extreme monitoring of the environment (Grillon, 2002; Grupe and Nitschke, 2011, 2013). In threat contexts, hypervigilance can help respond quickly to threats, however such a state of hyperarousal can become difficult to manage if it persists for long-periods of time and can also produce inappropriate behaviour in non-threat contexts. Hypervigilance is linked to hyper responding of the sensory systems designed to inform the organism about changes in the environment. There is a large amount of evidence that supports this notion in anxiety. Questionnaire studies have related anxiety symptoms with differences in daily sensory experiences, showing that people with higher trait levels of anxiety display greater sensory hypersensitivity (Engel-Yeger and Dunn, 2011). Moreover, high levels of sensory processing sensitivity is a strong predictor of anxiety and depression (Liss et al., 2005) and can explain high levels of variance in symptoms (Ahadi and Basharpoor, 2010). Experimentally, hypervigilance has been related to amplified attention to fear-relevant stimuli (Weymar et al., 2014).

Few experimental studies have directly manipulated sensory uncertainty in order to examine anxiety. Herry et al. (2007) compared the neural response to predictable vs unpredictable sound sequences which were unrelated to the task goals. They found that the unpredictable sound sequences increased anxiety-like behaviours measured by attentional bias to emotionally threatening (angry) faces. Subjects in the unpredictable condition had faster RTs to the faces than the predictable conditions. The authors investigated where the effect of the unpredictable sound sequences was represented neurally and found that there was heightened sustained activity in the amygdala. This enhanced sustained activity is noteable given that

repeated stimulus presentations typically results in habituation. I investigate the relationship between sensory uncertainty and anxiety in **Chapter 3**.

Differences in processing outcome uncertainty, heightened risk aversion

Clinical anxiety often presents with decision-making difficulties and behavioural experiments have consistently found increased risk avoidant behaviour across anxiety disorders (Butler and Mathews, 1983; Maner et al., 2007; Mueller et al., 2010) and has been suggested as a trait feature of anxiety (Giorgetta et al., 2012). Theoretical accounts of outcome uncertainty rest largely on behavioural economic theory. Most widely applied models such as expected utility theory (von Neumann and Morgenstern, 1944) subjective utility theory, (Savage, 1972) and prospect theory (Kahneman, 1979) feature a non linear value function that maps the objective value of an outcome (usually monetary gains) to the subjective value of that outcome. In line with this framework, a study by Charpentier et al. (2017) used a computational model in order to determine whether the heightened risk averse behaviour was a feature of loss aversion or risk aversion. They found differences in the risk aversion parameter, not loss aversion, suggesting people with anxiety do not weight losses more heavily, but instead weight the uncertainty of the calculation differently.

In addition to associations with decisions based on known outcome uncertainty, processes governing the *learning* of outcome uncertainty have also been implicated in symptoms of anxiety. In probabilistic learning tasks, subjects must accurately learn the probability of receiving positive or negative outcomes in an inherently uncertain environment. Biases in the way these outcomes are incorporated into ongoing probability estimates can result in accurate estimation of uncertainty, for example leading the individual to believe outcomes are

more uncertain than they actually are, or biasing the estimation of uncertainty in a negative way (for example estimating the probability of an aversive outcome as 90% rather than 70%). Symptoms of anxiety have been linked to a negative learning bias, resulting in pessimistic expectations (Aylward et al., 2019; Huang et al., 2017), although not all studies have found such a bias (Gagne et al., 2020; Wise et al., 2019). As discussed previously, it might be that this bias is specific to cognitive anxiety specifically, such as worry (Wise and Dolan, 2020).

Negative bias towards ambiguous stimuli

Individuals with HA process ambiguous stimuli differently to neurotypicals, however ambiguity aversion as defined by a neuroeconomic framework has not been rigorously tested with computational modelling. Outside of a neuroeconomic framework, stimuli is often termed ambiguous if there is high uncertainty about which category the stimulus falls into, for example whether a facial expression is angry or sad, or if there are many potential interpretations for a given scenario. When tasked with providing interpretations to emotional events, highly anxious individuals give negative interpretations to ambiguous events, whereas non anxious individuals will give positive interpretations (Hirsch and Mathews, 1997). When presented with hypothetical future scenarios that were ambiguous, people with anxiety rated these scenarios as having a higher probability of a negative outcome occurring. Furthermore, they predicted these negative events would be more costly (Mitte, 2007; Butler and Mathews, 1983). This suggests anxiety is related to a pessimistic bias for ambiguous stimuli, such that there is a tendency to evaluate ambiguous information more negatively.

Using computational modelling, a study by Aylward et al. (2019) paired ambiguous and non-ambiguous tones with either low or high rewards. They observed that for ambiguous but

not unambiguous tones, individuals with mood and anxiety symptoms were less likely to associate this stimulus type with high reward, indicative of a negative bias towards ambiguous stimuli. Using a DDM, they were able to identify the source of the bias as a slower drift rate for the ambiguous tone sequence for high rewards. This bias was not specific to anxiety, however and adds to the growing amount of evidence that uncertainty processing is especially relevant for affective disorders generally (Pulcu and Browning, 2019).

Differences in volatility adjustment

The ability to adjust ones expectations to changes in the environment is essential for being able to adapt to changing conditions. There is converging evidence that people with anxiety are slower to adjust to changes in the statistical environment. Neurotypicals adjust their learning rate in response to the environment, increasing the speed of their updating with higher volatility conditions and decreasing the speed of their updating for lower volatility conditions (Behrens et al., 2007). Browning et al. (2015) showed that for participants with high trait anxiety, this adjustment of learning rate according to the levels of volatility in the environment was slower, meaning they showed worse performance in the highly volatile conditions. Recent research (Gagne et al., 2020) shows that this is not specific to anxiety and is a general feature of internalising disorders such as depression. Importantly, this slower volatility adjustment was observed across both primary (shock) and secondary reinforcers (money) and was even observed in appetitive contexts, suggesting volatility adjustment isn't only a prominent feature of threat contexts.

Uncertainty around self-concept and other-concept

The aforementioned uncertainty processing differences so far have mainly been discussed in relation to neuroeconomic theory, often using monetary outcomes as rewards or losses.

Humans however, operate in a decidedly social context and how we relate to ourselves and others is a crucial part of our everyday lives. We live in an incredibly complex social world and need to hold, update and maintain models of not only ourselves, but ourselves in relation to others. The most obvious relationship to self and other concept is with social anxiety, which is defined by a fear of social interactions. Clinical social anxiety disorder is one of the most common forms of anxiety disorder and causes significant disruptions to an individual's ability to function (Taylor and Montgomery, 2007).

A number of studies point to uncertainty as being a crucial factor in the negative interpretation bias observed by people with social anxiety across different situations. People with social anxiety tend to classify faces that have ambiguous expressions as threatening, indicative of an interpretation bias that is present with stimuli that are high in uncertainty (here, sensory ambiguity), (Yoon and Zinbarg, 2007, 2008). Similarly, when asked to give interpretations of social events, people with social anxiety will evaluate ambiguous events more negatively (Clark and Wells, 1995). This suggests that people with social anxiety process uncertainty differently to their neurotypical counterparts in that high uncertainty and negative interpretations tend to co-occur.

However, dispositional anxiety has also shown to be related to low self-esteem, with a meta-analysis by (Sowislo and Orth, 2013) demonstrating the reciprocal relationship, with high anxiety predicting low self-esteem and vice versa. Computationally, it appears as though anxiety and self-esteem are related through a reduced uncertainty in self-concept, which results in beliefs about the self that are more easily shifted. A study by Will et al. (2017) investigated self-esteem and showed that anxiety significantly loaded onto an interpersonal

vulnerability factor, which was related to computational parameters within the study. Namely, people with greater anxiety were shown to have greater weights on social PEs, meaning they were more easily shifted by violated expectations about being approved. This results in greater variability in the response to approval, with people with high anxiety having not only lower, but more variable expectations about approval.

In a subsequent study utilising the same paradigm, Will et al. (2020) recruited subjects from the top and bottom 10% of the Rosenberg self-esteem scale (Rosenberg, 1965) from a large population based cohort. Again using a computational model, they were able to show that subjects with low self-esteem placed greater weight on social PEs when giving their self-esteem ratings. Importantly, the fluctuations were more volatile, meaning subjects with low self-esteem shifted their ratings to a larger degree. This study did not directly quantify trial-by-trial uncertainty in the computational modelling, thus it cannot be determined whether value updating through associative learning is the best way to conceptualise the social evaluative learning process or what the role of uncertainty about the self is. Belief-based models offer an alternative approach and have been applied to this data in a recent study (Low et al., 2021), which sees belief-based modelling showing superior performance to AL models. Here, self-esteem is conceptualised to be influenced by updating beliefs about self approval, incorporating both the likelihood of approval and the uncertainty surrounding approval.

A study by Koban et al. (2017) was one of the first to use computational modelling to elucidate the learning process in people with SAD. Using AL models based on the Rescorla-Wagner formulation, they found that socially anxious people displayed a lower positivity

bias than neurotypicals, who had a larger update for self-relevant positive feedback. This difference manifested in a greater negative learning rate when performing self-evaluations for the socially anxious group. In this study, uncertainty was not directly quantified and associative models were not directly compared to belief-based models, therefore there is no way to establish the role of uncertainty.

Differences in social interactions and social learning, although not specific to anxiety, may therefore play an important role in the disorder, especially given how ubiquitous social interaction is. Models that quantify uncertainty directly would be a useful way to understand what role uncertainty has to play in this process and I address this question in **Chapter 4**.

1.6 Neural uncertainty representations

The complex, multilevel nature of uncertainty representation has been a subject of great interest in neuroscience for decades. A key question is whether different forms of uncertainty are encoded and held in brain regions whose specific task it is to encode and process that uncertainty, or whether there is a general 'uncertainty processing' region that is always involved with tasks that require high levels of uncertainty processing. The answer appears to be both, as evidence from multimodal imaging studies support both modality-specific and modality general uncertainty processing (Nastase et al., 2014, 2018), with the frontoparietal and orbitofrontal cortices in particular showing cross-modal representation.

This combination of specific and general processing areas appears to also apply to the decomposition of uncertainty within the uncertainty hierarchy. The insula, particularly the anterior insula features prominently in a large number of decision-making under uncer-

tainty studies, highlighting its role in outcome uncertainty (Preuschoff et al., 2008; Elliott et al., 2000; Schultz et al., 2008), ambiguity (Huettel et al., 2006) and volatility (Jiang et al., 2015). Similar involvement across multiple levels of uncertainty can be seen in the OFC and amygdala (Hsu et al., 2005). For example, the OFC is thought to play an important role in outcome uncertainty, with increasing uncertainty around the reward value correlating with increased OFC activity, independently of expected value (Tobler et al., 2007).

For higher-order uncertainty, key work by Behrens et al. (2007) implicated the ACC as a key region for uncertainty monitoring, specifically for unexpected changes in reward contingencies, or volatility. Their study indicated that participants updated their reward rate contingencies in an optimal manner according to how much volatility was in the environment. Accordingly, higher ACC activation corresponded to more volatility, and greater adjustment of the learning rate. Payzan-LeNestour et al. (2013) investigated the neural representation of unexpected uncertainty and disambiguated this from volatility, by keeping the rate of change of the probability changes constant throughout the task. The task employed was a six-arm bandit task, in which participants had to choose between two options in order to gain the most reward. They disambiguated between different types of uncertainty through a Bayesian model, which allowed the separation of representation in the neural activity. The results of the study indicated that unexpected uncertainty held a negative association again within the insula, such that higher unexpected uncertainty corresponded to reduced activation. Low uncertainty outcome trials showed the converse relationship. Crucially, the pattern of this activation was distinct from other estimates of uncertainty, such as risk and volatility, suggestive that indeed these forms of uncertainty have a shared but distinct pathway.

1.6.1 Individual differences in uncertainty representations

We know that individuals differ in their attitudes towards uncertainty and that this affects behavioural and physiological markers of anxiety, but how does this shape their neural responses? The few structural studies that have examined this have observed increased grey matter volume in the right superior temporal pole in high IUS individuals (Hilbert et al., 2015) and a positive correlation between striatal volume, especially in the putamen specifically and IUS (Kim et al., 2017), although further research is needed to determine the relationship between this region and others that show functional differences. Further studies have highlighted differences in functional neural activity in high IUS individuals. The anterior insula in particular shows heightened activity to uncertain emotional faces (Somerville et al., 2013; Simmons et al., 2008) and rewards (Gorka et al., 2016) and this increased activity is positively correlated with IUS. A further key area in relation to IUS has shown to be the amygdala. Greater and more sustained amygdala activity to uncertain images (Somerville et al., 2013; Schienle et al., 2010) and uncertain rewards (Krain et al., 2008) was found to be positively correlated with high IUS across a number of studies.

Given the prominence of the amygdala in uncertainty research and its theorised computational role of uncertainty representation (Pearce and Hall, 1980), the following chapter provides a review of the amygdala in detail, taking a computational perspective and highlighting, where-ever possible, studies that link to psychopathology. This review chapter sets the context for and motivates the study presented in **Chapter 3**, which aimed to investigate the role of the amygdala and other brain regions in sensory uncertainty and threat perception. Following this, **Chapters 4** turns to investigate the computational mechanisms in involved in social evaluation using novel uncertainty based models, which are again used in **Chapters 5**, where I attempt to understand the relationships between different computational mechanisms.

2 Chapter 2: The Human Amygdala: A Computational Perspective

2.1 Acknowledgements

I would like to thank Laurence Hunt, Dominik Bach, Miriam Klein-Flugge, Rob Ruttledge, Jochen Michely and Zeb Knuth-Nelson for invaluable discussions, advice and suggestions for this review.

2.2 Abstract

Limited research into value-based decision-making has examined the neural computations taking place in the human amygdala, yet such research has huge potential to enrich our understanding of this important structure. We review this research, which suggests that the amygdala represents the expected value of stimuli, with stronger activity for more positive values in both the aversive and appetitive domains. Studies hint at an important computational specialisation of amygdalar subnuclei. The basolateral amygdala may more closely track appetitive value, while the central nuclei may track elements of aversive learning, and the need to change action choice. Mixed evidence implicates prediction-error (PE) encoding, with most studies finding no PE representation, but key findings suggesting an important role for aversive PEs. Unpredictability appears to play a crucial computational role, with increasing errors inducing increases of associability, i.e. effective learning rates, by the amygdala. There is evidence that the amygdala implements associative, (model-free) computations, but may also approximate sophisticated modelling of the environment (model-based control in dual-systems learning theory). The latter may be particularly important in signalling uncertainty

and associability. We relate current non-computational theories of amygdala function with the computational roles of affect, and link this to computational biases seen in psychopathology. Finally, we point out a number of gaps in understanding and opportunities for discovery that future research may address.

2.3 Introduction

Over the last two decades human computational neuroscience has flourished, bringing unique insights into the mechanisms employed by the brain during learning, experiencing affect and making decisions (Kriegeskorte and Douglas, 2018; Charpentier et al., 2016a; Rutledge et al., 2014). The tools of human computational neuroscience have great potential to elucidate the undisputed roles of the amygdala in the processing of salient stimuli, the orchestration of responses to them, the construction of complex neural patterns such as 'emotion constructs', and of course in learning useful maps between stimuli and responses (LeDoux, 2012). Here, we overview the contributions of value-based learning, inference and decision-making research in elucidating key functions of the amygdala, through reviewing the limited but important human neuroscience work in this area. Hence, we provide guidance as to the most needed next steps. We focus on studies that employ computational models to understand the function of the human amygdala, especially human cognitive neuroimaging. However, we also draw selectively from the rich multi-disciplinary literature on this structure, to interpret and inform computational theory.

We will discuss the likely computational interplay of the amygdala with regions which play key roles in value-based computations (VBC), especially the striatum, as the amygdala is densely connected to many of these regions. We will take into account the extensive

neuroimaging literature which indicates that amygdala structure and function, as well as the function of these key areas with which it intimately interacts, is affected in clinical populations (Stein et al., 2007; Siegle et al., 2007; Armony et al., 2005). Computational studies may elucidate the mechanistic processes that are awry in these populations, and relate them to the fine structure of the amygdala (see Box 2). This is important because human neuroimaging research, like cellular level studies, indicates that distinct subregions of the amygdala may have substantially different neurocomputational roles, with recent research showing dissociable roles in aversive learning (Michely et al., 2020). In particular, the functional separation of sub-regions may provide crucial information for distinguishing valence dependent and valence independent computations.

However, research on the human amygdala needs to take heed of methodological concerns regarding mapping indirect measures of neural activity to the computational quantities likely to be represented in this structure. Only when these methodological issues are addressed, which are arguably more serious for the amygdala than e.g. the striatum or neocortex, will the computational functions of the amygdala be elucidated with high confidence, analogous to invasive recordings. One key example is that the haemodynamic response function of the amygdala appears to be more complex than that of the neocortex. We discuss consequences of these methodological issues, suggesting that further research is needed to clarify the best methods to study different tasks and paradigms involving value based decision making.

Box 2: The amygdaloid complex

The amygdala is an almond shaped structure that sits within the evolutionarily ancient limbic system. Evidence from neuroimaging studies in humans has begun to elucidate the specialised anatomical and functional subregions that have been previously observed in animal studies (Sah et al., 2003; Qin et al., 2014). In particular, there are three major subregions that display some degree of functional specialisation, described below.

Basolateral Nuclei
Consisting of the lateral, basolateral (BLA), basomedial and basoventral nuclei that receives input from cortical regions such as the prefrontal cortex and subcortical regions such as the thalamus and hippocampus (LeDoux, 2003). Thus, the BLA is the seat of substantial sensory afferent connections.

Emerging evidence from model-free learning studies appears to implicate the BLA with (un)predictability, particularly at the time of *cue presentation*. Boll et al. (2013), observed reduced activity with higher associability parameter, meaning increased cue-outcome predictability suppressed activity, or conversely when there is high cue-outcome uncertainty, the BLA increases its activity. It also must carry a value representation as it is involved in the encoding of aversive PEs (Michely et al., 2020). Taken together, it appears that the BLA is substantially involved in tracking expectations.

Centromedial nuclei
Consisting of the central and medial nuclei (CMA), it has projections to the brainstem and striatal regions such as the caudate (LeDoux, 2003) and is the major output centre of the amygdala.

The CMA appears to be primarily concerned with *outcome*, especially for aversive events. In Michely et al. (2020), the CMA was insensitive to predictive aversive cues, but showed an increased response to aversive outcomes. In Prévost et al. (2013), there was a positive correlation with the value expectation signal for aversive events. Interestingly, the authors found a positive correlation for precision signals for both appetitive and aversive events, potentially indicating a role in general uncertainty modulation.

Superficial or corticomedial nuclei
The corticomedial nuclei are found at the most outer layer of the amygdaloid complex. They have a layered structure, similar to the structure of the cortical surface of the brain (Sah et al., 2003). The functional specialisation of the corticomedial nuclei has been less studied in humans relative to the BLA/CMA.

In a pavlovian conditioning paradigm with electric shocks as aversive outcome, Boll et al. (2013) observed positive correlations in the corticomedial nuclei with the unsigned PE, suggestive of a surprise signal that is concerned with magnitude but not valence.

2.4 Why is a computational approach needed?

"The purpose of computation is insight, not numbers" - Richard Hamming

Computational studies are increasingly important for the functional study of any specific regions of the brain. Cognitive science often tests whether a neural substrate such as an area of the brain responds to a contrast corresponding to a specific cognitive hypothesis (e.g., does this area respond more vigorously to grape or cucumber?) Hypotheses about brain function here are usually conceptual descriptions and the outcomes of interest are typically stimulus-brain-behaviour correlations. Computational methods translate conceptual description of the hypothesis into a mathematically specified, mechanistic theory with potential to elucidate how exactly information is processed by the brain (e.g. probability of choosing grape vs. cucumber $\leftrightarrow$ activation $\leftrightarrow$ subjective sweetness). Computational modelling is not a homogeneous field however, and there is a wide range of computational models that have a different focus, depending on whether they aim to capture biological detail, or to give an algorithmic description of the putative cognitive process. For example, many neural network models are biologically plausible, considering how populations of neurons may respond to stimuli and interact with each other to excite or inhibit neural activity, but do not directly quantify psychological constructs. Whereas, cognitive computational models typically quantify a particular theory through precise parameterizations, and often utilise correlation analysis to relate a particular parameter to brain activity. Kriegeskorte and Douglas (2018) suggested models of the mind/brain should attempt to capture both elements and should integrate biological dynamics with cognition (LeCun et al., 2015; Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Eliasmith et al., 2012; Hassabis et al., 2017). However, neuroscience is not yet ready to explain brain function, and indeed the function of the amygdala, across all levels

of description.

We argue that a very fruitful level to focus on is that of computations subserving *valuing* perceived states of the world, possible actions that one can take, and *selecting behaviours*. This is precisely because the amygdala is indisputably involved in evaluation, learning and acting – classically, in defense of the organism. We focus on neuro-computational models which can help analyse and understand human brain data. Such neuro-computational models are 'generative', that is, they can reproduce behaviour in cognitive tasks. If the computational mechanism has been captured well by the model, we should be able to reverse engineer the process. The model parameters can be adjusted ('fitted') so that the model explains experimental data, and the values of the fitted parameters then test whether a particular quantitative hypothesis is true. The models can then be used create synthetic data, replicating the mechanism of interest and reproducing the detail of experimental data (Palminteri et al., 2017), including, most importantly, patterns of neural activity.

2.5 Value-based decision-making

Value based decision-making (VDM) is the part of computational neuroscience which aims to understand how humans estimate value and weight different options, integrating their value into their decisions. Much evidence suggests that the brain instantiates VDM. A set of 'classic' brain areas have been identified, including the ventral striatum and frontal cortical areas, whose neural activity reflects value-related quantities (Pagnoni et al., 2002; McClure et al., 2003; O'Doherty et al., 2003; Ramnani et al., 2004). At the same time, there is increasing recognition of amygdala involvement in value calculations. Consistent with this,

the striatum and amygdala share a high degree of structural connectivity, as the amygdala is a substantial seat of projections to the striatum (Sah et al., 2003). Relatedly, a number of functional connectivity studies have shown a linkage between the two regions (Koelsch and Skouras, 2014; Satterthwaite et al., 2011). Roy et al. (2009) demonstrated that resting state spontaneous activity in the centromedial nuclei was predictive of striatal activity.

Reinforcement Learning (RL) (Sutton et al., 1992) is an efficacious framework for understanding value-based decisions. RL simply assumes that agents, in our case people, can tell what a valuable or 'rewarding' outcome is, and strive to optimize outcomes with respect to these rewards. They take an action while being in a particular state, and can represent the value of both states, $V(state)$, and of actions, $Q(action, state)$. Our first question, therefore, will be whether the amygdala represents such values at the voxel level, evidenced by functional neuroimaging. Agents may work out what actions may lead to the best outcomes; and having acted upon their environment, they can learn from experience. They can observe how good the outcome is, and hence update its action-value, in order to discourage or promote (reinforce) its future use[3]. Reinforcing actions has to be done very judiciously, as the ways in which outcomes flow from actions contain key *uncertainties*. The amygdala may have a very important computational role to help agents take uncertainties into account in evaluating actions.

The statistically optimal, yet often impractical, way to learn from uncertain outcomes is provided by the Bayesian framework. According to this, agents learn by using observations to improve their knowledge depending on how uncertain this knowledge was in the first

[3]In RL, reward values are usually considered as one-dimensional (scalar), and negative reward values are usually called 'punishments'. It is best to be cautious about this terminology, as the behaviourist, social-psychological and other traditions use more complex definitions of 'punishment'.

place. The same observation will have a big impact when we are ignorant, but much less if we already have much evidence. Structured – yet uncertain – knowledge forms the agent's model of the environment, technically a 'generative model'. Upon observing data d, the pre-existing (or 'prior') strength of our conviction (or 'belief') that a cause c obtains, $p(c)$, is modified proportionately to the *likelihood* that this cause can indeed give rise to the data, $p(d|c)$. The result is an updated or 'posterior' belief, $p(c|d) \propto p(d|c) \times p(c)$. Unfortunately, such model-based (MB) decision-making using Bayesian inference is difficult to implement. Yet its potential for optimal decisions means that in biologically common scenarios, evolution may have furnished good approximations to it (Dayan and Abbot, 2001; Huys et al., 2015; Gershman, 2019; Sanborn and Chater, 2016). As an important example, the amygdala may help estimate uncertainty in multi-stage processes found in nature. Here, variability in upstream stages produces higher-order uncertainty and influences downstream stages, which contribute their own uncertainties, eventually producing observed rewards.

Even before considering learning, evidence shows that people do not choose so as to maximize total rewards, as basic RL might predict, but use systematic biases to guide their choices. Such biases have been well described by behavioural economic theories, especially Prospect Theory (Kahneman, 1979), which has received substantial empirical validation in humans (Charpentier et al., 2018, 2016a; George et al., 2019). RL models thus make heavy use of such neuro-economic biases. In a fascinating twist, if actions are chosen probabilistically, then biases may be understood as the pre-existing, or prior, strength of our conviction that a particular action is the best in the context at hand. The amygdala may be involved in determining these biases, as we shall see.

Thankfully, in many scenarios precise probabilistic decision-making can be approximated by implementing simple associations, without considering complex Bayesian models. This associationist framework is known as model-free (MF) control of action[4], and has been extensively applied to amygdala function. In MF systems, direct updates of action values are made, by associating actions with the value of their outcomes (Box 3). Crucially, in both MF and MB frameworks, values are updated depending on the difference between anticipated and received outcomes, or *prediction errors* (PE). Hence, our next question will be whether prediction errors are represented in the amygdala.

Scientists can use MF models to describe behaviour, but more importantly the brain may use both MF and MB action control (Daw et al., 2011). There are important trade-offs between learning through MF vs. MB systems, and paradigms that can dissociate the contribution of each system suggest that people undertake a mixture of both types of learning, and may flexibly adjust this depending on task demands (Kool et al., 2017; Eppinger et al., 2017). MF learning is resource-efficient but statistically inefficient, as action-outcome associations require a lot of experience to develop. MB learning on the other hand, because of its knowledge of task structure, is able to learn quickly and generalise actions to novel situations, but is computationally expensive. In specific cases it is biologically implausible, e.g. due to memory constraints. It is thus of interest to elucidate the role of the amygdala in MF vs. MB control.

We now turn to the questions fundamental to VDM, namely the *representation of value* associated with a situation or an action, and the way that outcomes are observed to deviate

[4]This terminology has been criticised in that the set of action-values in an MF system implies a model of the world. Here, however, we follow the predominant convention in the field.

from expectations, that is, *prediction errors*. We then examine the role of *uncertainty* and the role of *model-based vs. model-free* computations in the amygdala.

Box 3: Key Computational Terms

Model-free and model-based learning. Instrumental conditioning is thought to operate under two modes of control. Model-free learning, in which action values are updated directly according to the PE and model-based learning in which action values are updating according to knowledge of the structure of the environment and the transition probability between states.

Value. Value is not a unitary concept, but depending on the study can be represented in many ways. The most common dimensions of value are magnitude and probability which when used in combination are termed expected value (EV), which is simply the product of the terms.

Prediction Error. The prediction error (PE) is a key concept in computational models of learning. The PE represents the difference between the expectation of value relative to the observed value, R. A positive PE occurs when the observed value is better than expected and a negative PE occurs when the observed value is worse than expected. These are examples of signed PEs, however PEs can also be unsigned, and represent purely the magnitude of the error.

Rescorla-Wagner models of associative learning. Value is updated trial-by-trial according to the signed PE multiplied by a learning rate (Rescorla and Wagner, 1972). The learning rate, often denoted λ, represents the amount by which the PE is weighted when performing the value update. In RW models, the learning rate is stable and does not adjust over time or according to uncertainty, although there can be different learning rates for different kinds of trials, for example positive vs negative words. The same formula applied to the action value $Q_t(action, state)$ rather than $V_t(state)$ is called the 'Q-learning rule'.

$$\begin{aligned} V_t &= V_{t-1}(1-\lambda)+\lambda R \\ &:= V_{t-1}+\lambda PE \Rightarrow \\ \Delta V &= \lambda PE \end{aligned} \tag{1}$$

Pearce-Hall and related models of learning. Value is updated trial-by-trial according to how much attention is paid to the CS, i.e. the associability α_t, which depends on the unsigned PE. If a PE is observed, more attention is given to the relevant CS, boosting learning. A commonly used formula is:

$$\begin{aligned} \alpha_t &= \eta|PE_{t-1}|+(1-\eta)\alpha_{t-1} \\ \Delta V &= S\alpha_t PE \end{aligned} \tag{2}$$

The learning rate thus that dynamically adjusts in proportion to the PH associability variable, α (Li et al., 2011). Equivalently, α can been thought of as gating the impact of PE on learning. When associations have not been adequately learnt, $|PE|$ is large, α increases the learning rate and facilitates learning. Conversely, when $|PE|$ decreases as associations have been better learnt, α adjusts the learning rate downwards.

2.5.1 The Value of Actions and States

Value is an important quantity in VDM (see Box 3) and classically it is calculated as expected value (EV), the combination of outcome probability and reward magnitude of a decision option. Gottfried et al. (2003) provided evidence that the amygdala encodes the current value of representations of reward. The first studies to investigate this using computational models, such as Schiller et al. (2008), concurred that both the striatum and amygdala tracked the aversive value of predictive cues. In a key study, Yacubian et al. (2006) were the first to use both gains and losses to examine EV computation and map their representation in the brain through systematically varying the probability, magnitude and valence of outcomes. In line with previous results, they found that the ventral striatum encoded EV, however this was only the case for reward-related EV. In contrast, the amygdala encoded loss related EV, such that greater amygdala activity represented less aversive punishments.

Value representation in the amygdala is not restricted to aversive VDM, as might be anticipated from the amygdala role in processing stimuli associated with threat. Instead, there is evidence for the computation of both positive and negative value (Baxter and Murray, 2002; Jenison et al., 2011). Some evidence suggests a skewness towards negatively valenced stimuli, as reported in Canessa et al. (2013), where amygdala activity was associated with the magnitude of a single loss but not a single gain. This result is not consistently found in the literature in studies investigating the involvement in single gain or loss calculations (Sokol-Hessner et al., 2013; Tom et al., 2007). When presented together, gains and losses need to be weighted to form the expected value of the decision (see Box 3).

A study by Gelskov et al. (2015) disentangled whether the amygdala performed this computation by tracking the individual magnitudes of monetary gains and losses (Salzman and Fusi, 2010; Canessa et al., 2013), or whether it integrated the magnitudes of both valences into the decision process. They found that the amygdala integrated both gain and loss magnitudes, and that activity increased with this gain-loss ratio. This activity was tuned to a decision boundary parameter, λ, with greater amygdala response the further the gain-loss ratio was away from the decision boundary, thus this can be seen as a signal that increases with value.

VDM models such as Prospect Theory (Kahneman, 1979) offer an important computational framework for understanding the neuro-computational relationships between EV and the amygdala. A key computational parameter derived from Prospect Theory is that of loss aversion. Loss aversion represents the tendency for people to prefer to avoid costs rather than acquire gains of the same magnitude (Kahneman, 1979). There is increasing evidence linking the amygdala to loss aversion and in patients with amygdala lesions, there is a complete absence of a loss aversion bias (Martino et al., 2010). Charpentier et al. (2016b) used a Prospect Theory based computational model to investigate the role of the amygdala in value decision making, using an experimental paradigm that included an emotional manipulation which involved priming participants with happy or fearful faces. Here, they showed that the amygdala was more involved in encoding losses than gains, and that the amygdala could be thought of as encoding a loss aversion signal. This loss aversion amygdala response was amplified in trials where participants were primed with an emotional face, suggesting that the amygdala integrates value representation with emotion signalling – to which we shall return.

However, the evidence on value representation in the amygdala is not fully consistent, and standard neuroimaging analyses may not offer sensitive enough measurements. A detailed meta-analysis of human fear-conditioning delineated a widespread network of brain areas where CS+ activity differed from CS-, but found no amygdala activation or de-activation, despite specifically looking for it (Fullana et al., 2016). Null findings extended specifically to the early stages of CS+/CS- presentation. Taken at face value, this challenges the theory that value representation in the amygdala is particularly important for fear *learning*. Some null results may be due to the interleaved arrangement of neuron types in the amygdala rendering its activity less visible to conventional functional MRI. Thus only powerful (de)activations associated with actually delivered, highly valuable outcomes may be detectable. In support of this, Tom et al. (2007) used an experimental design in which participants had to choose between risky gains and losses, but were not shown the outcomes of their gambles. They were instead told that one of their gambles would be selected for payment. These researchers found no amygdala response to expected value, even at a very liberal uncorrected threshold of $<.01$. This may be an issue especially for secondary reinforcers, as primary reinforcers such as shock or liquid delivery typically elicit stronger responses in the amygdala (Delgado et al., 2011). It is also reminiscent of the relatively weak impact of EV on emotion (Rutledge et al., 2014; Will et al., 2017), which may be important for amygdala function.

Clarification may be offered by recent advances. Only recently has it has become possible to investigate single- and population-neuron encoding of value in the human amygdala. In an important study using an associative learning model in combination with invasive recording in neurosurgical patients, Aquino et al. (2020) found that neural activity tracked the value of both cues and outcomes, and, remarkably, different neuronal populations tracked learning for

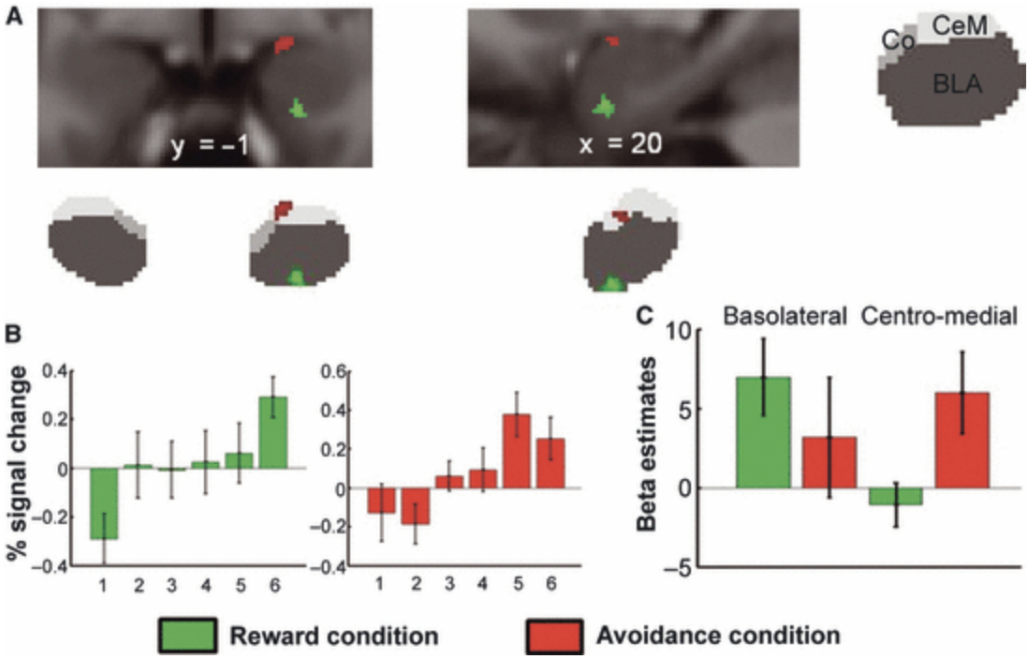

Figure 2.1: Reproduced with journal permission from Prévost et al. (2011). Structural dissociation of action-value depending on valence. (A) Blood oxygen level dependent signals correlating with the magnitude of the expected reward value of the chosen action were found in the basolateral complex in the reward condition (in green) and in the centromedial complex in the avoidance condition (in red). (B) Percentage signal change for six categories of action values (category 1 representing the lowest action values, and category 6 representing the highest action values) in the reward (green) and avoidance (red) conditions in the clusters activated. (C) Beta estimates showing an interaction between the sub-region (basolateral vs. centromedial) and the condition (reward vs. avoidance) ($P<0.05$).

the same stimulus depending on whether learning was experiential (i.e. the participant was rewarded) or observational (someone else was rewarded). This also highlights that amygdala value coding may be important for empathy and social decision-making. Non-invasive neural recordings, namely magnetoencephalography (MEG), also show promise. Using high precision MEG, Tzovara et al. (2019) found that anticipation of threat (vs. safety) elicited lower theta-power activity from both amygdala and hippocampal sources.

2.5.2 Value learning in amygdala nuclei

Amygdala sub-regions are likely to play distinct roles in both state-value (e.g., the value of a cue) and action-value learning. In a recent study using aversive learning, Michely et al. (2020) found that the BLA, but not the CMA, tracked the expected value of aversive cues. Prévost et al. (2011) delineated the contributions of amygdala subregions in both reward and aversive learning. These authors found that in the reward condition, an anterior region of the right amygdala showed increased activity to increasing action-value. In the aversive condition, this correlation was found in the dorsal region (Figure 2.3). Suggestive of a structural dissociation depending on valence, they found that action -values correlated more strongly with BLA activity in the reward condition, whereas the CMA showed a stronger correlation in the aversive condition.

2.5.3 The role of Prediction Errors

One of the most robust discoveries in human neuroscience is the link between the striatum and prediction error (PE) signalling. A significant number of computational studies have correlated striatal BOLD activity, especially within the Ventral Tegmental Area (VTA) with modelled PE signals (Heekeren et al., 2007). The role of the amygdala in PE signalling has not been subject to as thorough investigation and is less clear. Using a PE based value model, Yacubian et al. (2006) was able to dissociate between PE representations in the striatum and the amygdala. They found that, consistent with previous literature, the appetitive PE was located in the striatum. The aversive PE, however, was located in the amygdala, suggestive of a valence dependency between these structures. The balance of these structures in value computation may be particularly important in environments with a mixture of rewards and

losses. A number of studies however, have reported no signal in the amygdala in response to PEs (Seymour et al., 2007).

Rutledge et al. (2010) tested out a formal anxiomatic model of different brain areas with the aim to falsify whether they can represent an RPE signal. This was based on the principle of falsification (Popper, 1959), in which three necessary conditions are specified for an RPE signal. Alongside the striatum, they found that the amygdala, medial prefrontal cortex, and posterior cingulate cortex satisfied all the conditions for the entire RPE model class, meaning the amygdala is a candidate region for RPE signal representation in probabilistic reward tasks. The authors also critically examined the BOLD responses for all candidate regions when fitted to a typical RPE model using a regression based approach, as has often been employed in computational studies (Yacubian et al., 2006; Seymour et al., 2007). Using this approach at liberal thresholds, they failed to find RPE model correlations within the amygdala. As can be seen in Figure 2.2F, the signal from the amygdala does not appear to follow a classical canonical HRF response, in contrast to other regions, which suggests that studies employing a standard HRF response in the amygdala have reduced power to detect true effects that they could detect if using an adjusted HRF function.

Taken together, it appears that the amygdala is an important region for the computation of VDM and is especially important for decisions that include a mixture of reward and loss, such as in mixed gambling paradigms. Here, the amygdala seems to contribute to the weighting of losses more heavily than gains (Charpentier et al., 2016b). Studies have shown that the amygdala encodes both aversive expectations and an aversive PE (Yacubian et al., 2006), such that greater amygdala activity reflects lower punishment. However, the

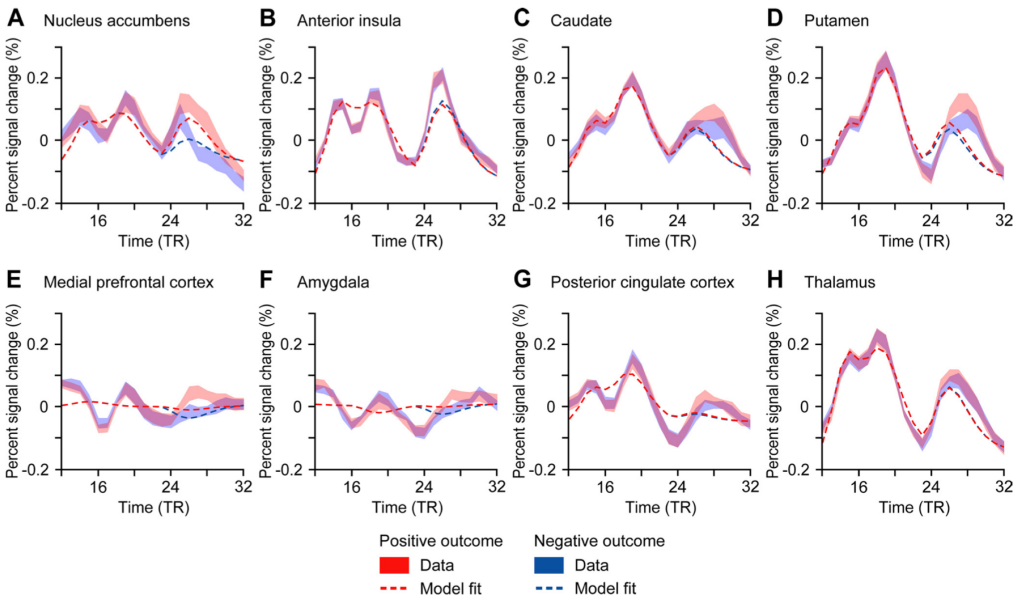

Figure 2.2: Reproduced with journal permission (Rutledge et al., 2010). BOLD responses to positive and negative outcomes in a task including three two-prize lotteries. A–H, BOLD responses for positive (red) and negative (blue) outcomes are plotted, averaged across subjects. Error bands reflect $+/-$ SEM across subjects. Dotted lines represent best fits for a typical regression model with regressors for option onset, choice, and outcome onset, convolved with the canonical two-gamma hemodynamic impulse response function with fits averaged across subjects. We see that the response of the amygdala (F) is poorly described by the standard regression model.

amygdala also satistifes conditions for a reward related PE (Rutledge et al., 2010), but due to non-canonical HRF, studies employing a regression approach may not be able to detect RPE signals. Crucially, amygdala recruitment might rely on actual receipt of outcomes (Tom et al., 2007) and may be especially relevant in contexts that evoke emotions (Charpentier et al., 2016b), as we shall discuss.

2.6 Model-free (associative) analysis of learning

As we have seen, the updating of action-values can often be described through directly associating them with their returns. For example, a 'reward' that follows an action becomes *associated* with it and thereby *reinforces* it, i.e. increases the probability that it will be

chosen. Related computational models have been very successful in describing phenomena associated with the amygdala (Li and McNally, 2014). In VDM, associative models can be used in two ways. The first is *descriptive*: Such models have proven invaluable in capturing patterns of learning, and mapping the relevant expected values and prediction errors to the brain, whether or not the brain uses a model of a mediating sequence of events (a.k.a. 'state transitions').

The second is *mechanistic*, or true 'model-free learning'. Here, the neural machinery, often described in classic Hebbian terms (Johansen et al., 2011) is hypothesized to implement this direct linking between states or actions on the one hand, and their value on the other, without a mediating model of the situation at hand. Both these flavours of associative modelling are important. Using the associative framework, computational work has sought to disentangle the roles of expectation and outcome, and their divergence i.e. the PE, in the learning process. Here, we first use the descriptive framework to map key computational quantities in the amygdala. Then we ask whether the associative framework gives a comprehensive mechanistic account of amygdala function.

2.6.1 Types of prediction error that may be represented in the amygdala

We saw that PEs could play an important role in learning the values represented in amygdala activity, yet the evidence for PE signaling within this structure is not as strong as it is for learning in the corpus striatum (McGuire et al., 2014; McClure et al., 2003; Roesch et al., 2012b). Still, a meta-analysis by Chase et al. (2015) showed that although PE signals were predominately found in the striatum, they also extended into the superficial regions of the amygdala. Importantly, this was only the case in studies using pavlovian conditioning and was not observed with instrumental learning. In other studies, Prévost et al. (2011) reported

PE signals in the amygdala for both reward and aversive learning conditions, while Boll et al. (2013) reported a positive correlation to unsigned PEs. Again using associative models to estimate PEs, Eldar et al. (2016) were able to dissociate better- vs. worse-than-expected PE activity while learning to avoid pain in the striatum, amygdala and periaquaductal grey matter (PAG). Whilst the PAG adjusted its response to outcomes of relief, the amygdala only responded to shock outcomes, suggesting that it represents an aversive PE signal. This was in contrast to the striatum, which tracked both appetitive and aversive PEs. Furthermore, there were important individual differences in the amygdala response to aversive PEs, with a stronger response for individuals who predominately learnt from negative outcomes (shock), compared to those who learnt more from positive outcomes (relief from shock) (Eldar et al., 2016). Significant functional connectivity between the amygdala and insula characterized 'negative learners', while significant connectivity between the striatum and amygdala was found in the 'positive learners'. On the other hand, a majority of relevant neuroimaging studies report no correlation between aversive PEs and amygdala activity (Seymour et al., 2005; Li et al., 2011; Delgado et al., 2008; Schiller et al., 2008; Homan et al., 2019). Moreover, a key human intracranial recording study found no significant representation of PEs in the amygdala (Aquino et al., 2020). The amygdala may have a role in modulating PE activity in the striatum as a function of the neural correlates of emotion that it encodes, rather than directly representing PEs. This was investigated by Watanabe et al. (2013) by presenting an emotional face stimuli before a cue that predicted reward appeared. They observed that compared to the neutral face, an emotion face increased the learning rate by which the cue-reward association was learned.

In summary, there is fairly consistent evidence from associative-learning studies that the amgydala encodes values, with numerous studies reporting increased activity to both appetitive and aversive value (Schiller et al., 2008; Atlas, 2019; Prévost et al., 2011). However, the evidence regarding the representation of PEs in amygdala activity is inconsistent, although it may be important for modulating the impact of PEs in interconnected structures such as the striatum, depending on emotion.

2.7 Attention and associability

The amygdala has long been associated with vigilance and the related constructs of attention and tracking salience. Computationally, salience may increase the associability of outcomes to actions or states, in effect increasing the learning rate. Associative models of the Pearce-Hall (PH) family attempt to capture this by increasing the learning rate λ if large (absolute) prediction errors have recently been experienced (Hall, 1991). The intuition here is that attention must be paid and learning strengthened, when one has got predictions wrong, in any direction (see Box 3).

PEs thus boost associability upon entering an unknown environment, or upon reversal of contingencies in reversal-learning tasks. Both have have provided evidence for a role of the amygdala in modulating associability. Li et al. (2011) investigated amygdala activity during a reversal learning task. They compared different associative models, and found that a model containing an associability term best fitted the data. Associability was significantly positively correlated with amygdala BOLD activity, and was fully dissociable from striatal PE representation. Thus, when α was high, reflecting higher uncertainty or absolute PEs,

amygdala activity was also high. Conversely, when contingencies were learnt well and α was low, amygdala activity was also low. Thus the amygdala may control of the rate of learning, rather than representing the learning signal itself. A limitation of this study was the inability to disentangle whether this amygdala activation arose as a function of PEs, independent of expected value. Boll et al. (2013) addressed this limitation through the inclusion of a pavlovian element to the reversal learning task. Again, a model equipped with associability (a hybrid RW-PH model) provided a better fit than a simple RW model. These authors found that specific sub-regions of the amygdala corresponded to processes with a different experimental time course, derived from their computational model. The corticomedial amygdala showed correspondence with unsigned PEs at the *outcome* stage, indicating a pure surprise signal. However, BLA activity at the *cue presentation* stage increased as the predictability of the stimulus-outcome pairings decreased, showing a negative correlation with associability.

Associability was also investigated by Zhang et al. (2016), who characterised amygdala activity using pain stimuli without reversal learning. The paradigm was a pavlovian conditioning task, in which shocks were administered to either the right or left arm. Again, a hybrid model combining elements of RW and PH updating was found to be the best model. Using model derived associability measures, the authors were able to show that amygdala activity significantly represented associability, however this time showing a positive correlation. The difference in direction of effect to Boll et al. (2013) may be a result of the absolute-PE based heuristic that PH and related models use to capture salience or attention. In essence, these models have no way of estimating whether large PEs occur because the environment is novel (in which case associability must increase) or simply noisy (in which case it should not). In other words, these models cannot distinguish between *unexpected*,

or second-order, informative uncertainty and *expected*, or first-order, uninformative uncertainty. The two paradigms different in several aspects, importantly using different shock reinforcement schedules. Zhang et al. (2016) employed an 85% shock or 30% shock schedule, while Boll et al. (2013) utilised a 50% schedule for some of the conditions. A 50% reinforcement schedule induces maximum expected uncertainty, affecting the fitting of associability models across conditions. Thus in the non-reversal, relatively low noise paradigm of Zhang et al. (2016) associability neatly captures the representation of higher-order uncertainty reflected in contingency reversal, as well as initial learning, unlike Boll et al. (2013).

Another interesting finding of Zhang et al. (2016) was the presence an amygdala-striatal connection specifically learning the preparatory response associated with pain, such as arousal of the autonomic system. This is largely in line with previous literature that has shown that individuals with amygdala lesions have impaired autonomic nervous system responses, resulting in freezing and startle behaviour (LeDoux, 2014).

In summary, associability as formulated in the PH family models is for understanding conceptually the influence of (unexpected) uncertainty on learning, and the role of amygdala activity in the adjustment of learning rates. However, the specific computational formulation of the hybrid PH models currently appears more descriptive than mechanistic, in that relationships found are tied to the particular paradigms, rather than a being reliable property of the amygdala. We suggest that replication studies are needed in order to determine whether these are stable properties of the amygdala. However, the amygdala may also differentiate expected from unexpected uncertainty better than these models assume, relying on more sophisticated, possibly model-based, estimators of uncertainty. If this is the case, PH type

models fitted to data generated by more sophisticated models of uncertainty should reproduce the discrepancy between the two studies discussed here (Palminteri et al., 2017).

2.8 Model-based vs. Model-free information processing

Learning from associations as formulated in MF learning is prominent in studies of how humans are able learn and adapt their behaviour under flexible control. However, MF control is not the only learning system in the brain (Balleine and Dickinson, 1998; McDannald et al., 2012). MB learning is a distinct but complementary learning system, which, unlike learning simply by association, utilises knowledge of the *statistical structure* of the environment to boost learning. In the case of reversal learning, for example, a MB system would reduce learning in the presence of outcomes of relatively high but irreducible uncertainty, would track the probability that the state of the world may have reversed, and only then increase learning rates. This makes for optimal inference and learning, but requires much computational power. A MF system on the other hand requires very little computation, but is inflexible and may take much more experience to learn.

Because of their different advantages and limitations, the brain is thought to operate both MF and MB (or approximate MB) systems in tandem, flexibly switching between the two. Work by Lee et al. (2014) delineated this arbitration, arguing that it is based on an estimate of the reliability of the predictions made by the two systems. Frontal brain regions such as the lateral prefrontal cortex are likely to implement this arbitration, utilising the dense connections to value-related areas to rapidly update decision strategy.

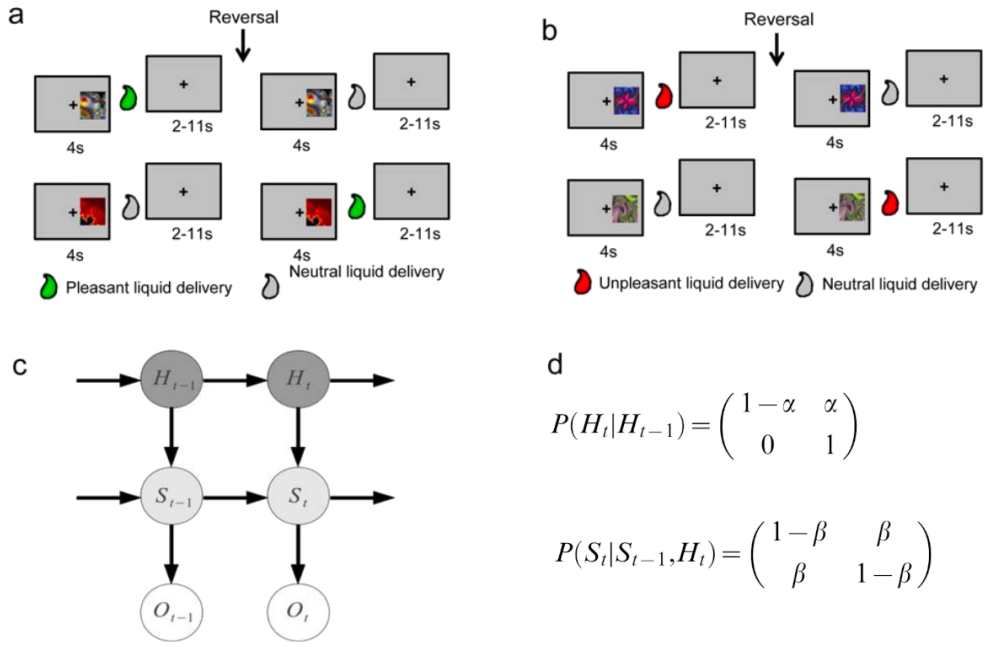

Figure 2.3: Reproduced with journal permission (Prévost et al., 2013). Structure of reversal learning task used in Prévost et al. (2013). Trials belong to either of **a**) appetitive or **b**) aversive type. **c**) Graphical representation of state transitions. There are two high-level states $H = 1$, when reversals can take place, and $H = 0$, where no reversals are allowed. In state $H = 1$, a cue (say, the red fractal in a) may be associated with 60% delivery of pleasant liquid, constituting lower-level state $S = 1$; or 40% pleasant, $S = 2$. In state $H = 0$, the contingency of S is fixed. **d**) Shows the transition probabilities that an MB agent can uses to facilitate learning. Transitions are denoted from row states to column states. β depends on H, so that it is zero for $H = 0$. An MB agent which knows about the structure of transitions and infers that $H = 1 \rightarrow H = 0$ has occured (entry '1' of top matrix in d)), say after a long run of consistent $Red \rightarrow pleasant$, will consider the contingencies fixed and will no longer learn new values for *Red* etc. In contrast, an RW agent will keep updating $Value(Red)$ according to the further rewards received.

As most computational studies utilised MF algorithms in their analysis there is only a limited amount of evidence that the amygdala may indeed be involved in MB learning. In addition, many of the aforementioned studies utilised an experimental paradigm that could have entailed a MB analysis, e.g. through inferring contingency switches in reversal learning (Boll et al., 2013; Li et al., 2011; Schiller et al., 2008), however the relative simplicity of these paradigms might have still favoured MF algorithms. Prévost et al. (2013) performed one of the few studies that have directly examined MB computations in the amygdala. In this study, participants performed a Pavlovian conditioning task, in which they had to learn associations between fractal cues and pleasant, neutral and unpleasant liquid. Importantly, participants were instructed that associations may change, or reverse, and the structure of the task included three probability changes between anti-correlated stimulus pairs. Thus, the structure of this task is such that the MB system would provide a learning advantage following the probability switches, as the MF system would require re-learning from scratch the associations. The authors indeed directly compared these two algorithms and found that a MB algorithm outperformed MF algorithms (Figure 2.3).

Many of the quantities derived from this MB algorithm were strongly associated with amygdala activity. Firstly, at the time of cue presentation, there was increased activity in the right BLA for the appetitive condition and increased activity in the left CMA for the aversive condition, in response to larger EV. Thus, greater amygdala activity was associated with increasing value, irrespective of whether increased value represented larger benefit or reduced cost. Further, a model derived precision value (a.k.a. inverse of uncertainty) was positively associated with increased activity in the bilateral CMA, with a conjunction analysis indicating that this was an overlapping area between the valence conditions. The

superiority of the MB algorithm and the association between amygdala activity and precision suggested that it does have a role in using uncertainty to track reversing contingencies. The precision signal in particular, may finesse the measure of associability derived from the aforementioned MF learning models (Li et al., 2011), which was directly compared in this study. Two uncertainty-related signals were dissociable from one another in the CMA, with one reflecting the expected uncertainty associated with knowledge of task structure, while another, associability-like signal tracked unexpected uncertainty.

Evidence for the possible involvement of the amygdala in model-based inference has also been provided by Wise et al. (2020). They used MEG to identify transitions between state-representations that could lead to an aversive outcome. They found that reactivation of task states in memory underpinned MB learning of aversive contingencies. Reminiscent of the study of Tzovara et al. (2019), they found that during replay of states associated with shock, a cluster of regions including the amygdala, anterior hippocampus as well as classical value-coding regions showed reduced theta power. They concluded that these regions are likely to participate in model-based learning of aversive outcomes.

We would like to emphasize, however, that fractionating uncertainty by using MB systems is computationally expensive, and may not always be implemented in the amygdala. For example, model comparison in the reversal-learning task employed in the aforementioned intracranial study (Aquino et al., 2020) clearly judged in favour of a simple RW model being employed, rather than sophisticated associative or indeed Hidden-Markov MB accounts. One potential explanation is that this study task also required sophisticated social inference,

which may have limited the scope for MB processing.

Study	n	Paradigm	Valence	Stimuli	Model	Parameter association	Time	ROI
Seymour et al. (2005)	19	Pavlovian 1st-order conditioning	Aversive	Capsacin--heat	TD learning	+ve correlation w. appetitive PE (relief)	Outcome	Left
						No correlation with aversive PE		
Schiller et al. (2008)	17	Pavlovian reversal learning	Aversive	Shock	TD learning	+ve correlation w. aversive value	CS	Whole
						No correlation with PEs	Outcome	Whole
Delgado et al. (2008)	14	Pavlovian conditioning	Mixed	Money	TD learning	No correlation with PEs	Outcome	Whole
Li et al. (2011)	17	Pavlovian reversal learning	Aversive	Shock	Hybrid	+ve correlation w. associability	Outcome	Bilateral AMY
						No correlation with aversive PE	Outcome	Bilateral AMY
Prévost et al. (2013)	19	Pavlovian reversal learning	Mixed	Liquid	HMM	+ve correlation w. precision (appetitive and aversive)	Cue	CMA
						Positive correlation EV (appetitive)	Cue	BLA (right)
						Positive correlation EV (aversive)	Cue	CMA (left)
						No negative correlation with EV	Cue	BLA/CMA
Boll et al. (2013)	22	Pavlovian reversal learning	Aversive	Shock	Hybrid	Negative correlation associability	Cue	BLA
						Poisitve correlation unsigned PE	Outcome	CMA and SP
Eldar et al. (2016)	41	Aversive learning task	Aversive	Shock	Adapted RW	Positive correlation with aversive PE	Outcome	Whole
						No correlation w. appetitive PE or relief	Outcome	Whole
Zhang et al. (2016)	15	Pavlovian conditioning	Aversive	Shock	Hybrid	Positive correlation associability	Outcome	Whole

Table 2.1: Key computational studies involving learning and key neuro-computational modelling results. Mixed valence represents studies that present both appetitive and aversive stimuli.

2.9 Uncertainty processing and integration

The computational studies of associability and MB probabilistic decision-making bring new understanding to the extensive non-computational literature which suggests a substantial role for the amygdala in the processing of uncertainty. Patients with amygdala damage are impaired in tasks related to uncertainty and risk (Bechara et al., 1999); though see **?**. Sarinopoulos et al. (2010) observed greater amygdala activity to aversive images following an uncertain cue relative to a certain cue. Under the classic hypothesis that the amygdala is a 'fear centre', a certain cue for an aversive stimulus, would be expected to provoke a greater response, as a negative experience is guaranteed to occur. However, the data were more in line with a vigilance function, in that it was the surprise generated by the uncertainty that modulated the response. The presence of an uncertain cue also produced exaggerated threat response behaviour, as almost 75% of participants reported a higher frequency of aversive pictures following an uncertain cue presentation, indicating that uncertainty influenced the subjective perception of threat. There is also evidence that acute stress interacts with the amygdala response to both threatening and positively valenced stimuli, possibly as salience is enhanced under conditions of stress (van Marle et al., 2009). There is also evidence that the amygdala carries representations of sensory uncertainty, unrelated to instrumental goals (Herry et al., 2007). It is as yet unclear how these different levels of uncertainty may relate together to each other in the amygdala.

These non-computational findings may be understood in the light of the role of the amygdala to modulate the impact of PEs. As we have seen, both PH and MB computational accounts of the amygdala suggest that in the presence of unpredictability or uncertainty, the impact of PEs on learning is increased (Prévost et al., 2011, 2013). It is therefore only

a small step to hypothesize that such a modulation of PEs will have an analogous impact on subjective perception of threat, as PEs have been found to play a crucial role in the subjective perception of reward and social evaluation (Will et al., 2017; Rutledge et al., 2014), consistent with the findings of van Marle et al. (2009). Second, we may hypothesize that a stressful environment makes it more important to avoid errors, and hence there is an increased presumption that the impact of PEs must be increased, thus helping to explain the findings of Sarinopoulos et al. (2010).

It is yet to be ascertained how computational accounts of uncertainty correspond to the rich literature on learning and attention in the amygdala. Specifically, in light of studies such as that of MB control (Prévost et al., 2013), there may be an important functional and structural distinction between expected and unexpected uncertainty. Here we claim that the amygdala has important roles both in the perception of events and in learning associated with the latter, unexpected uncertainty. Its role in aiding decision-making with respect to other levels of the uncertainty hierarchy, including perceptual uncertainty, may be similar but this remains to be tested.

2.10 Affect computation and psychopathology

Perhaps the most prominent label historically given to the amygdala the so-called ‘seat of emotion’ (see Costafreda et al. (2008) for a meta-analysis), suggesting that the computational role of affect in VDM has much to contribute to amygdala research. As a brief overview of the background, amygdala function was first associated with negative emotionality (Craske et al., 2006) Indeed, the amygdala is consistently activated in paradigms eliciting fear and

anxiety (Davis et al., 2010; LaBar et al., 1998; Robinson et al., 2019). However, this focus on negative emotion has been refined in light of a multitude of studies showing amygdala activation to positively valenced emotional stimuli (Sabatinelli et al., 2011; Adolphs, 2010; Janak and Tye, 2015), suggesting the amygdala responds to emotionally arousing stimuli in general (Weymar and Schwabe, 2016). This idea was supported by Bonnet et al. (2015), who systematically varied emotional arousal while controlling for valence and found that amygdala and hypothalamus activity, as well as physiological arousal measured by skin conductance, scaled with arousal ratings. However, not all neuronal activation in the amygdala is valence-independent, and such activation may be part of emotion processing (Russo and Nestler, 2013). Computational studies that make use of subregion analysis may help to inform the debate about valence differences, as the results displayed in Table 2.1 suggest that amygdala sub-regions may be in part dissociable with respect to valence.

A key computational study specifically linking affect to the amygdala suggests a modulatory role of emotion on the computation of value. The role of affect on VDM was investigated by Charpentier et al. (2016b) by including an emotional manipulation that involved priming subjects with happy, neutral or fearful faces before presenting a probabilistic gamble. They showed that the amygdala was more involved in encoding losses than gains, consistent with valence-dependence processing as mentioned above. More importantly, the amygdala encoded a loss aversion signal which was modulated by emotion, in that emotional cues elicited a heightened loss aversion signal. This coincided with heightened striatal-amygdala coupling, but only for low-anxious individuals, who importantly did not differ in baseline loss aversion. The emotional priming here could act as a context cue for the amygdala, which then serves to modulate the subsequent value computation towards more heavily weighting losses. The rela-

tionship to trait anxiety suggests that there are differences in subjective experience in addition to behavioural differences in loss aversion in people with reduced striatal-amygdala coupling.

This modulation can be understood by considering the likely computational utility of emotion for decision-making (Moutoussis et al., 2017). Specifically, emotions elicited by specific features of the stimulus, such as its valence, context, social nature etc. guide the decision-maker as to which scenarios are likely to unfold, depending on her or his actions. Thus, emotion is likely to be an *empirical prior* over such scenarios and the likely best actions to take. In the study of Charpentier et al. (2016b), the emotion induced by the facial expressions acts as the prior over actions, with a higher prior probability allocated to actions that avoid loss outcomes. This would be consistent with avoidance of losses being more important in stressful environments.

As we have seen, Eldar et al. (2016) found individual differences in learning from aversive vs. relief outcomes which were reflected in both activation and connectivity of the amygdala. In the context of emotion, we will consider whether these neural differences reflect neurodiverse experience of emotion, as in the study of Charpentier et al. (2016b). Non-computational studies have consistently linked a variety of mental health problems, such as anxiety (Stein et al., 2007), depression (Siegle et al., 2007) and PTSD (Armony et al., 2005) to dysregulated amygdala activity, and inroads are starting to be made to link those with VDM. In PTSD, where dysregulation often manifests as amygdala hyper-reactivity to emotional or negatively valenced stimuli (Armony et al., 2005; El Khoury-Malhame et al., 2011), Homan et al. (2019) investigated the role of computational mechanisms in a sample of combat trauma veterans. The veterans were asked to perform an aversive reversal learning

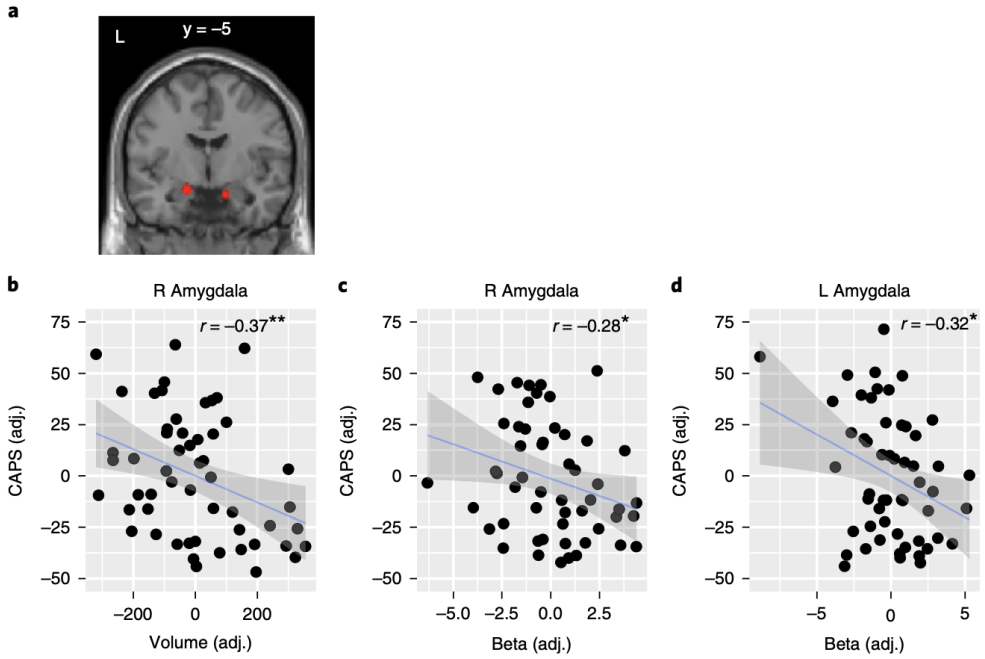

Figure 2.4: Figure reproduced from Homan et al. (2019) with journal permission. Amygdala structure and value computation correlate with PTSD symptoms measured with the CAPS scale. **a**, Region of interest (ROI) used in the computational imaging analysis. The ROI (red) was defined functionally, using the contrast of conditioned stimuli (both CS+ and CS–) versus baseline. **b–d**, Amygdala volume and value-dependent neural activity independently correlate with PTSD symptoms. Pearson partial correlations are shown (N = 54). Right amygdala volume and activity as well as left amygdala activity correlated negatively with PTSD symptoms. Lower value tracking in the amygdala and smaller amygdala volume correspond to higher symptom severity. Regressions were adjusted ('adj') for several covariates. Error shadings correspond to standard errors. adj., adjusted; **P <0.01; *P < 0.05., all two-tailed

task in which they would receive shocks as negative outcomes. Using a hybrid computational model, the authors were able to track trial-by-trial estimates of value and associability. They found no relationship to PE tracking, but they did find that for people with higher PTSD symptoms, there was a weaker amygdala response to value. This effect was independent of the fact that the volume of the (right) amygdala was also smaller in people with PTSD (Figure 2.4) and significantly predicted greater symptom severity.

Psychopathy has classically been associated with abnormally low fear, with reduced learning from aversive outcomes and blunted amygdala responses to aversive stimuli (Birbaumer et al., 2005). This classic view of psychopathy is related to the idea that moral decision-making, which is impaired by definition in psychopathy, is largely based on (learnt) fear of punishment. Indeed, one of the earliest computational (but not imaging) studies provided evidence that a specific deficit in learning from loss, as opposed to reward, was necessary to reproduce *in silico* the learning patterns of people with psychopathy (**?**). However, the low-fear theory of psychopathy has been challenged. Amygdala reactivity and fear was found to be greater, not lesser, in those with *low-anxious* or 'primary' psychopathy, as opposed to those with *high-anxious* (secondary) psychopathy during fear learning employing shocks[5] (Schultz et al., 2016). However, the above studies used diverging methodologies and small samples, while sophisticated computational analyses of amygdala function in psychopathy and related conditions are lacking.

2.11 General discussion

Our understanding of the unique role of the amygdala in value-based decision making has benefited from computational methods that seek to understand the mechanisms employed when making computations. In this review we saw rich evidence that the value of perceived states of the world is represented in the amygdala, especially in the context of adverse outcomes, that its activity may be modulated by uncertainty important for decision-making, partly by influencing the associability of stimuli, that it is activated during learning well-described by associative models, yet it may also be involved in sophisticated MB computations. Evidence

[5]Indeed, there is no evidence for acquired psychopathic behaviour in patients with early extensive amygdala damage despite serious fear deficits.

is mixed as to whether PE signals are represented in the amygdala. Neurotypical emotionality may influence its encoding of value, while many psychiatric disorders prominently involving emotion, from post-traumatic stress to psychopathy, may disrupt its role in value-based decision making. We now discuss the implications of these findings.

Value signalling

We saw that the amygdala is involved in encoding expected value even in the absence of learning, especially in the aversive domain (Yacubian et al., 2006). This adds to the established literature for value representation in the ventral striatum, the putamen, vmPFC and other 'reward network' areas (Adrián-Ventura et al., 2019), even in the absence of learning . Furthermore, the amygdala appears to encode a loss aversion signal which informs the value of outcomes, and this is amplified in the context of emotional stimuli (Charpentier et al., 2016b). However, value-dependent amygdala activation is not robust across all experiments. It appears to be sensitive to the specific paradigm used, as some studies reported null results (Seymour et al., 2007; Tom et al., 2007), perhaps because the rewards and losses utilized are not powerful enough to engage amygdala activity detectable by conventional neuroimaging. A complementary explanation is that contrasts may be stronger in situations where a change of decision or action is needed. It is important for advanced functional imaging and MEG paradigms to strive for replication, and to delineate the circumstances under which each technique can be relied upon to measure value-based quantities in the amygdala. Then, research must clarify whether the amygdala preferentially encodes values within specific contexts, such as with respect to primary vs. secondary reinforcers, or the need for skeletomotor or

visceromotor action.

There is fairly consistent evidence supporting the amgydala's role in value computation during AL, with numerous studies reporting increased activity to both appetitive and aversive value (Schiller et al., 2008; Prévost et al., 2013; Atlas, 2019). In all studies, the direction of the relationship is positive, indicating that greater amygdala activity is related to cues that signal more positive outcomes. This is in line with invasive studies of value representation in the amygdala, but not with the null findings of the meta-analysis of Fullana et al. (2016), which require clarification. Overall, it appears that the amygdala is tracking value in a similar manner to the striatum, which also shows a positive relationship with value computation. This also supports the notion of valence dependency in expected value calculation, at least at the cue presentation stage. This picture raises further interesting questions for further research. Namely, how does the amygdala calibrate its responses, especially to less-aversive vs. more-aversive outcomes? Its connectivity with structures such as the hippocampus, which may provide a memory-based model of the local situation, as well as with the 'reward network', may play a role here. The robust evidence from invasive studies suggests that value pertaining to the self vs. to other agents needs study.

There is inconsistent evidence surrounding the specific role of the amygdala for the computation of PEs in simple learning tasks, both appetitive and aversive. The majority of studies report no correlation between aversive PEs and amygdala activity (Seymour et al., 2005; Li et al., 2011; Delgado et al., 2008; Schiller et al., 2008; Homan et al., 2019). However, Yacubian et al. (2006) and Eldar et al. (2016) did find evidence of aversive PE signaling in the amygdala. Boll et al. (2013) also reported a positive correlation, but this time to unsigned

PEs, suggesting that the amygdala is sensitive to magnitude of error rather than valence per se.

Methodological considerations become important in the case of measuring activity related to PEs, as direct neural recordings in humans have not, so far, furnished good evidence that PEs are represented in the amygdala. A significant issue is the reported non-canonical HRF that the amygdala displays during a classical probabilistic decision making task (Rutledge et al., 2010). Using a standard regression approach, the authors found no significant amygdala activity in response to PEs. However, an alternative, 'axiomatic' approach indicated that the amygdala does satisfy conditions required to represent PEs. This suggests that some of the studies reporting null results could be underpowered, though studies reporting positive results require replication.

Few of the studies in this review decomposed the amygdala region into subregions. Different subregions of the amygdala may perform different roles in the learning process (Michely et al., 2020). Prévost et al. (2013) performed a thorough subregion analysis, and found a valence dissociation between the BLA and CMA at the time of cue presentation, with the BLA representing positive EV in the appetitive domain and the CMA representing positve EV in the aversive context. The CMA also had a positive correlation with precision for both appetitive and aversive contexts, suggesting the CMA plays a role in uncertainty monitoring. Structurally, the BLA and CMA differ in their predominate afferent/efferent connections to other brain regions.

Salience, uncertainty and model-basedness

The reason for the discrepant findings regarding PEs is not clear, but may relate to the ways in which amygdala activity may engage in MB computations to estimate different types of uncertainty based on PEs. A crucial differentiation is that between irreducible or expected uncertainty on the one hand, which results in uninformative PEs, and unexpected or reducible uncertainty in the other, giving salient PEs. Thus unexpected uncertainty is salient, should enhance learning, and the amygdala may have an important role in reporting it. In associative models, PE salience is approximated by the 'associability' variable. This has been detected in the amygdala, especially in studies that use non-stationary action-outcome mappings, such as reversal learning (Boll et al., 2013; Li and Daw, 2011). This underlines its possible computational role in modulating learning in down-stream brain regions. However, sparse but intriguing evidence indicates that the amygdala may in non-stationary situations employ sophisticated, MB computations to estimate when attention and learning should be boosted (Prévost et al., 2013). This could explain discrepant findings between studies which used standard associative models and found different directions of correlation of amygdala activity with associability (Boll et al., 2013; Zhang et al., 2016; Li et al., 2011). Further research, including replication of the MB findings in invasive studies, should delineate more clearly when the amygdala performs MB processing and especially uncertainty estimation. Hierarchical paradigms (Mathys et al., 2011; Berker et al., 2016) that can effectively discriminate between different levels of uncertainty to provide a robust way of assessing their independent contributions towards learning would be key. The finding that amygdala activity may reflect not only the statistical structure of uncertainty – expected, unexpected – but also the cognitive level whence it derives – perceptual (Herry et al., 2007), inferential etc. – suggests that future research should look carefully at its role with respect to these different cognitive levels

during decision-making. Theoretical work has promoted the idea of an uncertainty hierarchy, in which distinct but inter-related levels of uncertainty present in the environment need to be integrated in order to support optimal decision-making (Bach and Dolan, 2012; Bach et al., 2011). Studies should first examine the role of the amygdala in how perceptual uncertainty, especially that associated with salient, aversive events, may influence decision-making.

Emotion, subjective experience and symptoms

Theoretical and experimental considerations suggest that amygdala research may become a privileged arena for the computational understanding of emotion. Numerous studies have demonstrated that the amygdala is very sensitive to emotionally valenced stimuli, that its response begins before conscious awareness is established, but also that it may be important for the subjective, conscious experience of emotions such as socially induced fear (Calder, 1996). Thus affect must be understood as a complex, whole-brain phenomenon with components ranging from non-conscious perception and interoception, to cognitive-linguistic categorization (Smith et al., 2018). The connection between the amygdala and physiological measures of arousal has been reported in a number of studies and future computational work could seek to understand the mechanistic connection. Computational modelling of affect is not as advanced as that of overt behavioural choices, however computational modelling research has started to integrate conscious affect directly into the computations of PE, and also value (Rutledge et al., 2014; Will et al., 2017). This is *prima facie* consistent with results from non-learning paradigms, such as mixed gambling tasks. Here, it has been suggested that the amygdala may report subjective evaluation and experience of the outcomes of decisions, such that the amygdala acts as a bridge between the perception and experience of emotion

(Anderson, 2007). Intriguingly, early findings suggest that the amygdala may use emotion to refine the subjective value of outcomes, in this case by adjusting loss aversion (Charpentier et al., 2016b). These findings encourage a view whereby multiple amygdala afferents give rise to low-level components of emotion, which 'set the scene', or technically provide empirical priors, for value-based decision-making and affective experience.

It is interesting to hypothesize that neurodiverse amygdala function will be causally related both to different styles of inference and learning, and to the experience of emotion as has already been established behaviourally (Koban et al., 2017). We hypothesize that people reporting higher anxiety on questionnaires (Spielberger, 2010) may be the ones who preferentially learn from negative outcomes (Eldar et al., 2016) in an aversive context, whose amygdala responds more strongly to aversive PEs, and who have stronger functional connectivity between the amygdala and insula (as opposed to 'positive learners', found to have significant connectivity between the striatum and amygdala). We also hypothesize that such individual differences will be detectable in moment-to-moment subjective evaluation of threat, as has been modelled for other affective states (Rutledge et al., 2014; Will et al., 2020). Finally, we note the opportunity that the increasing sophistication of value-based computational studies affords to illuminate amygdala function in conditions such as PTSD, psychopathy, and many others.

2.12 Conclusions

We hope that this review assists researchers interested in further understanding the computational role of the amygdala. Computational research is needed to delineate between functional

hypotheses that share similar behavioural or neurological signals and will be instrumental for disentangling interdependent concepts such as value and surprise. First, however, discrepant findings need replication, especially with respect to value and prediction error signaling. We envisage that advances in invasive, magnetoencephalographic and high-field functional neuroimaging will be needed to obtain robust, replicable results. There is a clear need for tasks to discriminate between model-free and model-based learning in the amygdala. Instrumental learning, not just pavlovian conditioning. The regional specificity of this research may prove to be a key factor in understanding the discrepancies in the literature, for example about valence. Thus, further research should investigate the structural dissociation between appetitive and aversive expectation tracking. Furthermore, research that incorporates the multifacted nature of uncertainty are required, as is how uncertainty bears upon decision-making and with respect to neurotypical, neurodivergent and maladaptive emotion. Going forward, research needs to establish what computations the amygdala *performs* rather than which variables its activity merely reflects.

3 Chapter 3: Sensory uncertainty and threat perception

3.1 Acknowledgements

For providing guidance on the development of this experiment and for critical feedback on the design, I would like to give my thanks to Karl Friston and Tobias Hauser. For guidance on the neuroimaging analysis, I would like to thank Peter Zeidman and Jochen Micheley. This work is currently being written up in preparation for submission as a research article.

3.2 Abstract

Organisms have evolved to integrate various forms of uncertainty into their learning and decision making, however how different levels of uncertainty influence each other is still largely unknown. In this experiment we examined how sensory uncertainty is processed in the brain under context of threat and how this relates to outcome uncertainty and subjective threat perception. We were especially interested in whether higher sensory uncertainty modulated the response to threat related outcomes in brain regions implicated in uncertainty processing, such as the amygdala and insula. We further sought to explore whether this response was heightened for people who have a greater intolerance of uncertainty and trait anxiety. Our novel experiment involved participants learning the outcome probabilities of accumulating an aversive electric shock for four unique stimuli. Crucially, we varied the spatial distribution of the outcomes to reflect either a more or less unpredictable display. We combined functional magnetic resonance imaging (fMRI) with a novel hierarchical Bayesian model in order to investigate the neural representation of trial-by-trial beliefs about the expectation of shock and the uncertainty of those expectations. We observed that responses to stimuli with greater

sensory uncertainty was encoded in the insula, but only for people who are highly intolerant of uncertainty. This enhanced uncertainty encoding corresponded to greater uncertainty in subjective threat perception for more uncertain, low probability stimuli. Our results highlight the role of the insula in uncertainty processing, and that individuals who self-report that they find uncertainty aversive have an elevated response pattern in this brain region.

3.3 Introduction

Theoretical work (Bach and Dolan, 2012; Bach et al., 2011) (introduced in **Chapter 1**) has suggested that uncertainty can be disambiguated into multiple different levels that form layers of an uncertainty hierarchy. Uncertainty integration is needed to form accurate estimations of events in the world and is crucial for learning (Grupe and Nitschke, 2011). Outcome uncertainty in particular, has been widely studied and related to different behavioural and neural signatures (Bach and Dolan, 2012) and is thought to be prominently related to anxiety (Charpentier et al., 2017) and stress (Berker et al., 2016). However there is evidence that low-level processes such as environmental or sensory uncertainty may play a role in behaviour, even if that uncertainty is independent to the task process (Lake and Labar, 2011; Herry et al., 2007; Bach and Dolan, 2012). Previous research has indicated that sensory areas, such as the occipital cortex, show higher activation in correspondence with higher sensory unpredictability (Vilares et al., 2012). How this uncertainty is related to perception of threat however and anxiety however, is unclear.

A study by Herry et al. (2007) sought to elucidate the role of temporal unpredictability on behaviour. Using evidence from both mice and humans, they found that sound sequences that were more unpredictable elicited more aversive subjective feelings and anxiogenic

behaviour in both animals and humans. This was related to increased amygdala activity for unpredictable tones, even though the unpredictability of the sequence itself was not instrumental for the task. As the amygdala is a structure commonly associated with fear and anxiety (Davis et al., 2010; LaBar et al., 1998), this study bridges understanding of how unpredictability *per se* may relate to anxiety like processes. Given that anxiety is often characterised by an increase in feelings of dread, one might expect that unpredictability lends itself to an increase in subjective perception of the likelihood of threat as well as threat avoidant behaviour (Bishop and Gagne, 2018; Mobbs et al., 2007; Grupe et al., 2013).

Another important brain region implicated in the processing of uncertainty is the insula. In a study by Shankman et al. (2014), they manipulated the temporal unpredictability of aversive images and observed heightened insula activity in anticipation of the stimuli, suggestive of insula involvement in monitoring outcome uncertainty for aversive outcomes. Individual differences in intolerance of uncertainty were also reflected in greater insula activity to ambiguous faces (Simmons et al., 2008). This result was specific to intolerance of uncertainty and not other measures such as anxiety sensitivity or neuroticism, highlighting the link to subjective uncertainty perception and insula activity. We wish to further clarify the role of amygdala and insula in subjective threat perception by directly asking participants beliefs about threat probability, leveraging both choice behaviour and subjective reports.

One way of studying threat perception is to use associative reinforcement learning, which quantifies the value associated to different stimuli acquired through learning. Alternatively, Bayesian models can give optimal (but harder to compute) estimates of how uncertainty evolves over time. These methods allow the impact of uncertainty to be quantified in model-

based fMRI analyses (Kruschke, 2008). We hypothesized that under conditions of threat, induced by a paradigm framed around avoiding electrical shocks, subjects will show a heightened threat perception of more unpredictable stimuli, over and above its threat likelihood. We expected that this is related to heightened neural activity in sensory processing areas such as the occipital cortex (Vilares et al., 2012), threat related areas, such as the amygdala (Grupe et al., 2013) and areas involved in uncertainty processing, such as the insula (Limongi et al., 2016; Monosov, 2017). We further expected that individuals who differ in intolerance of uncertainty and trait anxiety will show a heightened response in these areas, indicative of an enhancement of uncertainty related processing (Simmons et al., 2008).

We first aimed to investigate the computational and neural representations of sensory unpredictability, and to understand their relationship with threat perception. We investigated the hypothesis that sensory unpredictability would impact threat perception in an aversive learning paradigm. This gave participants the goal of minimizing the number of electrical shocks that they would receive. We found evidence for individual differences in uncertainty aversion, with people that had higher trait intolerance of uncertainty perceiving higher threat levels from more unpredictable stimuli for low threat conditions. These individuals also displayed greater neural activity in the insula. We modelled participant assumptions and learning about uncertainty and probability of shock, as a function of perceptual uncertainty. We found that people activated different assumptions about threat in the context of greater uncertainty, modulating learning.

3.4 Methods and Materials

3.4.1 Participants

Forty-five participants (age (M = 28, SD = 9.6), 24 female, 2 left-handed) were recruited from the University College London Psychology Pool (SONA). Inclusion criterion for the study was age (min = 18, max = 60). Exclusion criteria were colour blindness, as the task required one to discriminate between different coloured stimuli, and cardiac or endocrine problems. This sample size was the maximum within available resources and was therefore chosen pragmatically with economical considerations. All participants were remunerated for their time, plus a performance-dependent bonus (20-25 UK pounds). Ethics were obtained through University College London local research ethics committee (ref: 9787/001) and informed consent was acquired before the experiment.

Two participants were excluded for failing to understand task instructions properly (revealed upon debriefing). Four further participants were excluded due to failure of electrical stimulation administration. The final sample included in the behavioural analyses was therefore 39 participants (age (M = 27.8, SD = 9), 23 female, 2 left-handed). One further participant was excluded from the fMRI analyses due to excessive motion, therefore the final fMRI sample consisted of 38 participants.

3.4.2 Materials and Stimuli

The experimental materials were programmed in Matlab and presented using Cogent 2000 (www.vislab.ucl.ac.uk/cogent.php;). The stimuli consisted of 4 uniquely coloured roulette wheels (Fig. 3.1). For each wheel, half always consisted of a red coloured shock zone, which indicated shock accumulation, the other half consisted of the unique colour. The colour of

the wheel and the red shock zone were matched for luminance and were displayed against a gray-patterned background to control for luminance-related effects on both pupil diameter response and occipital cortex activation (Eldar et al., 2016). One 4-choice button box was used to record responses during the task. Before the experiment, in order to assess uncertainty and anxiety, participants completed the Spielberger Trait and State Anxiety questionnaire (STAI)(Spielberger, 2010) and the intolerance of uncertainty scale (IUS) (Carleton et al., 2007).

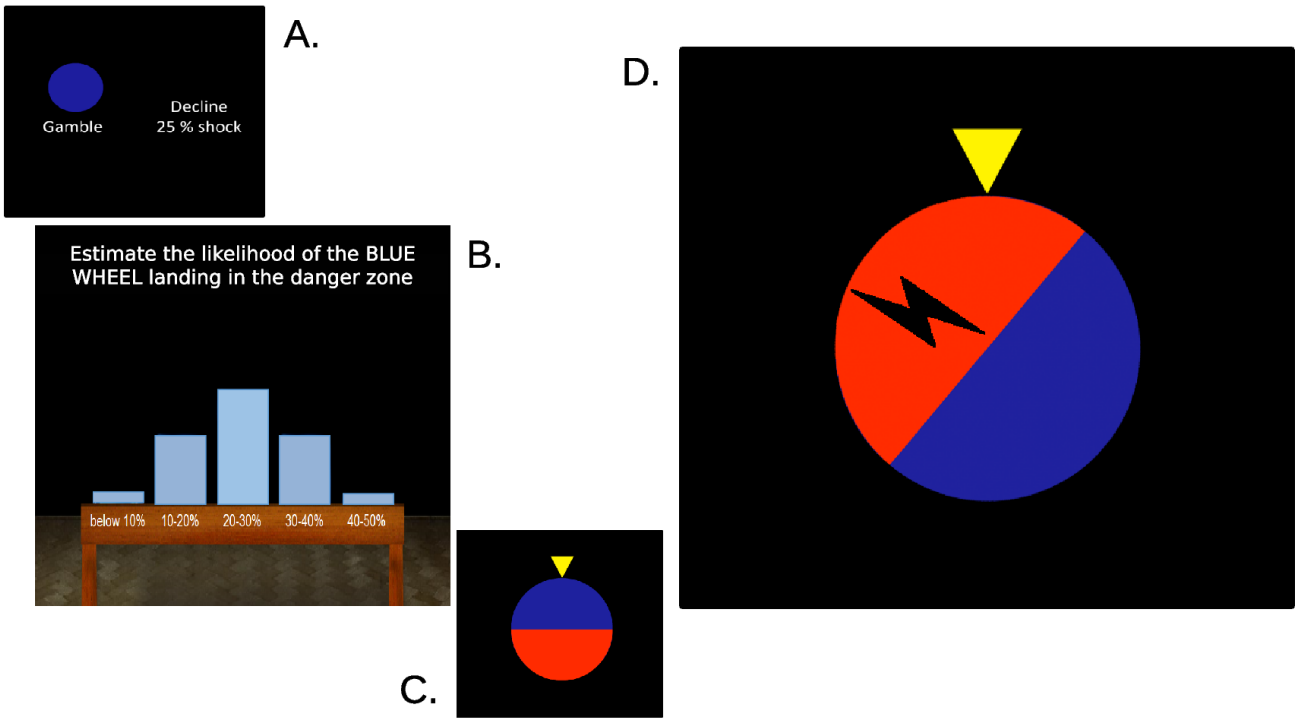

Figure 3.1: **Experimental task** trial structure (A.) participants chose between a gamble of unknown probability of shock outcome, and a known 25% probability of shock outcome. Four roulette 'wheels of misfortune' were used, the example here is in blue. (B.) Every third trial, participants were asked for their beliefs about the adverse outcome of the current unknown roulette. After declaring beliefs, the roulette wheel appeared on the screen and began to spin (C.). Participants then saw the outcomes of both chosen and unchosen options (D.). In this example, a 'shock' outcome is shown, as the roulette stopped on the red ('shock' zone).

3.4.3 Experimental task

In the task, participants observe a roulette wheel that would spin around and stop spinning at a certain position, either within a red coloured shock zone, or within one of the unique colours. If the wheel stopped within the shock zone, participants would accumulate half a shock to be administered at the end of each block. Sensory unpredictability was investigated by manipulating the mean and variance of the Gaussian distributions that controlled where the roulette wheel outcomes would be displayed. Importantly, the distributions were matched according to their overall probability of shock accumulation. Thus there were 4 conditions corresponding to the unique roulette wheel colour, that varied along two independent factors – probability of shock accumulation (Low, p = 0.15 / High, p = 0.35) and variance of the distribution (Wide/Narrow). The mean Gaussian distances away from the threshold were 90, 25, 70 and 15 degrees, with SDs of 70, 30, 70 and 30 degrees respectively. This created the four conditions, that differed in terms of probability and variance 'LowWide', 'LowNarrow', 'HighWide' and 'HighNarrow'. We used samples drawn from these distributions which were checked to ensure they were representative of the generative distribution. As the mean and standard deviation of the distributions are interlinked, we carefully selected the order of observations presented to de-correlate the estimates of the mean and standard deviation at each point in time. Ten thousand randomised sequences were generated and the correlation between the mean and standard deviation was calculated for each sequence. The sequences that had Pearson correlation $-0.01 < r < 0.01$ were chosen and put into a pool of sequences that would be randomly drawn from for each participant. The sequences were also checked for an even distribution of shocks across the 4 blocks.

The task consisted of 4 blocks of 40 trials each, totalling 160 trials. Participants could minimize the amount of shocks they would accumulate by correctly learning the probability that each roulette had of stopping in the shock zone. At the beginning of each trial, participants were presented with one of 4 possible roulette wheels, represented with a unique colour. Each trial started with a choice between whether to gamble on the roulette wheel, or whether to decline. Participants had 4 seconds to make their choice. The majority of participants did not have high rates of missed trials (median = 2), with the range being 0-24 and with no participants missing more than 15% of trials. All missed trials were excluded from further analyses. If participants chose to gamble, after a short interval (2-4s, uniformly distributed), the roulette wheel would be displayed. Every third trial, participants were first asked to rate their current belief about the probability of the wheel stopping in the shock zone and had 4 seconds to make their choice. They did this by moving bars that corresponded to different probability bins (Figure 3.1B.). Participants could also indicate their certainty by making the bars higher (indicating greater certainty) or lower (indicating lower certainty). Participants were incentivized by instruction that they would receive a monetary bonus the more accurate their probability ratings were relative to the true underlying probability. After another short interval (2-4s, uniformly distributed), the roulette wheel was displayed on the screen for 2s, before spinning for 2-4s (uniformly distributed), after which the wheel stopped and the outcome was displayed for 3s. If the wheel stopped anywhere within the shock zone, participants would accumulate half a shock to be administered at the end of each block. If the wheel stopped outside of the shock zone, they would avoid accumulating any shocks. If participants chose to decline, they would accumulate half a shock, usually with a fixed probability of 0.25, which would be indicated by the words ‘shock/ no shock’ displayed on the screen. We also included a small number of catch trials, presented after every four

presentations of each condition. The catch trials were included to present the decline option as the most obvious best choice. The catch trials were derived from half of the reported participants beliefs about each roulette probability. Thus, if the previous reported probability was 20-30%, the catch trial decline option would be 15%, thus the participant should choose to decline. Participants would still be shown the outcome of the gamble so that they can still learn about the probabilities. The first nine participants completed a version of the task that did not include catch trials and thus the decline option was always set to 25%. At the end of each block, the total number of shocks to be given was displayed, alongside a message that indicated the shocks would be delivered after 2 minutes, after which the total number of shocks would be administered. Upon completion of the experiment, participants were asked to rate the overall probability that each roulette had of landing in the shock zone, given in percentage form.

3.4.4 Electrical stimulation

Electrical stimulation was given via an electrode placed on the back of the participant's non-dominant hand using a Digitimer DS7a electric stimulator (Welwyn Garden City, UK). The titration process was done before participants were taken inside the scanner bore, but while they were inside the scanning room in order to best account for the final shock delivery environment. Pain intensity was calibrated to each participant's pain threshold using a well-established titration procedure (Eldar et al., 2016; Michely et al., 2020), which was explained thoroughly before the calibration began. Participants were asked to rate the intensity of the pain by giving a number ranging from 0 = cannot feel anything, to 10 = maximum can tolerate. The titration procedure was initialised with low-current (0.1 mA) shock and then

gradually the intensity was increased using small incremental milliamp changes (0.25mA). Participants gave subjective pain ratings after each shock, until a rating of 9 or 10 was reached, in order to find the maximum pain threshold. The final shock intensity was given as a 7 out of 10. As participants did not receive shocks during the experiment, habituation to shock was not expected, however participants were asked to rate the shock intensity after each block in order to check there were no changes due to differences in environment inside the scanner bore. If participants indicated a large reduction/increase in pain intensity, a short recalibration was done, in order to reach the same 7/10 shock level. Mean shock intensity across subjects ranged from 0.6 to 6.7 mA, with a mean of 2.11 mA (SD = 1.16). There was no relationship between shock amplitude and reported trait anxiety.

3.4.5 Choice modelling

Associative learning models. We first fit AL models to the data. These allowed us to parameterize intuitively the impact of the perceptually certain/uncertain conditions on the choice process. Here, participants learnt the value of the action a, based on feedback (accumulated shock or not). Action-values $Q_t(a,s)$ were updated after each outcome r_t, according to a prediction error multiplied by a time-independent learning rate. The state s indexed which wheel was updated at each trial. We allowed different learning rates for learning from good and bad outcomes (α_{shock} vs α_{relief}) and also fitted models with different learning rates for the narrow vs wide distribution wheels (α_{wide} vs α_{narrow}). Crucially, we also explored a 'certainty boost' parameter, which had the effect of boosting the correct choice option for the two narrow variance conditions. Thus, each associative model took into account, the action, gamble or decline, $a \in \{gamble, decline\}$, the roulette ($p \in \{lowwide, lownarrow, highwide, highnarrow\}$) and

the reinforcement value of the outcome, $r \in \{+1 = noshock, -1 = shock\}$ on each trial t. The shock/relief model contained separate learning rates for shock and no shock outcomes. τ is inverse decision temperature parameter (eq. 3).

$$
\begin{aligned}
Q_t(a_t, s_t) &= Q_{t-1}(a_t, s_t) + \lambda_{relief}\ (+1 - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad r_t = +1 \\
Q_t(a_t, s_t) &= Q_{t-1}(a_t, s_t) + \lambda_{shock}\ (-1 - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad r_t = -1
\end{aligned} \tag{3}
$$

The certain/uncertain model contained separate learning rates for the wide/narrow distributions (eq. 4).

$$
\begin{aligned}
Q_t(a_t, s_t) &= Q_{t-1}(a_t, s_t) + \lambda_{certain}\ (r_t - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad s_t = narrow \\
Q_t(a_t, s_t) &= Q_{t-1}(a_t, s_t) + \lambda_{uncertain}\ (r_t - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad s_t = wide
\end{aligned} \tag{4}
$$

The initial bias parameter, q_0, allowed the starting expectations to vary between -1 and 1. This is applied to both the wide and narrow distributions and impacts the first trial for every roulette. The gambling bias, β, is applied to each round of the expectation update as a constant term which is allowed to vary from -1 and 1, values above 1 indicate a gambling bias, whereas values below 0 indicate a decline bias. Again, this is applied to both wide and narrow distributions.

Response function: Actions were chosen probabilistically, as a function of a propensity variable for choosing each action. For models containing the certainty boost parameter, this propensity was the action value $Q(a,s)$ biased by a 'certainty boost' ρ, which quantified

biases in favour of choosing the more certain stimuli independently of learning (Eq. 5). For models without this parameter, the response function was simply the action value $Q(a,s)$ without any additional biases (the same as for the wide distribution conditions). $Q(a,s)+\rho$ then entered a standard softmax function, weighed by a 'decision noise' parameter $\tau > 0$:

$$\begin{aligned} P(a = narrow) &= z\, exp\frac{Q(a,s)+\rho}{\tau} \\ P(a = wide) &= z\, exp\frac{Q(a,s)}{\tau} \end{aligned} \tag{5}$$

Ideal Bayesian Observer models: We compared associative models with a novel Hierarchical Bayesian model that quantifies sensory uncertainty. This model was adapted from the standard conjugate-prior model (Figure 3.2) of inferring the parameters of a one-dimensional Gaussian distribution from samples (Gelman et al., 2013; Haines, 2011). This allows quantification of optimal (ideal Bayesian observer) estimates of sensory uncertainty and threat probability. The model frames learning as updating (upon each new observation) beliefs about the mean and variance of the output of the wheel generating the data. Participants were taken to correctly assume a Gaussian generative distribution. If the wheel were to stop at an angular position x sampled from such a Gaussian distribution, the posterior beliefs about x then become the familiar t-distribution as per equation 6 (see also Supplement, and Figure 3.2). Denoting the scale of the t-distribution after the j^{th} observation as s_j, its mean as μ_j and its degrees of freedom as n_j, we have:

$$p(x) = \frac{1}{s_j}\, t(\frac{x-\mu_j}{s_j}; n_j) \tag{6}$$

The probability of the wheel landing in the shock zone is then the mass of the distribution that falls beyond the 'shock boundary', i.e. the cumulative probability $x < 0$, i.e.:

$$p_s = \int_{-\infty}^{0} dx\ p(x) \tag{7}$$

Modified Bayesian Observer models According to the core hypothesis of the study, we considered that prior beliefs about the dispersion and location of variable outcomes may be correlated with each other, e.g more variable lotteries being more dangerous. To capture this, we allowed the model to utilise different prior belief parameters for the 'wide' and 'narrow' lotteries. In addition, we allowed for a 'memory decay', whereby old observations gradually decay. A memory parameter m resulted in forgetting information of the order of more than $N_{max} = 1/(1-m)$ observations ago (see Supplementary Information).

Response function The policy π for action $\in \{gamble, decline\}$ was then specified by comparing current beliefs about the probability of accumulating half a shock in the 'gamble' wheel vs. the known probability of the 'decline' lottery, via a softmax function:

$$\pi(gamble) \propto \exp \frac{P_{decline} - p_s + \rho_{choice}}{\tau} \tag{8}$$

Here, ρ_{choice} represents biases. In the simplest case it is a constant 'gambling bias' favouring the unknown lottery. τ is the decision temperature. All models could include an initial bias I, which allowed a separate policy for choosing at the first trial ($t = 1$) only. Each model could also have the gambling bias parameter ρ_{choice} above.

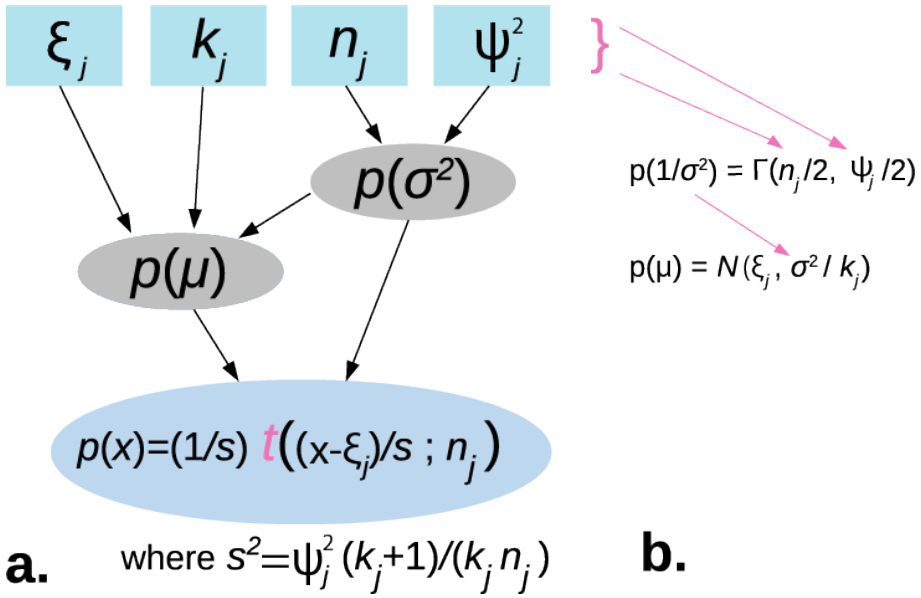

Figure 3.2: One-trial hierarchical Bayesian generative model for trial j. a) At the bottom level, the posterior belief about the probability of shock is parameterised by a Student's t distribution. This is informed by the level above, formed of a Gaussian distribution of the beliefs about the distribution of the outcomes. p(μ) is not the belief about the mean given all the known parameters, but given a sample drawn from the precision-distribution. b. How to generate a sample of σ^2 and μ.

3.4.6 Neuroimaging data acquisition

We acquired the data from all participants using a Siemens 3T Trio scanner with a 64-channel head coil. In order to maximise signal and minimise dropout in our *a priori* regions of interest, we acquired functional data using a specialised OFC-AMY 2D Echo Planar Image (EPI) sequence (Weiskopf et al., 2006) with a TR of 3.36s and a TE of 30ms. We acquired a 48 slice whole brain coverage with 3mm voxels isotropic, using a slice tilt of -30^o. To bias correct our EPI data, we acquired individual subject whole-brain field maps, 3mm isotropic, with TEs of 10ms and 12.46ms. We further acquired individual subject T1-weighted structural scans, using a 176 slice acquisition, 1mm isotropic, with a TR of 7.92ms and a TE of 2.48ms (MPRAGE). We applied a standard pre-processing pipeline, described in more detail in the Supplementary Information.

In order to investigate the neural activity in response to sensory unpredictability, a general linear model (GLM) analyses were performed in SPM12. Stick regressors included choice onset for all four conditions (entered separately), the decision made (gamble vs decline), stimulus onset which was the appearance of each wheel (entered separately) and the possible outcomes (the accumulation of half a shock or safety). The GLM also included 18 regressors for cardiac and respiratory phases to correct for physiological noise and 6 motion parameters regressors to correct for motion-induced noise. In order to investigate whether individual differences in neural activity related to sensory unpredictability, we ran further GLMs with intolerance of uncertainty, state and trait anxiety as between-subjects factors.

In order to investigate the role of expectations and outcome surprise, we ran another GLM with model-derived quantities. Parametric regressors included the expected value (EV) of each condition, time locked to choice onset and the prediction error (PE), timed locked to feedback. EV and PE were computed by applying the learning model to the participant's sequence of stimuli, choices and outcomes, using the median values of the group model parameters. The PE measure used was a classical information gain signal, the Kullback Leibler (KL) Divergence (Kullback, 1997) (more information in the Supplementary Information). In order to investigate whether individual differences in neural activity related to our model derived quantities, we again ran further GLMs with intolerance of uncertainty, state and trait anxiety as between-subjects factors. Alongside whole-brain analyses, we ran a series of ROI analyses using structural masks for the bilateral insula (from the SPM anatomy toolbox) and amygdala from Michely et al. (2020). Following recommendations outlined in **Chapter 2** regarding subregion specificity of amygdala computations, we further obtained subregion

masks from Michely et al. (2020) and used these to examine subregion CMA/BLA effects.

3.5 Results

3.5.1 Choice behaviour

To assess individual differences in intolerance of uncertainty and trait anxiety, we performed a median split on IUS and STAI-trait scores and included these 'low' (n = 20) and 'high' (n = 19) groups as between subjects factors (separately) in the subsequent analyses. We first performed descriptive analysis to ascertain that participant behaviour was modulated by the probability of shock, and test whether it was also affected by the sensory uncertainty ('wide') vs. certain ('narrow') conditions. A repeated measures ANOVA on the proportion of gambles participants took for each condition revealed a main effect of probability $F(1,37) = 6.77, p < .05$, with more gambles taken for low probability conditions, as expected. There was no significant main effect of uncertainty $F(1,37) = .13, p = .72$ against our hypothesis, or a significant interaction $F(1,37) = .80, p = .38$. There was a further main effect of IUS, with individuals with greater intolerance of uncertainty gambling less in general $F(1,37) = 4.26, p < .05$.

A repeated measures ANOVA on the number of correct choices revealed no significant effect of sensory uncertainty $F(1,37) = .68, p = .42$, but a significant effect of probability $F(1,37) = 33.56, p < 0.001$, with more correct choices taken for low probability wheels. There was further a significant probability X IUS interaction, with individuals with greater intolerance of uncertainty making more errors for the high threat conditions $F(1,37) = 4.78, p = .04$.

3.5.2 Reported beliefs

A repeated measures ANOVA on the final probability ratings, using IUS as a between-subjects factor, revealed a main effect of probability (low/high), $F(1,37) = 13.71. p < 0.001$, with higher probability wheels being rated as higher than low probability, as expected. There was no main effect of uncertainty (wide/narrow) contrary to expectations $F(1,37) = 0.40. p = 0.53$, or a significant interaction between variance and probability $F(1,37) = 0.23. p = 0.63$. There was no main effect of IUS $F(1,37) = 0.07. p = 0.79$, but there was a significant probability X IUS interaction $F(1,37) = 5.59. p < .05$, with high IUS individuals reporting a smaller low to high threat level difference in threat perception relative to low IUS individuals (Figure 3.5A). There was further a significant uncertainty X IUS interaction $F(1,37) = 6.29. p < .05$, with high IUS individuals reporting the more uncertain (wide) conditions as more threatening. The same analysis performed using trait anxiety as a between-subjects factor only revealed a main effect of probability $F(1,37) = 12.4, p < 0.001$, suggesting these results were specific to intolerance of uncertainty. Figure 3.3 shows the reported belief change trial-by-trial, split by each condition.

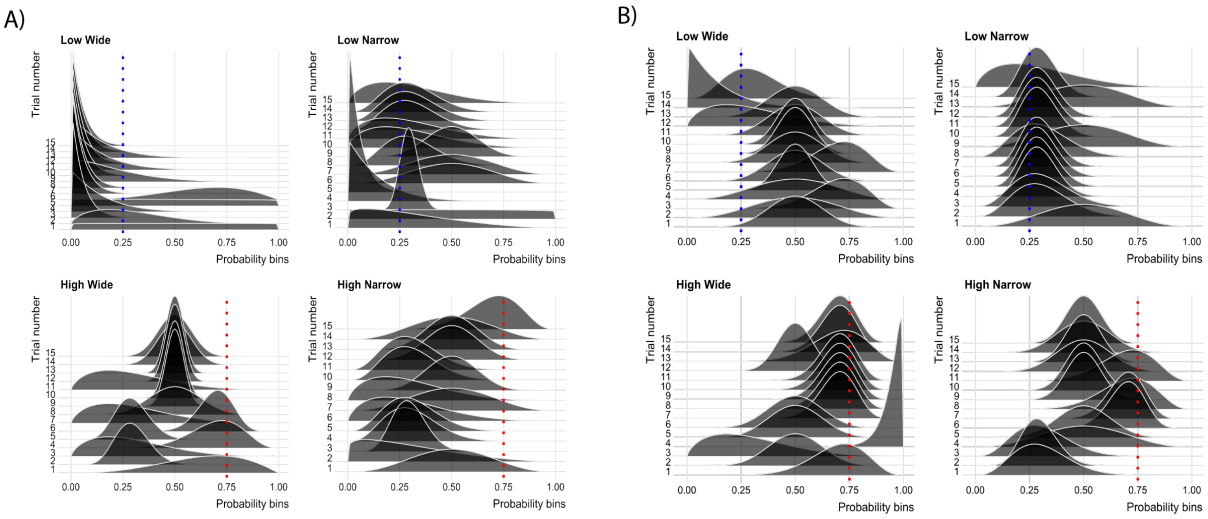

Figure 3.3: Trial-by-trial subjective belief ratings for 15 trials for all conditions A) A low IUS participant (lower 25% of IUS scores) B) A high IUS participant (upper 75% of IUS scores. Dotted lines represent the true probabilities for each condition, with low probability conditions at 0.25 and high probability conditions at 0.75.

3.5.3 Computational modelling of choice behaviour

All models (n = 36) were estimated using Maximum Likelihood Estimation (MLE) and model fits were compared using the Bayesian and Akaike information criteria (Bozdogan, 1987; Schwarz, 1978) at an individual level, then summed over participants. The best-fitting model in absolute terms according to BIC (Table 3.5.3; See also section 3.8.3 and the SI for full model specifications) was a hierarchical Bayesian model that had shared parameters for all conditions and a gambling bias, with 7 parameters in total. The best-fitting associative learning model was a model with 6 parameters in total, containing split learning rates for shock and relief outcomes, an initial bias parameter, a gambling bias, an inverse temperature and a 'certainty boost' parameter, highlighting the role of uncertainty on learning. However, even with fewer parameters, this model performed far worse than the best Bayesian model in terms of BIC.

Model Type	Model Family Name	NP	LL	AIC	BIC	Pseudo-r^2
Associative	Shock/relief LR	3-5	-3230.79	6773.58	7248.96	0.23
Associative	Shock/relief LR + certboost	4-6	-3115.64	6621.29	7215.51	0.26
Associative	Cert/uncert LR	3-5	-3444.91	7123.81	7480.35	0.18
Bayesian	**Shared**	**6-8**	-2717.64	5981.28	**6813.19**	0.35
Bayesian	Split prior mean	8-10	-2679.85	5983.71	6934.47	0.36
Bayesian	Split prior uncertainty	8-10	-2706.95	6037.89	6988.65	0.36
Bayesian	Split prior scale	8-10	-2696.04	6016.07	6966.83	0.36
Bayesian	Split prior location	8-10	-2614.50	5931.00	7000.60	0.38
Bayesian	Full split	10-12	-2595.08	6048.16	7355.46	0.38

Table 3.1: Fit statistics for the best models from each family. Shock/relief LR contained split learning rates for shock/relief outcomes. Cert/uncert LR contained seperate learning rates for the wide vs narrow conditions. The split prior models contained separate prior parameters for the wide vs narrow conditions. The final model selected, the shared prior Bayesian model, is given in bold, selected according to BIC. Note that AIC criteria and pseudo-r^2 suggests the split prior location model, however the final winning model was chosen as it was more parsimonious.

We wanted to further determine which, if any, of our best fitting model parameters were related to individual differences in intolerance of uncertainty. We had no *a priori* hypotheses about which parameters would carry the relationship, therefore we performed Bayesian regression on each of the model parameters, using IUS and trait and state anxiety as predictors. The results of the Bayesian regression models with separate predictors (3.4) indicated that state anxiety was related to a lower ξ_{bsl} (mean) prior parameter (M = -0.33, CI [-0.65 -0.02]). Lower ξ_{bsl} parameter values result in a heightened initial expectation about uncertain conditions having higher probability of shock. We also saw that individuals who are intolerant of uncertainty had a larger ψ_{bsl} parameter (M = 0.32, CI [0 0.63]). The ψ_{bsl} parameter is involved with determining the initial speed of learning, with larger values resulting in faster learning, therefore people with high IUS had faster learning of the shock probabilities.

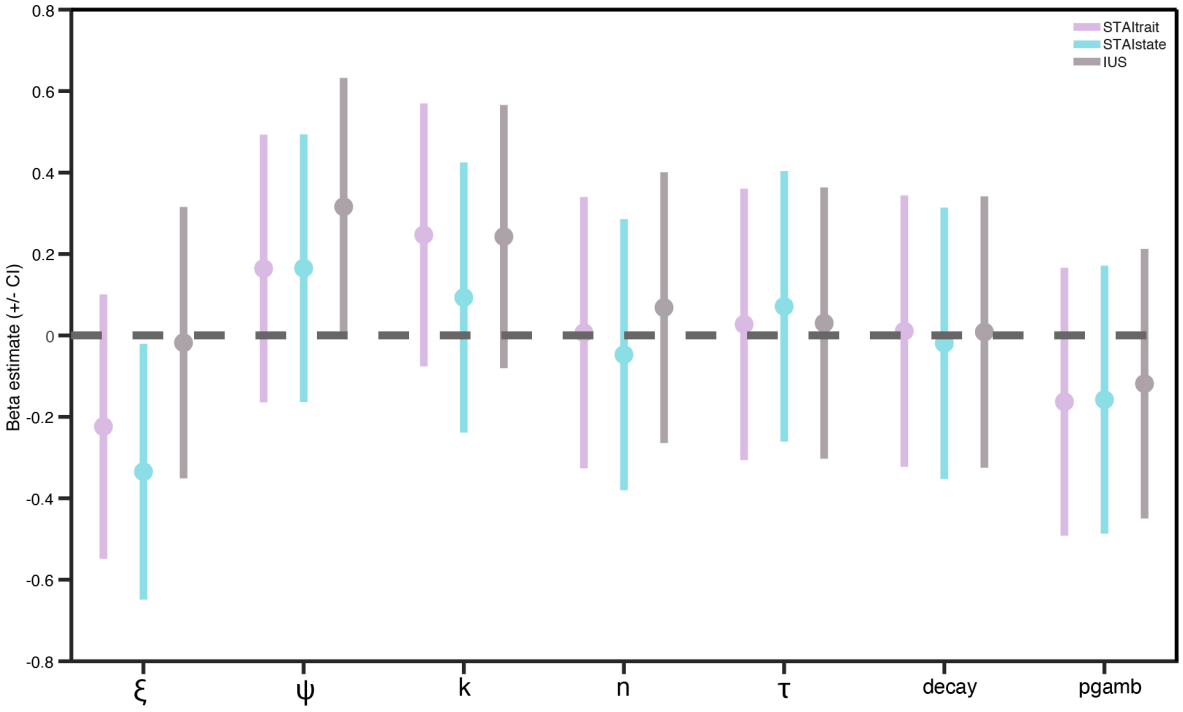

Figure 3.4: Results of Bayesian regression models with best-fitting Bayesian model parameters as outcome variables and trait and state anxiety, and intolerance of uncertainty as predictors entered separately into the model. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

3.5.4 Neural processes of sensory unpredictability

The first analysis focused on the representation of unpredictability as a function of the stimulus variance (wide vs narrow). We examined the response to stimulus onset at the choice stage and feedback using whole-brain regression analysis. To test the hypothesis that sensory uncertainty would impact neural activity, we computed difference contrasts for wide > narrow, and narrow > wide at the stimulus onset and feedback stage of the task. The results of glm analysis without parametric regressors investigating wide > narrow activation at stimulus onset, revealed no significant clusters (at FWE level), nor did the reverse contrast, narrow > wide at the group level. Including IUS as a between-subjects factor (wide > narrow) revealed significant bilateral activation in the insula at cluster level (FWE

corrected) (MNI peak x/y/z coordinates: left side [−42/5/−4], $t(36) = 4.98$, $p = 0.004$; right side [36/17/−1], $t(36) = 4.12$, $p < 0.004$, Figure 3.5). Thus, people who were more intolerant of uncertainty had greater insula activation for the more uncertain stimuli at onset. This result was specific to IUS and not state or trait anxiety. Contrary to our hypothesis, no activation was found in the amygdala or occipital cortex. At feedback stage, there were no significant activations at whole-brain FWE corrected level for any of our contrasts of interest (shock wide > shock narrow, shock narrow > shock wide, relief wide > relief narrow, relief narrow > relief wide). An ROI analysis using our *a priori* regions of interest, the amygdala and insula again saw significant cluster level insula activity in the left, but not right insula (left, -42/5/04, $p = 0.025$, right 36/17/-1, $p = 0.089$, uncorrected $p = 0.013$) in response to stimulus onset, but not feedback. There was no significant activity in the amygdala, either bilaterally or using left/right hemisphere masks. Using subregion CMA/BLA structural masks also yielded no significant activations.

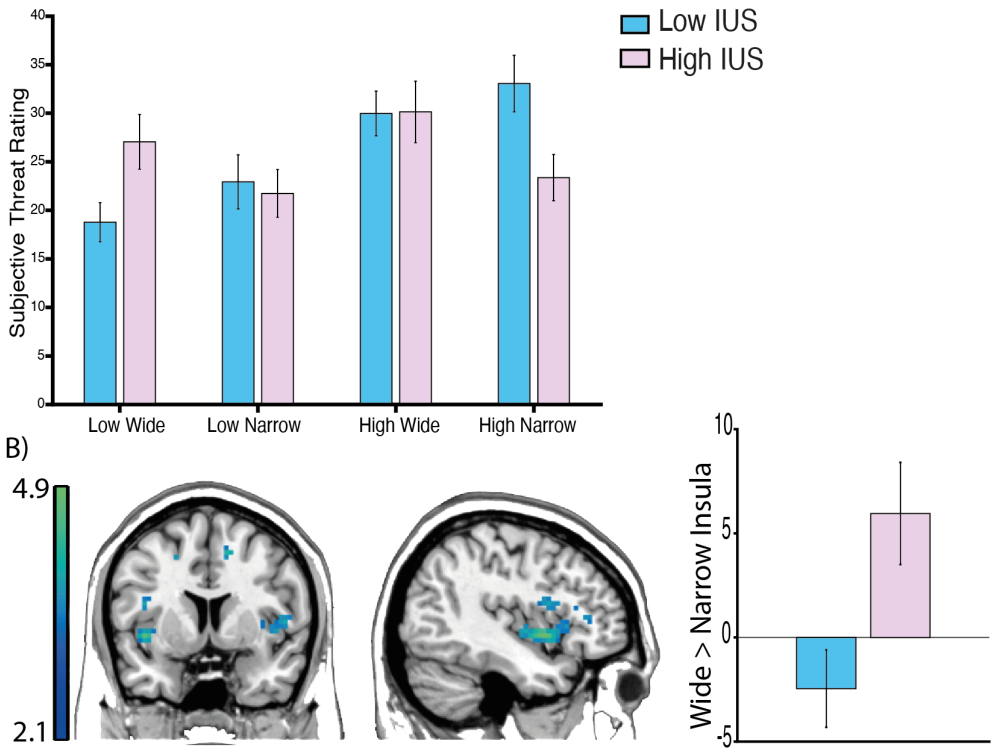

Figure 3.5: A) Subjective threat perception for each condition split by low and high IUS participants to visualise the interaction effect. B) Results of neuroimaging analysis visualised using low and high IUS. Response to onset of uncertain conditions (wide > narrow) correlated with activity in the bilateral insula in individuals high in IUS. Images are thresholded at t > 2.1 for display purposes. Error bars represent +/- SEM.

We next examined the encoding of the expected value (EV) of the stimulus (at stimulus onset) and the KL divergence (our measure of PE, Bayesian surprise) at feedback in a whole-brain regression analysis with trial-by-trial EV (inferred using our computational model and time-locked to stimulus onset) and PEs (inferred using our computational model and time-locked to feedback onset) as parametric modulators. This analysis revealed no significant activity that survived FWE correction. There were no significant activations even at uncorrected threshold level within canonical regions of PE representation, such as the SN/VTA (Nour et al., 2018) and canonical areas representing EV, such as the vmPFC (Bartra et al., 2013). There was no further modulation of any of our model derived parameters and between subjects measures of IUS or anxiety.

3.6 Discussion

Sensory unpredictability is an important element of uncertainty in the environment (Vilares and Kording, 2011; Herry et al., 2007) and may be relevant for threat related processing (Bach and Dolan, 2012). Here, we aimed to determine the relationship between sensory uncertainty and uncertainty about aversive outcome probability and whether this relationship is related to individual differences in anxiety and self-reported intolerance of uncertainty. We show that individuals who are highly intolerant of uncertainty report greater subjective threat estimates for unpredictable, low threat stimuli compared to predictable stimuli, but this was not reflected in their behavioural choices when assessed using traditional statistical measures. Using a novel Bayesian computational model, we were able to reveal differences in initial learning that originate from differences in prior beliefs about uncertainty. Neurally, the insula showed enhanced activity for conditions with greater sensory uncertainty, but only for individuals highly intolerant of uncertainty. These results highlight the connection between sensory uncertainty and individual differences in the subjective perception of uncertainty within threat contexts.

The insula has long been implicated in uncertainty processing (Shankman et al., 2014; Somerville et al., 2013), however we now extend understanding of the role of the insula to include visual sensory uncertainty within threat contexts. These results, however, were limited to individuals who self-reported that they were highly intolerant of uncertainty, providing a link between the subjective perception of uncertainty and differences in how the brain may track uncertainty in these individuals (Simmons et al., 2008). These results were specific to the stimulus presentation stage, before choices are selected, not outcome stage of the task, suggesting the insula is more involved with the anticipation stage, rather than the

response to the aversive outcomes themselves. This has been previously observed in other paradigms, for example using temporally unpredictable negative images (Shankman et al., 2014) and unpredictable shock outcomes (Alvarez et al., 2015). For these paradigms, the aversive outcome was itself unpredictable, whereas here, the unpredictability of the cue was informative for learning about threat, thus our results extend the role of the insula beyond outcome uncertainty to include sensory uncertainty. Our initial hypothesis also focused on the amygdala, as this region is thought to be key in vigilance monitoring and studies have linked heightened amygdala activation to unpredictable events, even those unrelated to task goals (Herry et al., 2007). However, we found no evidence for amygdala activation here, even using strongly aversive stimuli, which have previously been found to involve the amygdala (Michely et al., 2020). Our paradigm differs to other paradigms which have observed amygdala activation, in that shocks were not delivered at the time of outcome, they were accumulated to be delivered at the end of the task. The actual delivery of aversive outcomes may therefore be particularly important for the amygdala (FeldmanHall et al., 2012).

Associative learning models gave a good account of behaviour, but were substantially outperformed by Bayesian models. AL models do not directly quantify sensory uncertainty, and from trial to trial, participants simply keep a cached value of shock outcomes. However, the necessity for a 'certainty boost' parameter clearly showed that participants took account of differing perceptual uncertainty across contexts, and adopted different threat-evasion strategy policies. The 'certainty boost' parameter suggests that computations involving sensory uncertainty are performed differently. Our near-optimal observer (Bayesian) model, on the other hand, allowed us to track trial-by-trial estimates of uncertainty. Our participants had to

perform a number of computations in our task that we were able to track with our model. Their overarching goal, made explicit to them, was to avoid accumulating shocks by making the best decisions during the choice stage of the task. To do this they had to calculate and keep track of the probability of accumulating a shock from each roulette. That is, they were asked to learn and compute the outcome probability of an aversive stimulus and compare it to a stimulus with known outcome probability. The most optimal way to do this is to integrate information about the spatial distribution of the roulette outcomes, thus estimating the mass of the distribution beyond the 'shock threshold'. The superior performance of this model suggests that sensory uncertainty is integrated into learning about threat outcomes and illustrates the relationship between different levels of uncertainty (Bach and Dolan, 2012; Bach et al., 2011). Furthermore, the model revealed differences in prior beliefs about uncertainty for individuals intolerant of uncertainty, which impacted learning, highlighting the link between subjective beliefs and behaviour.

An interesting consideration arising from this task, is the discrepancy between subjective perception about threat level, as reported by participants, and their choice behaviour. High IUS individuals self-reported that they found the uncertain (low probability) condition to have higher threat probability, but this distinction was not observed in their choice behaviour in the form of fewer gambles, although they did gamble less overall. The uncertainty manipulation was only rated as more threatening for the low threat probability condition, suggesting at higher threat levels, sensory uncertainty becomes less informative as a signal for threat. A reduced tendency to gamble is suggestive of increased risk aversion, which has shown to be higher for more anxious individuals, although it has not been directly tied to IUS computationally (Charpentier et al., 2017). Another consideration of the subjective threat

ratings, is that high IUS individuals *underestimated* the threat level of the high threat, low variance condition, suggesting perceptual certainty may convey a safety signal, although again, this was not reflected in choice behaviour. This discrepancy between the neural signals, the subjective threat reporting and behaviour may suggest a potential mismatch between affect, cognition and action. This has been observed in other domains, for example using a metacognition task, Rouault et al. (2018) were able to show that people higher in depression and anxiety levels had similar levels of task performance, but persistently low confidence, highlighting the dissociation between feelings and action.

Surprisingly, our results were specifically related to intolerance of uncertainty, but not measures of anxiety. These measures are highly correlated (Gentes and Ruscio, 2011), but the IUS measure is normally distributed within population (Carleton et al., 2012; Gentes and Ruscio, 2011; McEvoy and Mahoney, 2011), whereas the STAI is usually positively skewed in university samples (Thomas and Cassady, 2021), reflecting a predominance of low anxiety symptoms. Therefore, it is possible that the low variance and small representation of symptomatic individuals using our anxiety measure gave us an inability to detect relationships to neural and computational processes. Using the IUS measure in unselected samples such as this one may therefore be useful for studying processes that are not reflecting mental health problems *per se*, but which reflect individual variation on specific processes thought to be transdiagnostic rick factors for anxiety (Tanovic et al., 2018; Gentes and Ruscio, 2011).

In terms of limitations, we only found weak neural signals in general, possibly highlighting a weakness in the complex design of the task. There were also limitations in terms of the model fitting, which was conducted using maximum likelihood estimation, rather than hierar-

chical model fitting. MLE can result in noisy parameter estimates that may obscure parameter symptom relationships (Moutoussis et al., 2018) and the use of hierarchical methods can improve parameter estimates. Future work should apply hierarchical models to the data to improve parameter estimation. A limitation in our analysis is the use of median split, in which information is lost about individual variance (Iacobucci et al., 2015). Although this was done to simplify analysis, it may have reduced power to detect any effects (**?**). Another limitation of this study is the limited and narrow sample size consisting mostly of university students, which not only limits the generalisability of the findings but limits our ability to assess the test-retest reliability of the findings. As discussed, we only included a sample with subclinical anxiety levels and little variance on these measures. It would be pertinent to study the process in individuals that meet the diagnostic criteria for anxiety disorders, as differences in uncertainty learning is believed to be especially prominent in this group (Bishop and Gagne, 2018).

3.7 Conclusion

We find evidence that people who report being intolerant of uncertainty also hold elevated threat perception of more uncertain, low probability stimuli relative to their non-intolerant counterparts. Computationally, the same individuals also had differences in prior beliefs about uncertainty that led to faster initial learning. Neurally, this was reflected in greater insula activity in response to stimuli with larger variance.

3.8 Supplementary Information

3.8.1 Hierarchical Bayesian Model Descriptions

Basic Hierarchical Model. We modified a well-established hierarchical Bayesian computational model to quantify estimates of sensory uncertainty and, ultimately, perceived threat likelihood. The model frames learning as updating beliefs about the mean and variance of the output of the roulette generating the observations, upon making such observations. The beliefs about mean and variance form the middle layer of a three-level hierarchy: Priors over belief distributions of the mean and variance; beliefs about the mean and variance; and beliefs about the observations. The belief about shock is then the proportion of the lowest-level distribution that exceeds the 'shock threshold'. A link function then specifies the policy (gamble vs decline), by comparing the current beliefs about the probability of receiving accumulating half a shock in the 'gamble' vs. 'decline' lotteries.

The generative distribution of observations is Gaussian with mean and standard deviation unknown to the agent. We represent the agent's beliefs about the probability of outcome S occurring as a Gaussian distribution, parameterized by μ and σ^2. At the top layer, we parametrise the prior beliefs over the two parameters of the generative Gaussian. We parameterise the distributions so they are amenable to conjugate-prior updating, with a Gaussian distribution over the mean, μ, and a Gamma distribution over the precision of the Gaussian, $1/\sigma^2$ (Figure 3.2).

$$\begin{aligned} p(x) &= \frac{1}{s_j}\, t(\frac{x-\mu_j}{s_j};n_j) \text{ with} \\ s_j^2 &= \sigma_j^2 \frac{k_j+1}{k_j n_j} \end{aligned} \tag{9}$$

The trial-by trial belief updating equations were based on the standard conjugate-prior formulation, but were modified to allow for imperfect or 'lossy' updating, expected in real people. For every trial, j, participants update their beliefs upon observation of new evidence. λ is a memory parameter that allows for lossy updating and gradual reversion to baseline beliefs, indicated by θ_{bsl}. Four free baseline parameters, k_{bsl},n_{bsl},ξ_{bsl},ψ_{bsl} were fitted to the data, along with λ.

$$
\begin{aligned}
n_{j+1} &= n_{bsl} + \lambda(n_j - n_{bsl}) + 1 \\
k_{j+1} &= k_{bsl} + \lambda(k_j - k_{bsl}) + 1 \\
\xi_{j+1} &= \xi_{bsl} + \lambda(\xi_j - \xi_{bsl}) + \frac{1}{k_{j+1}}(x - (\xi_{bsl} + \lambda(\xi_j - \xi_{bsl}))) \\
\psi^2_{j+1} &= \psi^2_{bsl} + \lambda(\psi^2_j - \psi^2_{bsl}) + \frac{k_{j+1} - 1}{k_{j+1}}(x - \mu_{j+1})^2
\end{aligned} \tag{10}
$$

Complete description of models can be found in Table S3.1.

3.8.2 Link Functions

The probability of choosing each action, a, (gamble vs decline) as derived from the posterior probability is then entered a standard softmax function, weighed by a 'noise' or temperature parameter $\tau > 0$.

$$
p(a) = \frac{1}{1 + exp - \frac{(Q(gamble) - Q(decline)}{\tau}} \tag{11}
$$

3.8.3 Modified Hierarchical Models

Additional parameters

The initial bias parameter, I, allowed starting values to vary from -1 to 1, reflecting a starting tendency to either gamble or decline in the absence of any prior knowledge. This situation reflects a starting tendency towards either action in conditions of complete uncertainty. Values above 0 indicate a gambling bias, values below 0 indicate a decline bias.

Models had separate initial starting beliefs applied to the first round of the experiment, formalizing the intuition that individuals may have expectations upon entering a new environment which differ from their general beliefs. At trial 1 (t=1) policy was evaluated according to the free parameter, I, which is the same for all states.

$$Q(a,s)_{t=1} = I \tag{12}$$

$$Q(a,s)_{t+1} = p(x) \tag{13}$$

At trial 2, we assumed that individuals converged to policies determined by their belief-distribution, $p(x)$.

The gamble bias parameter, β, added a constant term onto the probability of choosing the gamble action $-1 < \beta < 1$. Values above 0 indicate a gambling bias, values below 0 indicate a decline bias.

$$Q(a,s)_t = p(x) + \beta \tag{14}$$

Table S3.1: Parameter details and fit statistics for all Hierarchical Bayesian Models.

	Model	**NP**	ξ	ψ	k	n	τ	**decay**	**initbias**	**pgamb**	**LL**	**AIC**	**BIC**	**Pseudo-r^2**
1	Shared	6	1	1	1	1	1	1			-2824.92	6117.85	6830.92	0.33
2	Shared	7	1	1	1	1	1	1	1		-2786.66	6119.32	6951.23	0.34
3	**Shared**	**7**	**1**	**1**	**1**	**1**	**1**	**1**		**1**	-2717.64	5981.28	**6813.19**	0.35
4	Shared	8	1	1	1	1	1	1	1	1	-2667.53	5959.06	6909.82	0.37
5	Mean	8	2	1	2	1	1	1			-2679.85	5983.71	6934.47	0.36
6	Mean	9	2	1	2	1	1	1	1		-2628.48	5958.97	7028.57	0.38
7	Mean	9	2	1	2	1	1	1		1	-2614.89	5931.78	7001.38	0.38
8	Mean	10	2	1	2	1	1	1	1	1	-2543.89	5867.79	7056.24	0.40
9	Uncert	8	1	2	1	2	1	1			-2706.95	6037.89	6988.65	0.36
10	Uncert	9	1	2	1	2	1	1	1		-2653.26	6008.52	7078.13	0.37
11	Uncert	9	1	2	1	2	1	1		1	-2615.20	5932.40	7002.01	0.38
12	Uncert	10	1	2	1	2	1	1	1	1	-2565.21	5910.42	7098.87	0.39
13	Scale	8	1	2	2	1	1	1			-2696.04	6016.07	6966.83	0.36
14	Scale	9	1	2	2	1	1	1	1		-2630.50	5963.00	7032.61	0.38
15	Scale	9	1	2	2	1	1	1		1	-2670.16	6042.32	7111.92	0.37
16	Scale	10	1	2	2	1	1	1	1	1	-2567.59	5915.17	7103.62	0.39
17	Location	8	2	1	1	2	1	1			-2716.75	6057.50	7008.26	0.35
18	Location	9	2	1	1	2	1	1	1		-2631.00	5964.01	7033.61	0.38
19	Location	9	2	1	1	2	1	1		1	-2614.50	5931.00	7000.60	0.38
20	Location	10	2	1	1	2	1	1	1	1	-2554.55	5889.10	7077.55	0.39
21	Wide/narrow	10	2	2	2	2	1	1			-3923.92	8627.83	9816.28	0.07
22	Wide/narrow	11	2	2	2	2	1	1	1		-2604.47	6066.93	7374.23	0.38
23	Wide/narrow	11	2	2	2	2	1	1		1	-2595.08	6048.16	7355.46	0.38
24	Wide/narrow	12	2	2	2	2	1	1	1	1	-2541.88	6019.76	7445.90	0.40

Separate parameters for wide/narrow distributions

The four free model parameters that comprise the prior belief distributions were separated according to the wide/narrow roulettes. This was done to formalise the notion that people have different expectations about predictable vs unpredictable stimuli. Different combinations of parameter separations were considered. The wide/narrow mean models were split according to the priors over the mean, μ, giving two ξ and two k parameters. The wide/narrow SD models were split according to the priors over the variance, σ^2, giving two ψ and two n parameters. The wide/narrow scale models were split according to the scale parameters of the prior distributions, giving two ψ and two k parameters. The wide/narrow location models were split according to the location parameters of the prior distributions, giving two ξ and two n parameters. Finally, the wide/narrow all model allowed all 4 prior parameters to vary, resulting in two of each parameter.

Table S3.2: Parameter details and fit statistics for all Associative Learning Models.

	Model	NP	lambda	temp	init	pgamb	certboost	LL	AIC	BIC	Pseudo-r^2	PP
1	Shock/relief	3	2	1				-3381.09	6996.17	7352.71	0.20	0.58
2	Shock/relief	4	2	1	1			-3239.89	6791.78	7267.16	0.23	0.60
3	Shock/relief	5	2	1	1	1		-3157.07	6704.13	7298.36	0.25	0.61
4	Shock/relief	4	2	1		1		-3230.79	6773.58	7248.96	0.23	0.60
5	Shock/relief	4	2	1			1	-3241.78	6795.56	7270.94	0.23	0.59
6	Shock/relief	5	2	1	1		1	-3125.51	6641.02	7235.24	0.26	0.61
7	**Shock/relief**	**5**	**2**	**1**		**1**	**1**	**-3115.64**	**6621.29**	**7215.51**	**0.26**	**0.61**
8	Shock/relief	6	2	1	1	1	1	-3048.77	6565.53	7278.60	0.28	0.62
9	Cert/uncert	3	2	1				-3444.91	7123.81	7480.35	0.18	0.58
10	Cert/uncert	3	2	1				-3415.83	7143.66	7619.04	0.19	0.58
11	Cert/uncert	4	2	1		1		-3379.65	7071.29	7546.67	0.20	0.58
12	Cert/uncert	5	2	1	1	1		-3362.10	7114.20	7708.42	0.20	0.58

The best-fitting associative learning model according to BIC contained 2 learning rates, one for learning from shock, another for learning from relief, an initial bias parameter, a gambling bias parameter and a certainty boost parameter.

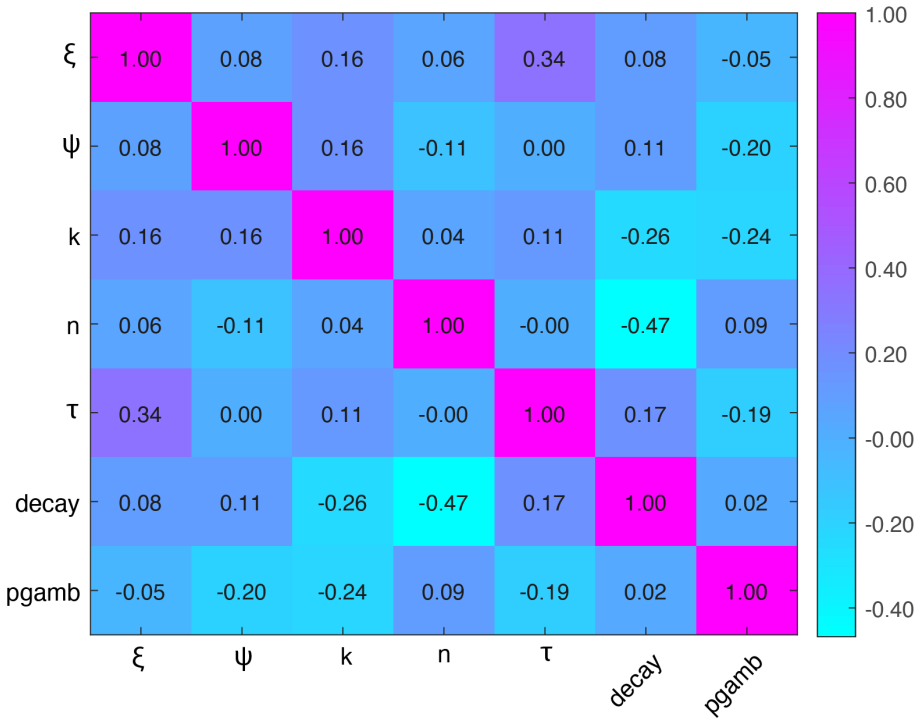

Figure S3.1: Hierarchical Bayesian Model Parameter Correlation Matrix: To investigate whether parameters were trading off against each other, we inspected the parameter correlations. Parameters were largely uncorrelated, with the largest correlation at -0.47.

KL Divergence

The PE measure used was the Kullback Leibler (KL) Divergence (Kullback, 1997) which

quantifies how far the posterior beliefs shift away from the prior beliefs in the light of each outcome observed, and is thus a 'model update' or 'information gain' signal (Nour et al., 2018). The KL divergence was chosen as PE signal because we were interested in the trial-by-trial updating of the distribution representation, rather than a better or worse than expected signal in terms of outcome. To calculate the KL, we took the posterior distribution of the probability of shock before and after outcome:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)}) \tag{15}$$

3.8.4 Pre-processing of fMRI data

In order to prepare the data for analysis, we implemented a standard pre-processing pipeline, using the default settings in SPM 12, unless otherwise specified (Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm). Before the pre-processing pipeline was implemented, we discarded the first six 'dummy' scans for each run and then reoriented all scans to the anterior commissure, to aid in the alignment and normalisation stages. All pre-processed images were individually inspected to check for abnormalities and ghosting (none reported).

Bias correction In the first step, we corrected for bias due to the differences in signal acquired from different brain tissue. For each subject, we applied individually acquired field-maps to our EPI images to correct the field inhomogeneities.

Re-alignment and unwarping During the re-alignment stage, our EPI images were aligned to the first volume acquired for each run. We used a 6 degrees of freedom affine 'rigid body' transformation. We individually inspected the realignment parameters to check for subjects that moved more than 3mm, who were subsequently removed from the analysis for excessive motion. During realignment, unwarping was also performed again applying the field-maps to our EPI data, in order to correct for non-linearities in the motion correction procedure (Andersson et al. (2001)).

Co-registration During the co-registration stage, the subject level EPI images were overlaid onto the structural T1-weighted image in order to anatomically localise the functional data for each subject.

Spatial normalization In the next pre-processing step, we normalised our EPI images to Montreal Neurological Institute (MNI) space. This was done in order to orient images from different individuals to a common space, whilst maintaining the resolution of the images from acquisition (3mm-isotropic functional, 1mm-isotropic structural). Normalisation was implemented by applying the deformation field parameters that were acquired during structural normalisation.

Smoothing In the final step in our pre-processing pipeline, we smoothed our data using a 8mm full-width half maximum (FWHM) kernel. This was done to meet the assumptions of the statistical inference models used for analysis.

4 Chapter 4: Models of self and other beliefs in sub-linical social anxiety

4.1 Acknowledgements

I would like to thank my co-authors for their encouragement towards publishing this work and would also like to say a special thank you to Geert-Jan Will who taught me so much about social learning. The work presented in this chapter has been published in a peer−reviewed journal.

https://www.cpsyjournal.org/articles/10.5334/cpsy.57/

4.2 Abstract

Positive self-beliefs are important for well-being, and are influenced by how others evaluate us during social interactions. Mechanistic accounts of self-beliefs have mostly relied on associative learning models. These account for choice behaviour but not for the explicit beliefs that trouble socially anxious patients. Neither do they speak to self-schemas, which underpin vulnerability according to psychological research. Here, we compared belief-based and associative computational models of social-evaluation, in individuals that varied in fear of negative evaluation (FNE), a core symptom of social anxiety. We used a novel analytic approach, 'clinically informed model-fitting', to determine the influence of FNE symptom scores on model parameters. We found that high-FNE participants learn more easily from negative feedback about themselves, manifesting in greater self-negative learning rates. Crucially, we provide evidence that this bias is underpinned by an overall reduced belief about self-positive attributes. The study population could be characterized equally well by belief-based or associative models, however large individual differences in model likelihood

indicated that some individuals relied more on an associative (model-free), while others more on a belief-guided strategy. Our findings have therapeutic importance, as positive belief activation may be used to specifically modulate learning.

4.3 Introduction

"We don't see things as they are, we see things as we are" - Anaïs Nin

In **Chapter 1**, I examined the relationship between uncertainty and anxiety and introduced the notion that the explicit modelling of uncertainty might be important for understanding how we learn from social information and maintain beliefs about oneself and others. In this chapter, I examine whether models of uncertainty help understanding of social learning by comparing two popular computational frameworks, associative learning and a belief-based framework.

Interpersonal interactions and how others evaluate us are thought to be crucial for shaping one's self-view (Cooley, 1902; Beck, 1971, 2008; Will et al., 2017). The nature of the social information individuals receive, and what they do with that information, is key to understanding how self-beliefs develop and are maintained (Spence and Rapee, 2016). In general, individuals have a positive, optimistic bias, tending to overestimate our competence and likeability (Sharot et al., 2011). This bias appears useful, allowing individuals who hold a positive self-view to benefit from better psychological well-being and mental health (Korn et al., 2014; Conversano et al., 2010; Moore and Fresco, 2012). For people with social anxiety, however this positive bias appears to be reduced.

Cognitive theories of depression and social anxiety hold that repeated exposure to social adversity can teach an individual that the world is an unpredictable and hostile place, where they should expect criticism and poor social outcomes (Beck, 2008; Clark and Wells, 1995). This negative learning forms the schema, a system of beliefs and expectations through which future self-relevant social information is processed (Clark and Wells, 1995; Rapee and Heimberg, 1997). Once activated, the self-schema acts as an information filter, influencing attention, perception, learning and memory, such that the dysfunctional self-views are maintained (Beck, 2008). Schemas are disorder-specific; for social anxiety, their content relates to the core fear of being negatively evaluated by others.

It is important to understand the psychological mechanisms behind inferring evaluation of self and others, and how this integrates into our self-schema. Evidence indicates that the activation of self-beliefs, or self-schema, and the updating of such beliefs in response to social feedback is key (Korn et al., 2014, 2012; Button et al., 2012). However, temperamental preparedness and operant learning routes to anxiety, such as behavioural inhibition and reinforcement via safety-behaviours, are also postulated to be important (Spence and Rapee, 2016).

Understanding self-views is especially important for people who are highly fearful of negative evaluation (FNE), a core symptom of social anxiety (Stopa and Clark, 2001). People with high FNE typically show a negative-bias in learning from social information (Winton et al., 1995). Across a number of studies, utilising a social evaluation probabilistic learning task, it has been demonstrated that high FNE individuals choose positive evaluation words less often when being asked to predict what evaluation a computer persona would give about

them (Button et al., 2015, 2012). This negative bias was not found when they were asked to predict how the persona would evaluate an unknown other person, highlighting the specificity of this bias towards the self. This result is consistent with cognitive models that emphasise the importance of context and object of evaluation for integrating information (Beck, 1971; Cooley, 1902).

Computational cognitive studies have recently addressed self-evaluation (Koban et al., 2017; Will et al., 2017). So far, studies have mostly relied on associative learning models (Rescorla and Wagner, 1972) to capture phenomena such as healthy people giving more weight to positive, rather than negative, information about themselves. Koban et al. (2017) analysed self-evaluation using an associative model, to test whether learning rates – *association values* in learning theory (Hill, 1960) – depended on social anxiety. Social Anxiety Disorder patients were found to have higher learning rates for negative attributes about themselves, compared to healthy controls. Learning-rate based models give a good description of changes in moment-to-moment evaluation of the self, but learning rates are not stable psychological characteristics, depending on a host of factors (Dorfman et al., 2019; Browning et al., 2015; Mathys et al., 2011). Clinically, this maleability is useful, opening up maladaptive learning rates to therapeutic intervention (Kube et al., 2019).

Instead of focusing on behaviour assumed to be gradually reinforced, belief-based frameworks focus how evidence, here provided by social information, updates beliefs. This framework can accommodate the top-down role of self-schema/beliefs, including trait-like views about the self activated given a social context, more naturally than associationist approaches. It also explicitly accounts for the role of uncertainty, which may be especially

important for social learning (Kruschke, 2008).

A Bayesian approach is particularly well suited to modelling the top-down influence of beliefs (Stankevicius et al., 2014), as it has belief update at its core and explicitly represents different strengths of belief. For example, I may believe that I am '80-90%' socially competent but also allow for a socially incompetent 10-20%. Alternative beliefs are then strengthened or weakened as social information accumulates. The certainty of beliefs is informed by learning throughout an individual's history. Certainty then determines how open existing ('prior') beliefs are to change, i.e. determines learning rates. Intuitively, someone with a negative self-view may be more likely to integrate negative evaluations, as they are more in line with their own initial beliefs (see SI for a tutorial demonstration). Similarly biased belief-updating has been demonstrated in non-social reward-based tasks (Stankevicius et al., 2014).

We aimed to clarify the explanatory power of these two psychological frameworks in social-evaluation. We expected associative learning models to capture well the dynamics of learning, while a Bayesian cognitivist framework would provide insight into how beliefs evolve and affect learning. We were interested in mechanisms of biased learning in individuals with high fear of negative evaluation, and its potential basis in biased updating of beliefs about the self.

4.4 Methods and Materials

4.4.1 Measures

Published data was obtained from Button et al. (2015). Data consisted of a Social Evaluation Learning Task (Figure 4.1) completed by 100 participants and a range of questionnaires, of which the primary measure was the Brief Fear of Negative Evaluation (BFNE) scale (Leary, 1983). A higher BFNE score indicates greater fear of negative evaluation. For a full details of the task and sample please see (Button et al., 2015).

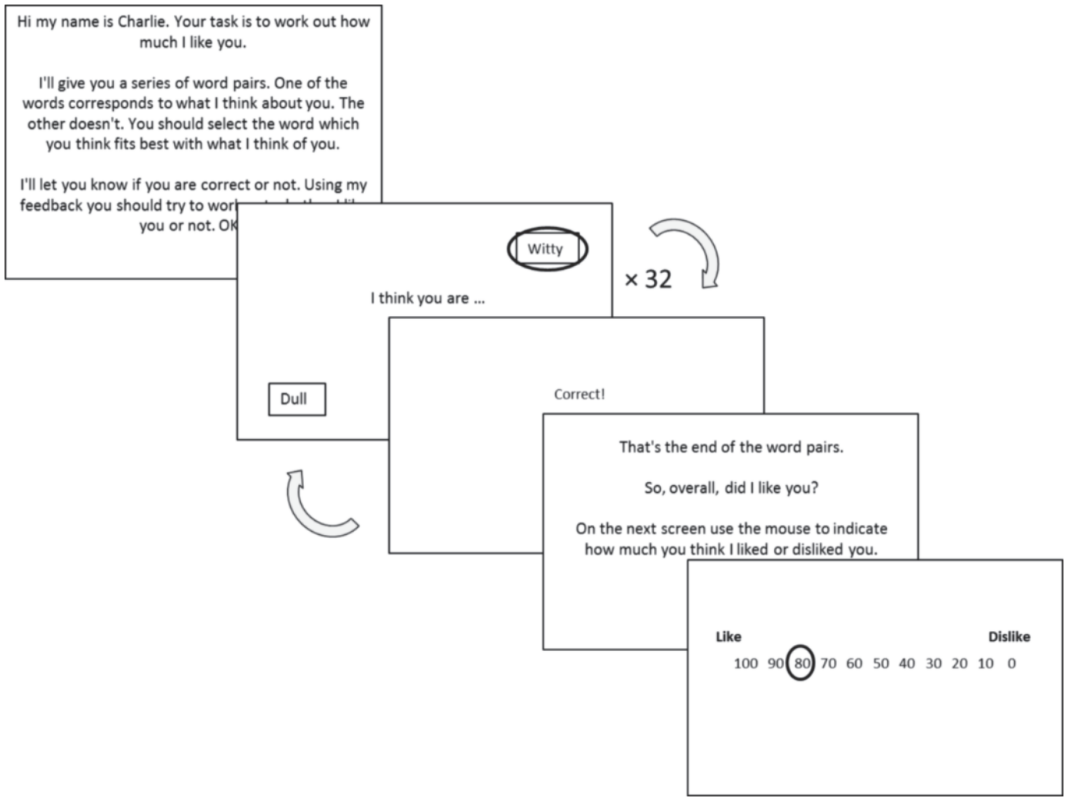

Figure 4.1: Each task block consisted of 32 trials. Participants had to choose between positive and negative words. There were 6 blocks in total, corresponding to 6 evaluative conditions, termed *personas* - Self-like, self-neutral, self-dislike, other-like, other-neutral, other-dislike. Self/other refers to who is being evaluated, like/neutral/dislike refers to the probability of a positive word being correct (0.8, 0.5, 0.2 for the like/neutral/dislike rules respectively)

4.4.2 Sample

In line with a dimensional approach to psychopathology, the original study recruited participants with a range of social anxiety symptoms using an efficient sampling approach to over recruit from the maximally informative extremes (high or low symptoms), ensuring a third of participants had scores in the bottom quartile of BFNE scores, a third from the top quartile and a third from the mid-range using random sampling to exclude one out of two participants with mid-range scores. Participants completed the diagnostic CIS-R (Lewis et al., 1992), which provides diagnoses in line with ICD-10 and DSM-IV. Seven participants met the diagnostic criteria for social phobia and 62 exceeded the cut-off for clinically significant social anxiety on the BFNE.

4.5 Associative and belief-based models

To assess how choices evolved as a function of social feedback, we used computational models. We formalised how social feedback influenced subsequent choices about the self and other using adapted Rescorla-Wagner reinforcement learning models (Rescorla and Wagner, 1972) and novel belief-update models. Here we describe the key features of the models, with technical details to be found in the Supplement.

4.5.1 Associative Learning models

Associative learning models describe learning in terms of value. Here, participants learn the value of the action 'choose the positive attribute' or 'choose the negative attribute', based on feedback. These action-values $Q(action, context)$ are updated after each outcome. A discrepancy between choice and outcome forms a 'prediction error', PE. The PE is then multiplied by a learning rate, λ_c, a parameter weighing the impact of new evidence on

Table 4.1: Model families, grouped according to their defining core parameters.

Model Family Name	NP	Core Parameters	Additional Parameters	
Valence model $-$ 2λ	3-5	$\lambda_{+ve}, \lambda_{-ve}, \tau$	Initial bias	Pos. bias
Self/other asymmetric valence - 3λ	4-6	$\lambda_{self\ pos}, \lambda_{self,-ve}, \lambda_{other}, \tau$	Initial bias	Pos. bias
Self/other valence - 4λ	5-7	$\lambda_{self,+ve}, \lambda_{self,-ve}, \lambda_{other,+ve}, \lambda_{other,-ve}, \tau$	Initial bias	Pos. bias
Belief-update	4	$\alpha, \beta, \eta, \tau$		
Belief-update self/other	7	$\alpha_{self}, \beta_{self}, \alpha_{other}, \beta_{other}, \eta_{self}, \eta_{other}, \tau$		
Belief-update self/other initial bias	9	$\alpha_{self}, \beta_{self}, \alpha_{other}, \beta_{other}, \eta_{self}, \eta_{other}, \tau$	$\alpha_{initial}$	$\beta_{initial}$

existing values, and the result added to update the existing action-value. High learning rates correspond to new evidence having a strong impact, quickly replacing old learning. The *context* s_t simply indexes which state, i.e. computer persona $\times$ (self vs. other), the trial t was about.

$$\begin{aligned} PE_t &= r_t - Q_{t-1}(a_t, s_t) \\ Q_t(a_t, s_t) &= Q_{t-1}(a_t, s_t) + \lambda_c\ PE_t \end{aligned} \tag{16}$$

We focused on learning rates, as these easily characterise which conditions have a major or minor impact on learning. Following Koban et al. (2017), we expected that learning could be valence dependent and therefore allowed separate learning rates for trials with a positive or negative outcome word (irrespective of what choice led to it). So, people might have $\lambda_{+ve\ outcome} > \lambda_{-ve\ outcome}$. Based on the descriptive findings of Button et al. (2015), we were interested in self/other distinction and therefore considered models that had separate learning rates depending on whether the object of learning was self or other, giving $\lambda_{self,+ve}, \lambda_{other,+ve}, \lambda_{self,-ve}$ etc. Models could include an initial value parameter, allowing

starting values $Q(+ve\ word, s_{t=0})$ to reflected an individuals starting tendency towards positivity.

Actions were chosen probabilistically, as a function of a propensity variable for choosing each action. This propensity was the action value $Q(a,s)$ biased by a 'positivity bias' ρ, which quantified biases in favour of choosing positive attributes independent of learning (Eq. 17). $Q(a,s)+\rho$ then entered a standard softmax function, weighed by a 'decision noise' parameter $\tau > 0$:

$$\begin{aligned} P(a = +ve\ word; s) &= z\, exp\frac{Q(a,s)+\rho}{\tau} \\ P(a = -ve\ word; s) &= z\, exp\frac{Q(a,s)}{\tau} \end{aligned} \tag{17}$$

Where z ensured that probabilities add up to 1.

4.5.2 Belief-update models

Belief-update models conceptualised participants as holding beliefs about how approving each computer persona was, from 0 to 1. Such beliefs do not contain just one value ('this persona will give me 80% +ve attributions') but also embody an uncertainty ('but it could be 70 to 90%). They can be formalized by a beta distribution, which conveniently describes beliefs through the amount of positive evidence α and that of negative evidence β held in mind. The mean probability of approval is then the average $p = \alpha/(\alpha + \beta)$.

The belief parameters were updated in every round by augmenting the evidence corresponding to the outcome (say, positive) by 1 piece of evidence. However, we sought to also model views about the self that participants brought to bear independent of learning. Greatly

simplifying clinical theory (Pinto-Gouveia et al., 2006), we represented this as the positive and negative evidence people brought to bear. People thus held two belief components. The first was trait-like, $(\alpha_{trait}, \beta_{trait})$, parameterized individual variability. It was fixed for the duration of the task, and represented the self- or other- view activated given the current context[6]. The second was state-like, $(\alpha_{state}, \beta_{state})$, and it accumulated task information.

$$\begin{aligned} \alpha_t &= \alpha_{trait} + \alpha_{state,t} \\ \beta_t &= \beta_{trait} + \beta_{state,t} \end{aligned} \tag{18}$$

Next, we considered that individuals may not integrate an indefinite amount of evidence, instead gradually discarding older task information. Memory decay parameters $0 < \eta < 1$ thus quantified a participant's effective working memory. Belief-update models could include separate initial values $\alpha_{state,t=0}, \beta_{state,t=0}$. They could also be separated into self/other with respect to $\alpha_{trait,self}, \alpha_{trait,other}$ etc., and with respect to initial values, or indeed the memory decay parameter.

Belief distributions inherently contain uncertainty, which can affect decision variability (Moutoussis et al., 2016). Hence, we considered two classes of probabilistic action choice. In the first, point estimates such as the mean of a belief distribution was used to determine policy. Here, choice variability was independent of belief uncertainty. In the second class, reduced belief uncertainty as a result of evidence accumulation resulted in reduced decision variability. We thus considered several 'link functions' from belief to choice (see Supplement), and

[6]Strictly, the models only contain *notional* or effective evidence, i.e. a numerical representation of the weight of affective memories, images etc. activated in real people

determined the best by model comparison. The winning action-choice function was the one which only depended on the mean of the belief distributions (Eq. 31):

$$P(a = +ve\ word; context = s) = z\ exp \frac{\alpha_s}{(\alpha_s + \beta_s)\tau} \tag{19}$$

A short summary of all models is displayed in Table 4.1. Detailed descriptions are given in the Supplement.

4.5.3 Modelling the relation to Fear of Negative Evaluation

We fitted all models using a hierarchical procedure that optimizes estimation of the relation between model parameters and symptomatic measures, i.e. by *clinically informed model-fitting*. Traditional hierarchical modelling reduces noise in parameter estimates, but we have found that empirical (population) priors which do not take adequately into account the possible correlations with external measures can increase the rates of Type 1 or Type 2 error, in subsequent correlation analyses with unmodelled psychometric measures (Moutoussis et al., 2018). Here, incorporating key psychological hypotheses in the model-fitting can give more accurate estimates of the relationship between model parameters and BFNE scores. As in traditional hierarchical modelling, individual parameters were estimated by taking into account the population distribution they came from, i.e. the 'group prior distribution'. This was in turn estimated from the data, including BFNE scores. We embedded FNE into model-fitting by including slope parameters that estimated a linear contribution of BFNE scores on the mean of the population distribution whence individuals were sampled from, as detailed below.

Let θ be a cognitive parameter that may correlate with BFNE. We modelled this correlation as a linear relationship between BFNE and the mean of θ over people with that value of BFNE:

$$\begin{aligned} \theta \sim N(\quad & \mu_\theta(FNE) \quad , \sigma) \\ & \mu_\theta(FNE) \quad = w\, BFNE + \theta_0 \end{aligned} \tag{20}$$

Where θ_0 is an intercept and in the first instance σ is taken to be independent of FNE. As a cognitive model is fitted using Eq. 20, the posterior distribution over the slope parameter w can be estimated, providing the credible interval over the dependence of of θ on FNE. The w parameter was related to all parameters within each model separately, for example, to test the hypothesis that α_{self} was related to FNE, w was applied to this parameter in one model and then to test an alternative hypothesis that β_{self} was related to FNE, w was applied to this parameter in another model.

We fitted the learning models under consideration (Table 4.1) using RStan (Carpenter et al., 2017). Following RStan convention, means over population-level parameters were scaled so as to be sampled from a standard normal distributions. The respective standard deviations were sampled from half-Cauchy distributions. The individual-level parameters were appropriately constrained in their native space (e.g. 0 - 1 for learning rates), then transformed so as to be subject to the Gaussian distributions informed by the relevant group priors. We initialised Markov-Chain Monte Carlo chains with random starting values. Posterior distributions were formed after 1000 burn-in samples from 4 chains, resulting in a total sample size of approximately 8,000. Convergence was determined by visual inspection

of the trace plots and monitoring the Gelman-Rubin statistic for each parameter (Gelman and Rubin, 1992), with values close to 1.00 implying convergence.

We compared the goodness of fit of different models via approximate leave-one-out cross-validation (LOO). This provides a measure of the likelihood of left-out data, suitable for estimating model-fit in hierarchical models (Carpenter et al., 2017). We then examine the credible intervals of correlation parameters (*w* above) between BFNE and specifically hypothesized parameters (learning rates, beliefs about the self and others) separately in the winning associative and belief-based models. A hypothesis that a parameter correlated with BFNE was tested by determining whether the credible interval of *w* included zero.

4.6 Results

4.6.1 Model fitting and model comparison

Model comparisons using left-out likelihood (LOO) (Vehtari et al., 2017) showed that associative learning models that included separate learning rates for self outperformed ones that did not distinguish between agents. There were also big improvements in model fit upon including an initial bias parameter that allowed individuals to vary in an initial propensity to choose a positive word, and upon including a constant 'positivity bias' boosting the action-value of positive information. Although the best-fitting associative learning model in absolute terms was the self/other valence model, LOO model comparison indicated weak evidence for this model over the next best-fitting model with fewer parameters. We thus also took account parameter recoverability, which was enhanced by having fewer parameters. We thus selected for further work the 'self/other asymmetric valence model', with 3 learning rates, an initial bias parameter and a positive bias parameter (see Supplementary Information for details of the full self/other valence model).

Table 4.2: The best models from each family according to approximate leave-one-out cross-validation. Final models selected are given in bold.

Model Family Name	**N. param**	**LOO**
Valence	3-5	-10026
Self/other valence	5-7	-9858
Self/other asymmetric valence	**4-6**	**-9862**
General learning rate	3-5	-9966
Belief-update IB	7	-9954
Belief-update self/other IB	**8**	**-9768**
Belief-update self/other full IB	9	-9762

Note: IB refers to models with Initial Bias parameters.

As shown in Table 4.2, the best-fitting model overall was a belief-update model with separate self/other alpha, beta and memory parameters and also had free initial bias parameters which also included starting beliefs to vary between individuals. Again, LOO model comparison indicated weak evidence for this model. Following a similar rationale as for the associative models, we selected for further work a 'separate self/other' model with a shared memory parameter. The belief-update model without separate initial α and β parameters also performed almost as well as the best models in their respective families. However, the parameters involved might relate to our hypotheses regarding self-Other activated schemata, and hence we proceeded simply with the best-LOO models. Belief models with separate 'trait' parameters for self and other performed much better than models without, emphasizing a necessary distinction between self and other in learning. We include more details for all models considered above in the supplement.

Although the belief-based model had better fit statistics overall, we asked whether this was because it fitted most people better than the associative models, or whether those that were better described by associative models were in the minority. To estimate this, we simply examined the distribution of the difference between maximum-likelihood (ML) estimates for the associative vs. belief-based models, shown in Figure 4.2. This indicates that for the

majority of participants there was no clear difference between the models, but for about a fifth there was conventionally strong evidence that one or the other model gave a better account of the data. We did not find a significant correlation between BFNE score and the belief-associative ML difference. Here, we computed the difference in log-likelihoods between the two models, with larger differences indicative of one model describing the data better than the other. There was no significant correlation of log-likelihood with BFNE score when models were analysed separately either.

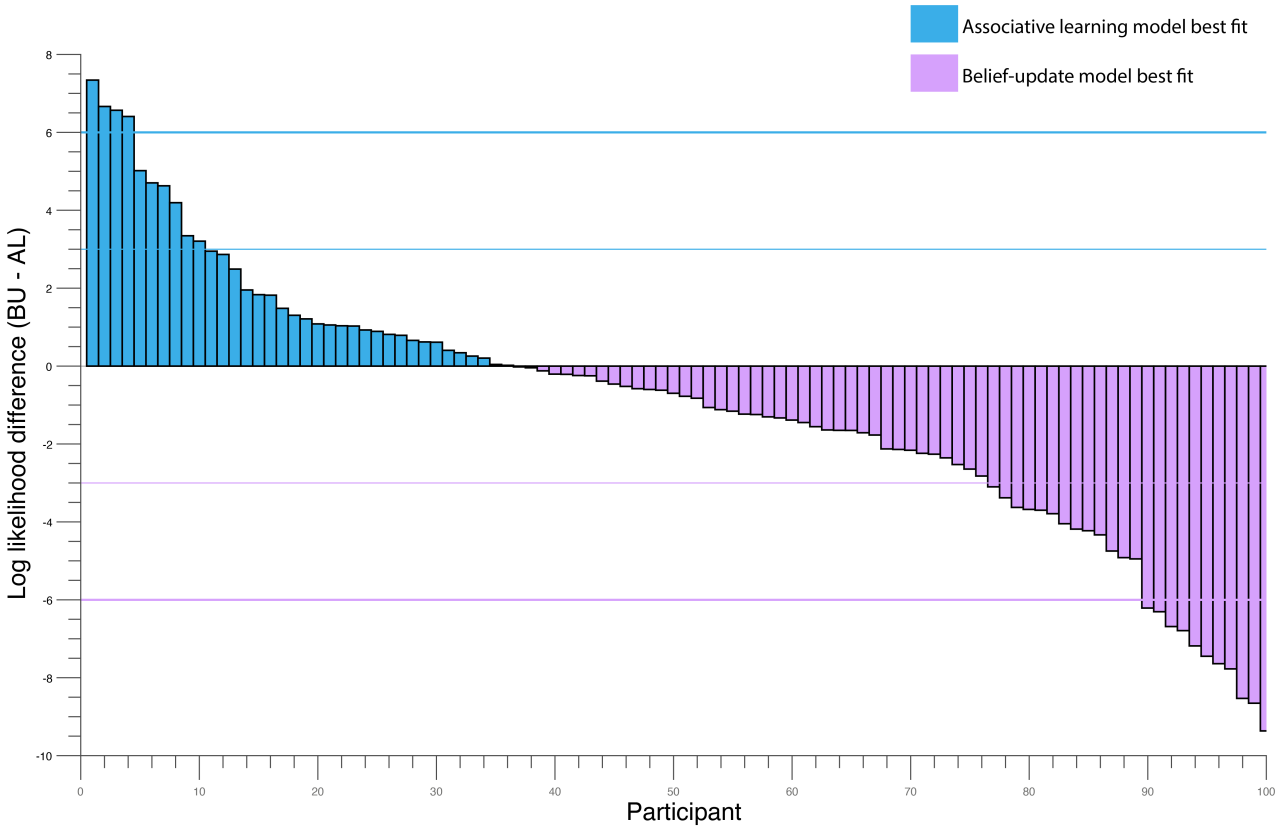

Figure 4.2: Individual log likelihoods for associative learning vs belief-update model. Positive values indicates greater evidence for the associative learning model. The horizontal bars indicate log likelihood differences of +/-3 and +/-6, conventionally mild and strong evidence in favour of one model over the other.

4.6.2 The relationship between BFNE and model parameters

Based on the literature (Carpenter et al., 2017) and the theory of self-schema, we examined the specific hypotheses that BFNE would relate to the trait-evidence in the self schema (α_{self} and/or β_{self}) or the corresponding learning rates $\lambda_{self,+ve}$ and $\lambda_{self,-ve}$ (See supplement for the theoretical derivation of this approximate correspondence). We also examined in an exploratory manner whether the other parameters of the winning models correlated with BFNE scores. We assessed each of the BFNE weight parameters to determine whether their credible interval overlapped 0, which would not support an effect of BFNE on that parameter (Table 4.3).

Table 4.3: Parameter weights on FNE, derived from clinically informed model-fitting.

Associative Learning parameter	Mean w [Lower CI − Upper CI 95%]	Belief-update parameter	Mean w [Lower CI − Upper CI 95%]
$\lambda_{self,+ve}$	0.01 [−0.09 0.09]	α_{self}	**−0.47** [**−0.87 −0.06**]
$\lambda_{self,-ve}$	**0.11** [**0.02 0.20**]	β_{self}	−0.24 [−1.55 1.08]
λ_{other}	−0.05 [−0.19 0.09]	α_{other}	−0.02 [−0.16 0.19]
τ	−0.07 [−0.01 0.15]	β_{other}	0.07 [−0.31 0.45]
Initial bias	−0.09 [−0.19 0.01]	η	−0.22 [−0.56 0.13]
Pos. bias	−0.09 [−0.19 0.01]	τ	−0.09 [−0.25 0.06]
		$\alpha_{initial}$	−0.39 [−0.99 0.22]
		$\beta_{initial}$	−0.97 [−5.07 3.13]

The only associative weight parameter that did not have credible intervals including zero was for the self-negative learning rate (see Table 4.3). This weight parameter was positive, indicating the higher the individual is in FNE, the larger the self-negative learning rate will be. Therefore, it appears that in an associative learning framework, fear of negative evaluation is specifically related to over weighting of negative information, while positive information processing appears intact.

The only belief-update weight parameter that did not have credible intervals including zero was was between BFNE score and the $\alpha_{self,+ve}$ parameter (see 4.3). This weight

parameter was negative, indicating the higher the individual is in FNE, the lower the amount of positive evidence in the self-schema, $\alpha_{trait,self}$, will be. The more negative balance of the self-schema then decreases the mean belief in approval in individuals with higher FNE.

We then explored whether the best fitted parameter values provided evidence for the theoretical correspondence between the two models. From the MLE fit parameters, indeed, $\alpha_{trait,self}$ was strongly anticorrelated with the $\lambda_{-ve,self}$, Spearman $r = -0.49$, raw $p = 3.006e-07$ and $\beta_{trait,self}$, Spearman $r = -0.3$, raw $p < .01$ (Spearman's rho was used due to non-normality). $\lambda_{-ve,self}$ was also correlated with $\beta_{trait,other}$, Spearman $r = -0.21$, $p = 0.04$, but none of the other parameters of the belief-model. Finally, $\lambda_{-ve,self}$ was also strongly anticorrelated to the *proportion* of activated positive self-beliefs, represented by the mean of the beta distribution (Spearman $r = -0.27, p < 0.01$), although this is of not, of course, an independent relationship. The best fitted parameter values from the MCMC fits indicated an even stronger relationship, with the key parameters $\alpha_{trait,self}$ being strongly anticorrelated with the $\lambda_{-ve,self}$, Spearman $r = -0.85$, raw $p = 1.5349e-29$, giving evidence that people with larger learning rates for self-negative information also have lower positive self-belief. Again, there was a strong relationship between the $\lambda_{-ve,self}$ parameter and the proportion of activated positive self-beliefs derived from the mean of the self beta distribution, Spearman $r = -0.78$, raw $p = 4.4583e-22$. There was also a positive correlation between the initial bias and $\alpha_{trait,self}$ parameter, suggesting they represent similar concepts (Spearman $r = 0.50$, $p < .001$) and suggesting people with lower positive self-belief have a prepotent starting tendency towards more negative responses. None of the other parameters indicated correlations.

4.7 Generative Performance

Crucially, good models not only statistically fit the data overall, but are also able to capture specific data features of interest that have not been privileged during modelling (Palminteri et al., 2017). We therefore tested this using our best-fit models. The best associative learning model and belief-update models were used to generate pseudo-data from 100 sample datasets consisting of 1000 participants each, simulating 'ideal experiment' conditions, here with more subjects than resource constraints allow. We checked whether these synthetic experiments reproduced the published findings from real people Button et al. (2015), ran the same formal statistical tests, and examined the credible intervals of each result over simulated samples. We computed the percentage positive response for each persona from the generated data as the number of positive word choices made/32 (number of trials). We ran linear mixed effects (LME) analyses including BFNE scores, persona (like/neutral/dislike) and referential condition (self/other) as predictor variables and percent positive response as outcome variable.

Table 4.4: Generative performance statistics.

	Associative learning model		Belief-update model	
Contrast	Mean β coefficient	% of sig samples	Mean β coefficient	% of sig samples
Main effect BFNE	-0.74 [-0.75 -0.73]	100	-0.73 [-0.74 -0.72]	100
Main effect self/other	-13.28 [-13.56 -13.00]	100	-13.52 [-13.84 -13.20]	100
Main effect persona: like	21.55 [20.98 22.11]	100	24.20 [23.53 24.88]	100
Main effect persona: neutral	19.36 [18.57 20.16]	100	15.97 [15.22 16.73]	94
BFNE X self/other	0.32 [0.32 0.33]	100	0.28 [0.27 0.29]	100
BFNE X persona: like	0.74 [0.73 0.76]	100	0.70 [0.68 0.71]	100
BFNE X persona: neutral	0.19 [0.17 0.20]	34	0.26 [0.25 0.28]	61
BFNE X self/other X persona	-0.30 [-0.31 -0.29]	100	-0.23 [-0.24 -0.21]	89

[a]Note: [Lower CI Upper CI 95%]

As illustrated in Figures 4.4 and 4.3, the generated data reproduced most key features of the real experiment. Table S4.5 shows that the LME results presented in (Button et al., 2015) were well reproduced. Using generated data from the belief-update model, we replicated almost all of the main and interaction effects in over 95% of the samples. The three-way

interaction, however was slightly underestimated. The associative learning model did better in this regard, not only replicating all of the main and interaction effects, but also providing evidence for the significant three-way BFNE $\times$ persona $\times$ condition interaction in over 95% of the samples. Both models slightly overestimated the BFNE difference for the neutral condition.

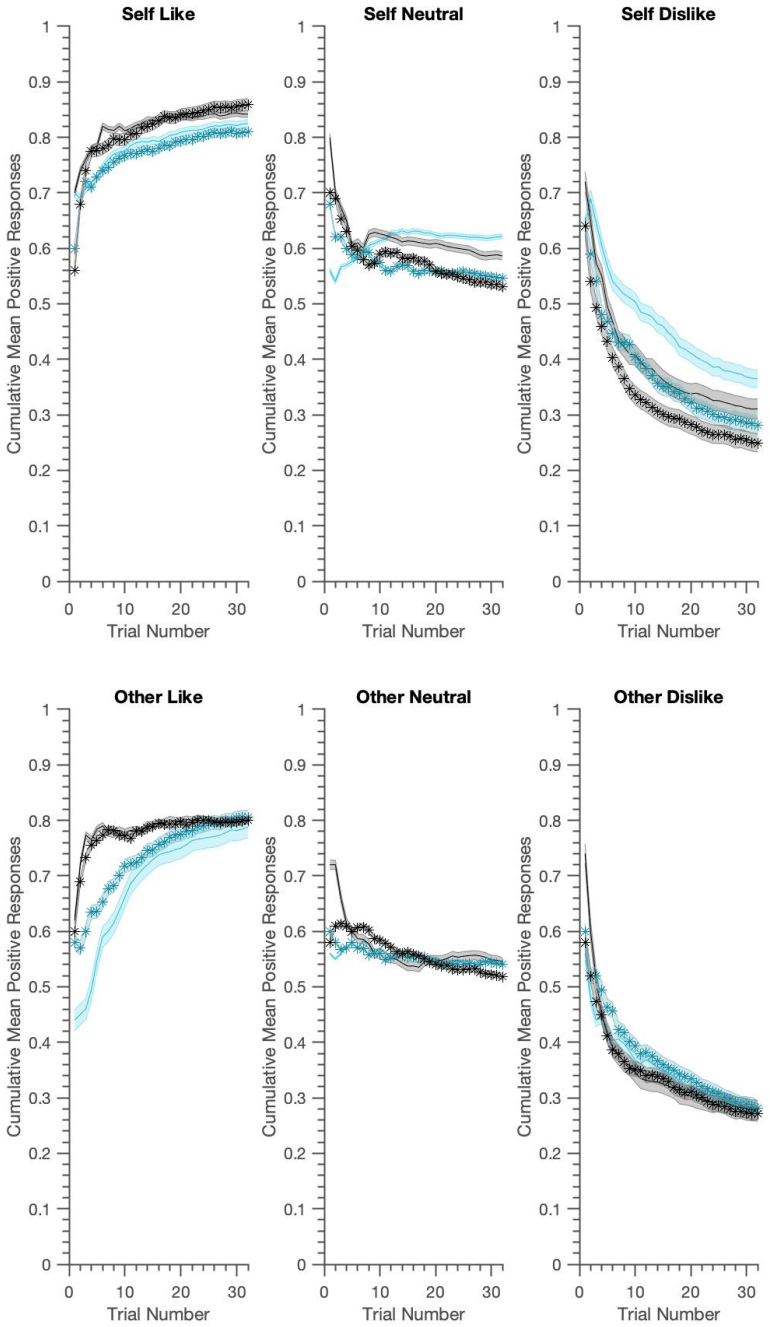

Figure 4.3: Generative performance for the Associative Learning S/O asymmetric model; mean cumulative positive words chosen for actual data (in black) vs. data generated from 'clinically informed fitting' (cyan). Data is visualised using median-split FNE scores (lighter=lower BFNE) and shaded zones represent +/- SEM. The generated data captures the asymmetries in positive vs. negative word selection and the group differences between high and low FNE for the self-referential condition well. There is slower initial learning, especially in the like condition and this model chooses over-optimistically, especially in 'dislike' conditions.

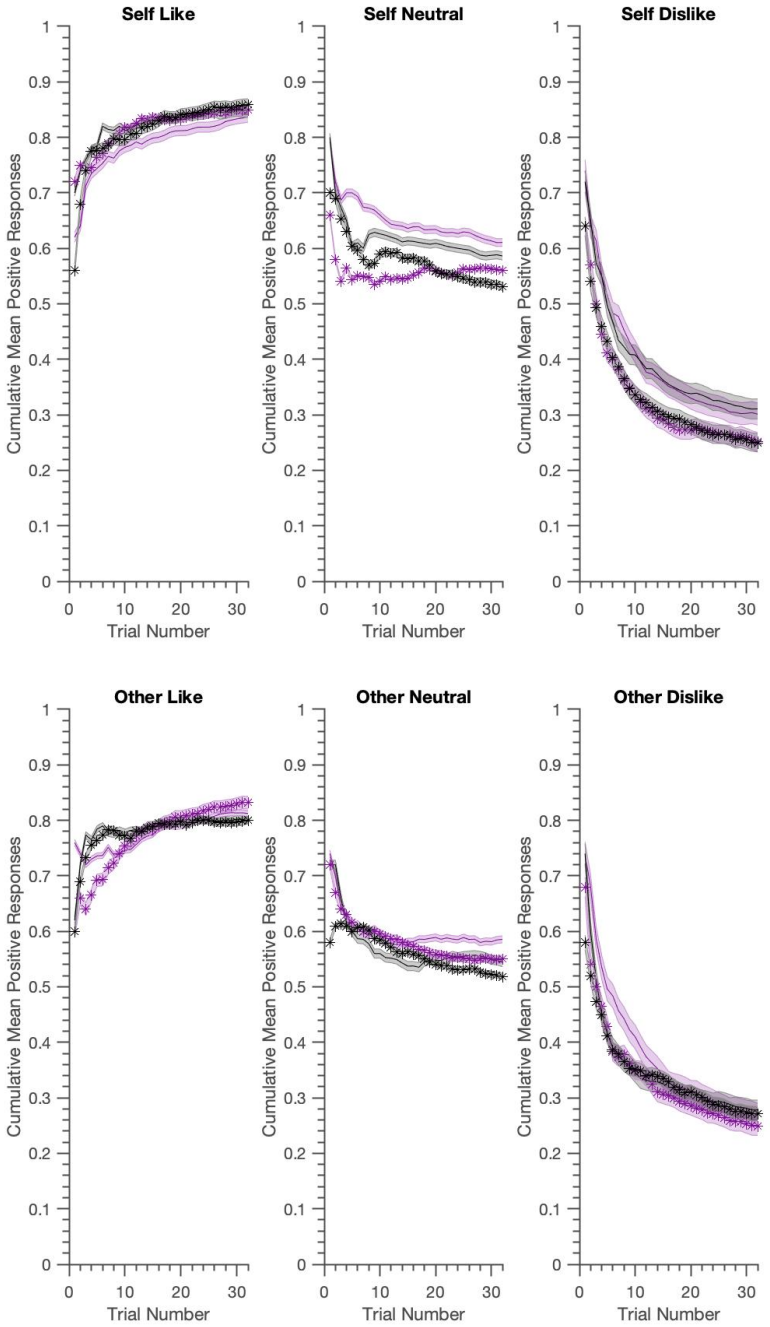

Figure 4.4: Generative performance for the Self/Other Belief-Update model; mean cumulative positive words chosen for actual data (in grey) vs. model (mauve). Again data is visualised using median-split FNE scores, with shaded zones representing +/- SEM for high (darker shade) vs. low (lighter shade) BFNE scores. The generated data captures well the asymmetries in positive vs. negative word selection and the group differences between high and low FNE for the crucial self-referential dislike condition.

4.8 Discussion

We aimed to understand learning about self and others in those fearful of social evaluation, by formalizing and comparing two classic psychological perspectives, associative learning and uncertainty based belief-update models. This is important, as the way in which belief-based accounts used by clinicians should be formalized is unknown, as is how valid they are and whether they are distinct from associationist accounts. Using a well-established Social Evaluation Learning task, we provide evidence that reduced positive content within activated self-schemata underpins increased sensitivity to negative evaluation in socially anxious individuals. Individuals with a less positive self-schema also had a larger self-negative learning rate when investigated using the associative framework. Both associative learning and belief-based models described social learning well, with belief-models especially able to capture the interaction between task context and participant disposition.

We replicated, and also refined, influential findings on associative learning in social anxiety (Koban et al., 2017). Using a task with evidence for reproducibility at the psychological level (Button et al., 2015, 2012), we reproduced the model results reported in (Koban et al., 2017). Namely, socially anxious (high-FNE) individuals had higher learning rates governing the impact of negative information on predictions about the self. We finessed this associative account by including a 'positivity parameter', thus better accounting for participants' optimism bias (Sharot et al., 2011). We also showed that learning rates for positive and negative feedback for the other-referential context were not distinguishable from each other, further pointing at the relevance of self-bias in social anxiety.

Detecting the dependence of task parameters on FNE in this subclinical sample was established through *clinically informed model-fitting*, which makes use of a fundamental property of hierarchical statistical models. These infer the characteristics of each individual not only from the data they provided, but also from the specific population from which they are drawn. Clinically informed model-fitting allowed (yet did not force) empirical priors over cognitive parameters like learning rates to be informed by clinical data, here BFNE scores (Moutoussis et al., 2018). It thus allowed more accurate estimation of the correlation between parameters and FNE. Research is starting to benefit from clinically informed model-fitting (Brown et al., 2020).

To examine whether key features of successful associative learning models were understandable in terms of self-beliefs, which statistically account for improvement during therapy (Gregory and Peters, 2017), we formulated a very simple model of social belief update. We assumed that upon entering a context of evaluation of self or other, individuals activate beliefs about themselves (or others), over and above the evidence gleaned during the task. We focus on the trait-like component of activated schemata, which are constant for the duration of the task but may differ according to the contextual focus of evaluation. This activated self-schema consisted of positive and negative 'notional' evidence that each individual brought to mind. We hypothesised that ordinary beliefs could be modelled as Bayesian beliefs, so that the strength of belief could be quantified much like in CBT ('I believe 70-80% that I will be judged positively'). This meant that belief change not only depended on evidence, but also on the certainty of prior beliefs (Moutoussis et al., 2016). Overall, the success of belief-based models suggested that this was indeed the case, emphasising the utility of explicitly modelling uncertainty. Next, we hypothesized that the amount of evidence that

each individual processed would be variable, in effect a working memory capacity. Again, the evidence supported this hypothesis. Another 'signature' of belief-based cognition might be that more uncertain participants would show increased decision variability. However, model comparison provided evidence against this, suggesting uncertainty was used to form beliefs, but did not directly impact action selection.

Most importantly, FNE was predicted by the amount of positive evidence about the self that was held in mind independent of task feedback. The variation in this positive self-evidence accounted for almost half the variance in self-negative learning rates. This was not, however, the only important model feature, as there was also evidence for reduced negative self-evidence. Combined, these two features may mean that social anxiety is associated with greater uncertainty in one's beliefs about the self. Such increased uncertainty would predict lesser stability of self-evaluation, reminiscent of the changeable self-evaluation found in individuals with low self-esteem (Will et al., 2017). Importantly, the proportion of positive to negative self-evidence was greater in those with lower self-negative learning rates. Thus an activated self-schema including more positive evidence correlated strongly with diminished association value for negative attributes, largely reconciling cognitivist and behaviourist perspectives.

Leave-one-out cross-validation measures suggested that belief-based models may give a better account of behaviour overall, but this finding is likely to hide important individual differences in learning mechanisms. Preliminary analyses indicated that a minority of individuals substantially favoured associative learning, while others belief-updating. Belief-based models are a simple case of model-based cognition, updating the probability of a transition

in the environment (that a persona will judge one positively), while the association models are model-free, incrementally associating values to actions. Thus, some people may be more model-based, whereas others more model-free in the domain of self-evaluation, as people are in impersonal cognition (Daw et al., 2011; Shahar et al., 2019).

This study has the potential to inform treatments for social anxiety. Simple tasks, like the one used here, may assess both the extent of biases and also the patient's predominant cognitive style (belief-based or associationist). Importantly, we describe cognitive mechanisms quantifying and lending support to self-schema theories of social anxiety, reproducing several features of self- and other- evaluation between groups with high and low fear of negative evaluation. Clinically, our results point towards strengthening psycho-education by incorporating rigorous research showing that patients are excessively influenced by negative feedback. In therapy, patients may benefit by learning to activate positive evidence about themselves 'on line', specifically upon exposure to negative feedback, consistent with the work of Kube et al. (2019). Ideally, however, testing such interventions should be guided by a reliable estimate of each individual's cognitive parameters, rather than by features of their condition in general. Here, as is often still the case with computational analyses, further progress is needed (Enkavi et al., 2019). Being able to quantify individuals' self-views may also prove to be useful for assessing the deeper changes that therapy has achieved, rather than just symptomatic change (Taylor and Montgomery, 2007).

There are important limitations to the modelling employed in this study. Our models include a number of hypothesis-driven additional parameters, which aim to capture well-known psychological phenomenon, such as the optimism bias Sharot et al. (2011) or initial

starting propensity towards positive or negative responses (Lockwood et al., 2018). When performing simulations to assess parameter recoverability, some parameters relevant to our hypotheses were difficult to recover. Limited recovery of the 'initial bias parameters' from the belief-update model and 'positivity bias' from the associative learning model (see supplement) suggest that our study may have lacked power to detect differences with respect to FNE with respect to these parameters. Aside from reduced power, the poor recoverability of some parameters renders the model less reliable at the individual level. Nevertheless, fit measures and synthetic data studies indicated that the more complex models, though over-parameterized given our concise data at the individual level, were best in describing the subtle differences in learning associated with FNE in our population. Future studies will need data capable of more fully constraining model parameters, and possibly alternative parameterizations of key models.

Despite the decreased reliability of specific parameters and possibly because of the increased accuracy of complex models, we are able to detect our main effects of interest, and found good recoverability for the positive self-belief and self-negative learning rate. Future studies using clinical populations with larger differences at the behavioural level could observe even greater effect sizes. Thus, our study is well able to detect group level differences in learning between the high vs low FNE groups (the main objective of the study), but poor at capturing individual level differences reliably (Shahar et al., 2019). An important consideration for our more complex models was the ability to reproduce key behavioural statistics of the data, which (Palminteri et al., 2017) recommend as a method of model falsification. Simpler models, despite showing good fit statistics, were unable to capture the key FNE group differences between self and other conditions (see supplement), thus we

preferred models with good fit statistics as well as generative performance. Finally, our modeling of evidence about the self was rudimentary compared to the sophistication of clinical research on self representations(Calvete et al., 2013; Pinto-Gouveia et al., 2006). Future studies modelling self-representations could combine our hierarchical clinically informed model fitting approach with this previous work.

In conclusion, individuals who are high in fear of negative evaluation (yet not care-seeking patients) are more affected by negative social feedback, compared to those unafraid of such feedback. The robustness of typical individuals is consistent with activation of more positive beliefs about themselves independently of feedback, acting as a 'buffer' against developing negative expectations. If replicated, this finding can inform therapeutic interventions aiming at activating positive views of self when people are in the crucible of social judgment.

4.9 Supplementary Information

Associative Learning Model Descriptions

Each associative model encoded:

- the action, the valence of the attribute that the computer persona will choose in a particular trial, $a \in \{+ve\ word, -ve\ word\}$,
- the state or context $s \in \{persona:\ like, neutral, dislike\} \times \{self, other\}$,
- and the reinforcement value of the outcome, $r \in \{+1 = correct, -1 = incorrect\}$ on each trial t.

The Valence model contained separate learning rates λ for positively and negatively valenced outcome words, a.k.a +ve and -ve information, regardless of which action led to them (and so regardless of r_t). The Valence model made no distinction between self and other.

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \lambda_{pos}\ (r_t - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad +ve\ word\ outcome \tag{21}$$

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \lambda_{neg}\ (r_t - Q_{t-1}(a_t, s_t)) \quad \text{for} \quad -ve\ word\ outcome \tag{22}$$

Learning rates λ further varied as follows, and are fully listed in Table S4.1.

Self/other valence model: separate λs for positive and negative information, and for self and other, resulting in 4 learning rates.

Self/other asymmetric valence model: separate λs for positive and negative information for self, but only one for other (i.e. 3 learning rates).

Self-valence model: separate λ for positive information for self only, and a general learning rate for all other information, resulting in 2 learning rates.

Initial values

1. Fixed uncertainty

Models without an initial bias parameter were initialised using two different methods. Fixed uncertainty operationalised the first trial for each persona, $q_{t=0}$ to be zero, expressing equal weight between the positive and negative word and capturing a state of equal uncertainty.

2. Initial bias free parameter

The initial bias parameter, q_0, allowed the starting expectations to vary between -1 and 1. This is applied to both self and other or to the self only and impacts the first trial for every persona.

The positivity bias, ρ, is applied to each round of the expectation update as a constant term which is allowed to vary from -1 and 1. Again, this is applied to both self and other or to the self only.

Model Fit

The fit of competing models for MLE was compared using the Bayesian and Akaike information criteria (Schwarz, 1978; Akaike, 1998) (see eq. 23) at an individual level, which were then summed over participants.

$$
\begin{aligned}
BIC &= -2\ln p(d|\theta_{ML}) + k\ln(n) \\
AIC &= -2\ln p(d|\theta_{ML}) + 2k
\end{aligned} \tag{23}
$$

Where k is the number of free parameters, n the number of trials, d the observed data and $p(d|\theta_{ML})$ the maximum-likelihood of the parameters given the data.

Although AIC and BIC assessments of model fit are the most commonplace reported model fit statistics, these raw values are non-interpretable on their own. Factors such as number of participants and the number of trials will influence the scale of the measures, rendering comparison across studies difficult. The scores can be standardized in order to produce the pseudo-r^2 value, which reflects the variance explained by the model relative to the log likelihood from a model of pure chance (Camerer and Ho, 1999; Daw et al., 2006).

$$
\begin{aligned}
C &= log(0.5)t \\
pseudo - r^2 &= 1 - L/C
\end{aligned} \tag{24}
$$

We can compute the log likelihood of the chance model just by inputting the total number of trials in our experiment, t = 192. We can then compute the pseudo-r^2 value, by taking the ratio of the two values, where L is the log likelihood of the fit model and C is the log likelihood of the chance model for each participant.

As explained in the main text, we fit all models hierarchically and report the appropriate fit statistic for hierarchical models, LOO. Table S4.1 gives all fit statistics for all models. We

Table S4.1: List of associative learning models and their fit indices. Each models also had one decision variability parameter, τ.

#	Model	NP	λ	q_0	ρ	LLs	AIC	BIC	Pseudo r^2	LOO
1	Valence	3	2			-10032.06	20664.13	21641.37	0.25	-10363
2	Valence	4	2	1		-9837.75	20475.51	21778.51	0.26	-10261
3	Valence	4	2	1 (self)		-9714.82	20229.64	21532.64	0.27	-10341
4	Valence	4	2		1	-9718.61	20237.22	21540.21	0.27	-10105
5	Valence	4	2		1 (self)	-9692.06	20184.12	**21487.12**	0.27	-10290
6	Valence	5	2	1	1	-9507.37	20014.74	21643.49	0.29	-10026
7	Self/other valence	5	4			-9568.46	20136.92	21765.67	0.28	-10074.8
8	Self/other valence	6	4	1		-9336.39	19872.78	21827.28	0.30	-9953.8
9	Self/other valence	6	4	1 (self)		-9383.95	19967.91	21922.40	0.29	-9970
10	Self/other valence	6	4		1	-9316.28	19832.56	21787.06	0.30	-9952.7
11	Self/other valence	6	4		1 (self)	-9373.80	19947.61	21902.10	0.30	-9910
12	**Self/other valence**	**7**	**4**	**1**	**1**	**-9215.91**	**19831.82**	22112.06	**0.31**	**-9858.4**
13	Self/other asymmetric	4	3			-9799.56	20399.12	21702.11	0.26	-10032
14	Self/other asymmetric	5	3	1		-9949.23	19898.46	21527.21	0.29	-9951.2
15	Self/other asymmetric	5	3	1 (self)		-9621.75	20243.50	21872.25	0.28	-9987
16	Self/other asymmetric	5	3		1	-9460.40	19920.80	21549.54	0.29	-10017
17	Self/other asymmetric	5	3		1 (self)	-9602.57	20205.14	21833.89	0.28	-10021
18	Self/other asymmetric	6	3	1	1	-9333.66	19867.32	21821.82	0.30	-9861.5
19	Self-valence	3	2			-10058.27	20716.54	21693.79	0.24	-10120.2
20	Self-valence	4	2	1		-9774.68	20349.36	21652.36	0.27	-10098.6
21	Self-valence	4	2	1 (self)		-9904.05	20608.10	21911.10	0.26	-10187
22	Self-valence	4	2		1	-9775.09	20350.18	21653.17	0.27	-10096
23	Self-valence	4	2		1 (self)	-9908.04	20616.08	21919.08	0.26	-10032
24	Self-valence	5	2	1	1	-9637.36	20274.72	21903.47	0.28	-9966

used a model comparison procedure reported in (Vehtari et al., 2017), which compares the expected log pointwise predictive density (ELPD) for each model. We determined the size of importance of ELPD difference by taking models greater than 5 x the SE of the estimate.

Belief-Update Model Descriptions

Belief models contained α and β trait parameters, memory parameters η and one decision variability parameter, τ. A full list of belief-update (BU) models is given in Table S4.2.

Assuming that a participant used an effective number of N items of information, of which at least one was positive and one negative, then asymptotic consistency (if only positive or only negative information were presented for very many trials) meant that the update equations for the state component of beliefs took the form:

$$\alpha_{state,t+1} = (1-\eta)\alpha_{state,t} + \eta + o_t \tag{25}$$

$$\beta_{state,t+1} = (1-\eta)\beta_{state,t} + \eta + (1-o_t) \tag{26}$$

We did not use additional parameters to weigh the feedback information o_t and $1-o_t$ to limit over-parameterising the model, allowing us to test whether preferential learning from positive and negative outcomes could simply be accounted for by activated evidence about the self, α_{trait} and β_{trait} as described in the main text. Belief-update self/other (BU S/O) models thus contained $\alpha_{self}/\beta_{self}$ and $\alpha_{other}/\beta_{other}$ trait parameters. The memory parameter was either the same for self/other, η, or separate (BU S/O full) ($\eta_{self}, \eta_{other}$).

Initial values

1. Trait belief

We can initialise each persona's starting α and β values to be the trait parameter values. This formalizes the intuition that in the absence of evidence, individuals may fall back to their default, stable beliefs. In the case of the separate self/other trait models, this meant that for the self and other personas, people will have different starting points. At trial 1, choice would only depend on the individual's trait values. By trial 2, some evidence accumulated and policy now according to ($\alpha_{state}, \beta_{state}$ + traits). At the start of each new persona, the α_{state} values reset and policy was again evaluated through just the trait values.

2. Initial bias(1)

Models had separate initial starting beliefs applied to the first round of the experiment, formalizing the intuition that individuals may activate modifiable as well as trait-like components to their schemata upon entering a new environment which differ from their general traits. At trial 1 (t=1) policy was evaluated according to free parameters α_{init} and β_{init} and default state values (1).

$$\alpha_{t=1} = \alpha_{init} + \alpha_{state} \tag{27}$$

$$\beta_{t=1} = \beta_{init} + \beta_{state} \tag{28}$$

At trial 2, we assumed that individuals converged to policies determined by (accumulated evidence + traits).

3. Initial bias(2)

Models again had separate initial beliefs, this time decaying over time.

$$\alpha_{t=1} = \alpha_{init} + \alpha_{trait} \tag{29}$$

$$\beta_{t=1} = \beta_{init} + \beta_{trait} \tag{30}$$

4. Fixed uncertainty

Here, we assumed a maximally uncertain starting state, given by beta distribution, $Beta(1,1)$. Again, we applied this only to trial 1.

Belief-Update Link Functions

We explored different ways in which belief uncertainty might impact choice variability.

1. Expectation only policy

Here, we assumed that belief uncertainty has no impact on choice variability. The individual chose solely on the basis of the expectation of each option, $\alpha/(\alpha+\beta),\ \beta/(\alpha+\beta)$ and a fixed decision noise τ, via a standard softmax function:

$$\pi_{+} \quad = \quad 1/(1+\exp\frac{\beta_{eval}-\alpha_{eval}}{\tau(\beta_{eval}+\alpha_{eval})}) \tag{31}$$

2. Strength-of-evidence based policy

Here, choice became more deterministic with the amount of evidence accumulated in favour of one or the other option, i.e. as $|\alpha-\beta|$ increased:

$$\pi_{+} \quad = \quad 1/[1+\exp((\beta_{eval}-\alpha_{eval})/\tau)] \tag{32}$$

3. Belief uncertainty using sampling

Here policy was randomly sampled from a distribution which was that of belief itself, albeit spread out through the decision noise τ:

Table S4.2: List of belief-update models. Each also contained one decision variability parameter, τ. n.f. refers to models not fit hierarchically.

#	Model	NP	α, β	η	α_0, β_0	Link	Initial	LL	AIC	BIC	Pseudo r^2	LOO
1	BU	4	1	1		11	Trait	-9704.08	20208.16	**21511.16**	0.27	-9962
2	BU	4	1	1		11	Uncert	-9747.96	20295.93	21598.93	0.27	-9969.1
3	BU	6	1	1	1	11	IB (1)	-9630.34	20460.68	22415.18	0.28	-9954.1
4	BU	6	1	1	1	11	IB (2)	-9630.47	20460.95	22415.44	0.28	-9958.6
5	BU	4	1	1		12	Trait	-9886.91	20573.83	21876.82	0.26	n.f.
6	BU	4	1	1		12	Uncert	-9915.74	20631.48	21934.48	0.25	n.f.
7	BU	6	1	1		12	IB (1)	-9796.54	20793.08	22747.57	0.26	n.f.
8	BU	4	1	1		13	Trait	-12783.94	26367.88	27670.88	0.04	n.f.
9	BU	4	1	1		13	Uncert	-13854.33	28508.65	29811.65	-0.04	n.f.
10	BU	6	1	1	1	13	IB (1)	-13325.75	27851.50	29806.00	0.00	n.f.
11	BU S/O	6	2	1		11	Trait	-9319.84	**19839.68**	21794.17	0.30	-10486
12	BU S/O	6	2	1		11	Uncert	-9350.03	19900.05	21854.55	0.30	-10044
13	BU S/O	8	2	1	1	11	IB (1)	-9230.29	20060.58	22666.57	**0.31**	-9777.5
14	BU S/O	8	2	1	1	11	IB (2)	-9230.05	20060.11	22666.11	**0.31**	-9768.2
15	BU S/O	6	2	1		12	Trait	-9601.99	20403.98	22358.48	0.28	n.f.
16	BU S/O	6	2	1		12	Uncert	-9620.43	20440.87	22395.37	0.28	n.f.
17	BU S/O	8	2	1	1	12	IB (1)	-9501.19	20602.39	23208.39	0.29	n.f.
18	BU S/O	6	2	1		13	Trait	-13375.24	27950.49	29904.98	-0.01	n.f.
19	BU S/O	6	2	1		13	Uncert	-14540.21	30280.42	32234.92	-0.09	n.f.
20	BU S/O	8	2	1	1	13	IB (1)	-13403.15	28406.29	31012.29	-0.01	n.f.
21	BU S/O full	7	2	2		11	Trait	-9267.28	19934.57	22214.81	0.30	-10305
22	BU S/O full	7	2	2		11	Uncert	-9292.34	19984.67	22264.92	0.30	-10021.3
23	BU S/O full	9	2	2	1	11	IB (1)	-9173.31	20146.62	23078.36	**0.31**	-9776.8
24	**BU S/O full**	**9**	**2**	**2**	**1**	**11**	**IB (2)**	**-9173.01**	20146.03	23077.78	**0.31**	**-9761.7**
25	BU S/O full	7	2	2		12	Trait	-9456.28	20312.57	22592.81	0.29	n.f.
26	BU S/O full	7	2	2		12	Uncert	-9474.01	20348.02	22628.26	0.29	n.f.
27	BU S/O full	9	2	2	1	12	IB (1)	-9357.03	20514.06	23445.80	0.30	n.f.
28	BU S/O full	7	2	2		13	Trait	-12979.24	27358.47	29638.72	0.02	n.f.
29	BU S/O full	7	2	2		13	Uncert	-13897.31	29194.62	31474.86	-0.04	n.f.
30	BU S/O full	9	2	2	1	13	IB (1)	-13472.47	28744.93	31676.68	-0.01	n.f.

$$
\begin{aligned}
\alpha_\pi &= \alpha_{eval}^{1/\tau} \\
\beta_\pi &= \beta_{eval}^{1/\tau} \\
\pi &\sim Beta(\alpha_\pi, \beta_\pi)
\end{aligned} \tag{33}
$$

The Expectation-only policy, which was agnostic to belief uncertainty, fitted the data best (Equation 31, Table S4.2. Belief uncertainty did not further increase decision noise. Given the absence of evidence for an effect at this level, we opted to not run this models hierarchically.

Self-schema: from intuition to key computational hypothesis

Inspired by the work of Koban et al. (2017), but also by the cognitive-behavioural hypothesis that activated, schema-based beliefs about the self appear to 'filter' information and facilitate congruent learning, we performed simulation experiments to establish *prima facie* validity of hypotheses. We set up a model based on the schema-based formulation, and asked whether such a model would behave, from an associative learning point of view, as if it learnt faster about congruent information. Here, we demonstrate the key simulation that led to the hypothesis tested in the main analysis.

In this demonstration, a person accumulates evidence about how two raters like them, one 80% positive and one 80% negative. They simply add 1 to their positive tally, α if they perceive a positive observation from social feedback $o_t = 1$, or 1 to their negative tally β, if they receive disapproving feedback $o_t = 0$.

$$\begin{aligned} \alpha_t &\leftarrow \alpha_{t-1} + o_t \\ \beta_t &\leftarrow \beta_{t-1} + 1 - o_t \end{aligned} \tag{34}$$

However our person has an amount of 'active background evidence' about themselves. For example, they already have positive information weighing as much as 10 items of positive observations in mind, and 2 items of negative information active about themselves. This is our proxy for the positive self bias in the active schema; however it could be reinterpreted as an assumption about whatever the object of learning is, depending on the experiment. So

their *total* active evidence is:

$$\begin{aligned} \alpha_{tot} &\leftarrow \alpha_0 + \alpha_{t-1} + o_t \\ \beta_{tot} &\leftarrow \beta_0 + \beta_{t-1} + (1 - o_t) \end{aligned} \tag{35}$$

Clinically, people seem to have self-schemas persistent in their influence and hard to overcome by new evidence, maybe because new information is often easily forgotten. We wondered if corresponds to Rescorla-Wagner learning, where the learning rate inherently induces *replacement*, that is, forgetting, of old association values with prediction-error based ones. We translated this clinical and computational intuition to the self-evaluation setting by assuming that new evidence, whether of approval or disapproval, is imperfectly remembered from trial to trial with a memory parameter $0 < m < 1$, giving:

$$\begin{aligned} \alpha_{tot} &\leftarrow \alpha_0 + m\,\alpha_{t-1} + o_t \\ \beta_{tot} &\leftarrow \beta_0 + m\,\beta_{t-1} + (1 - o_t) \end{aligned} \tag{36}$$

For the purposes of demonstration, we write the memory parameter in its simplest form and provide further explanation for the form actually used in the main work in equations 41 - 43 below. Finally, CBT-like belief strength in this simple model is simply the proportion of active evidence about the statement in question:

$$\begin{aligned} \mu_+ &= \alpha_{tot}/(\alpha_{tot} + \beta_{tot}) \\ \mu_- &= \beta_{tot}/(\alpha_{tot} + \beta_{tot}) \end{aligned} \tag{37}$$

Having formulated this simple model of evidence accumulation into beliefs, how can we map it onto a value-based, associative model? We could complete the simulation by linking beliefs to behaviour, and fitting this behaviour with an associative model, but greater conceptual clarity can be achieved as follows. In our task, we take the choice probabilities π to scale according to belief strength in a schema-model, but with $exp(action\ value)$ in a value-based model (consistent with, for example, the standard Gibbs Softmax):

$$\begin{aligned} \mu &\sim \pi \sim e^{Q} \text{ so, for example,} \\ Q_{+} &= \ln(\mu_{+}) = \ln(\alpha_{tot}/(\alpha_{tot}+\beta_{tot})) \end{aligned} \tag{38}$$

(We have omitted proportionality constants which do not materially affect this demonstration). Next, we can solve the Rescorla-Wagner learning rule as if the learning rate were the unknown, and obtain for each belief update in the cognitive model the equivalent associative learning rate:

$$\begin{aligned} Q_{t+1} &= Q_t + \lambda(r_t - Q_t) \Rightarrow \\ \lambda &= \frac{Q_{t+1} - Q_t}{r_t - Q_t} \\ &= \frac{\ln \mu_{t+1} - \ln \mu_t}{r_t - \ln \mu_t} \end{aligned} \tag{39}$$

These effective learning rates can be averaged to demonstrate if indeed they are higher for in the presence of $\alpha_0 > \beta_0$. This lead to the hypothesis that in the real data higher learning rates would indeed correspond to congruent activated schemas, with more positive or less negative information activated about the self and either of the two being relatively blunted in

those with high FNE.

Note, however, that the in this demonstration only self-schema parameters are considered, whereas in principle models fitted to the real data contained several more parameters that might account for descriptive differences in behaviour - or even fail to capture these altogether. Note also that the 'equivalent associative learning rate' above does not map exactly to the λ_{pos} and λ_{neg} in our associative models above.

Further details on belief models

Memory, or the leaky accumulation of evidence

Participants may gradually forget older evidence, as well as accumulate new evidence, about themselves. We thus took the evidence α_t and β_t to be subject to decay, as well as trial-dependent accumulation.

Furthermore, we wanted α_t and β_t to be no lesser than 1. If this condition is met, the belief distribution has a value of 0 at the extremes $p = 0$ and $p = 1$. This formalizes the commonsense assumption that people do not think that their most likely self-evaluation is perfectly positive, nor perfectly negative. U-shaped beliefs are also possible in principle, but here we adopted a more conservative framework.

We now consider how participants may update information in their working memory by partially forgetting older information. We denote the maximum size of effective memory to be $N = max(\alpha_t + \beta_t)$. We also need the beliefs represented by α_t and β_t to make sense if participants do not - as yet, or through memory decay - have any evidence about the current context. We assume that they then become agnostic about how they will be evaluated by others, which is achieved by $\alpha_t = \beta_t = 1$. If now $\alpha_t, \beta_t \geq 1$. Happily, this also excludes

nonsensical values for α_{total} and β_{total}. If n^+, n^- is the positive and negative evidence retained from the task itself, over and above the minimum of 1,

$$\begin{aligned} \alpha_t &= 1 + n^+ \\ \beta_t &= 1 + n^- \end{aligned} \tag{40}$$

Consistent with traditional views of working memory update, we take old observations to be chosen at random to replace. This is on average equivalent to a forgetting process operating on α_{t-1}. We now show that this forgetting rate is fully determined by the requirement for α to tend to N-1 (so β is still at least 1) if all $o = 1$, and to 1 if all $o = 0$. Now $max(\alpha_t) = N - 1$. Only the n^+ part of α decays before updating, so for example in the absence of positive feedback $n_t^+ = (1-\eta)n_{t-1}^+$. If the participant receives positive evidence all the time,

$$\begin{aligned} max(\alpha_t) &= [max(n_t^+)(1-\eta)+1] + max(o_t) \Rightarrow \\ max(\alpha_t) &= [(max(\alpha_t)-1)(1-\eta)+1] + 1 \Rightarrow \\ N-1 &= [(N-1-1)(1-\eta)+1] + 1 \Rightarrow \\ (1-\eta) &= (N-3)/(N-2) \Rightarrow \eta = 1/(N-2) \end{aligned} \tag{41}$$

Substituting this into the update equation for α_t gives:

$$
\begin{aligned}
\alpha_{t+1} &= [n_t^+(1-\eta)+1]+o_t \Rightarrow \\
\alpha_{t+1} &= (\alpha_t-1)(1-\eta)+1+o_t \Rightarrow \\
\alpha_{t+1} &= \alpha_t(1-\eta)+\eta+o_t \Rightarrow \\
&= \alpha_t\frac{N-3}{N-2}+\frac{1}{N-2}+o_t
\end{aligned}
\tag{42}
$$

When it comes to real data, η, is a free parameter to be fitted. Apart from $0 \leq \eta < 1$ (or equivalently $2 < N < \infty$) the value of η is not constrained (though it can be subject to regularizing or empirical priors). It is directly related to the effective memory. β_{t+1} is estimated in the exact same manner:

$$
\beta_{t+1} = (1-\eta)\beta_t + \eta + (1-o_t) \tag{43}
$$

Finally, we note that in the models fitted here o_t is either 0 or 1. In future work, we agents can be considered that over-count or under-count information.

Parameter recovery

Table S4.3: Parameter recovery correlations taken from a bootstrapped sample of 1000 parameter values. Key parameters from each model shown in bold.

Associative Learning Parameter	Spearman's Rho, p value	Belief-update Parameter	Spearman's Rho, p value
Self-positive learning rate	r = 0.38, p <.001	**Alpha-self**	**r = 0.72, p <.001**
Self-negative learning rate	**r = 0.64, p <.001**	Beta-self	r = 0.29, p = < .001
Other learning rate	r = 0.72, p <.001	Alpha-other	r = 0.38, p < .001
Tau	r = 0.75, p <.001	Beta-other	r = 0.37, p <.001
Initial Bias	r = 0.54, p <.001	Memory	r = 0.74, p <.001
Positivity Bias	r = 0.22, p <.001	Tau	r = 0.78, p <.001
		Initial bias alpha	r = 0.22, p = <.001
		Initial bias beta	r = 0.21, p = <.001

Correlations between fitted parameter values (Table S4.3) for self/other asymmetric model the self/other belief-update model and recovered parameter values. Parameter recovery was performed by taking the best-fitting parameters as derived by the MCMC hierarchical fits with the key parameter FNE correlations embedded and bootstrapping 1000 values, generating synthetic data and then re-fitting it. Initial values for both models were set to the mean parameter values derived from the MCMC fits and informed by the bounds of each parameter.

Simulated key parameter differences

In order to illustrate the effects of the key parameter, $\lambda_{selfneg}$, for the best-fitting associative learning model, we input different values of the parameter and generated synthetic data for 100 participants. The rest of the parameters were set to the best-fitting parameter values as obtained by MCMC. The results are displayed in Figure S4.1. The same procedure was done with the α_{self} parameter from the best-fitting belief-update model, displayed in Figure S4.1.

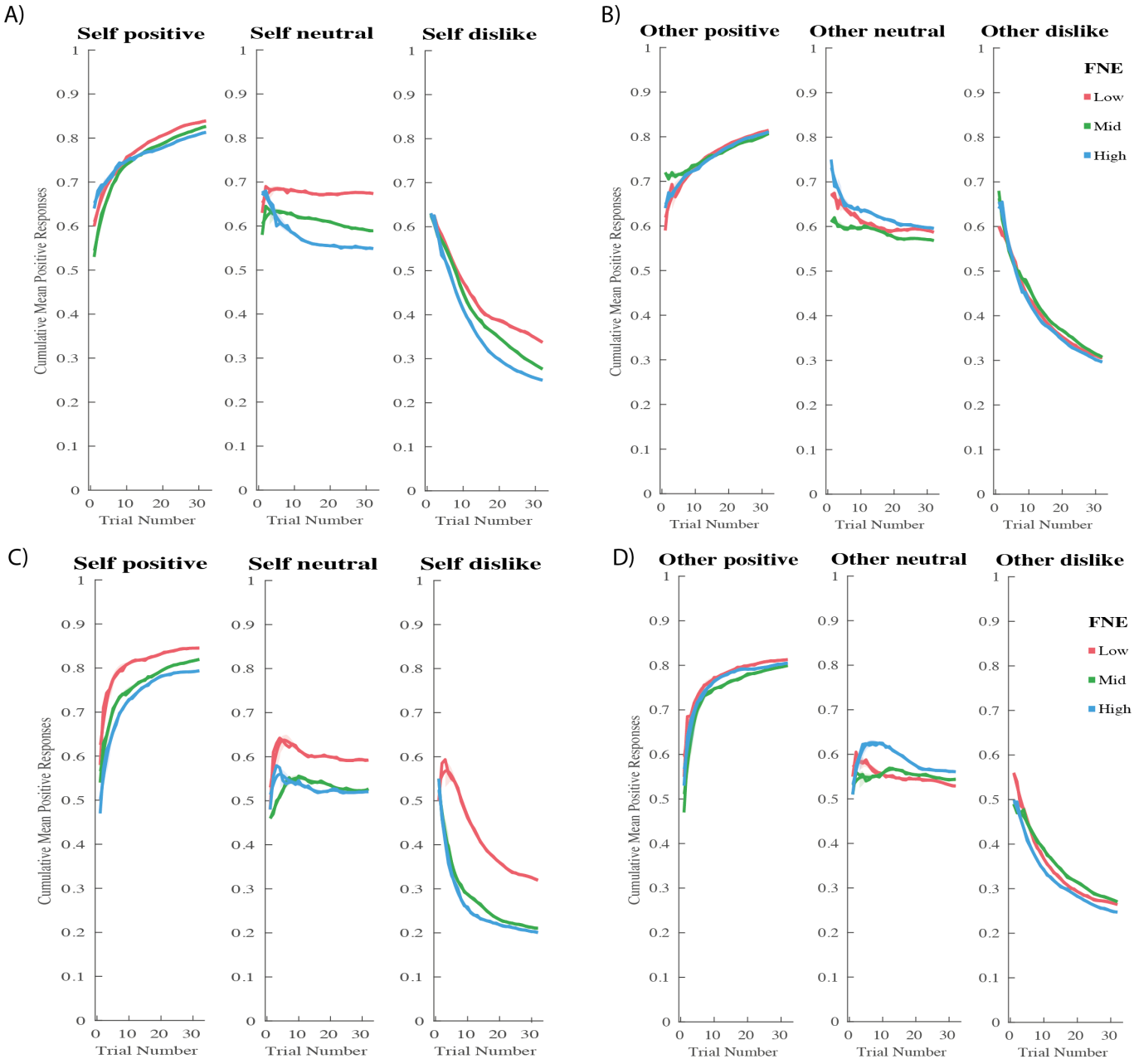

Figure S4.1: Simulated data for different parameter values of the $\lambda_{selfneg}$ parameter from the self/other valence model. Parameter values differentiate behaviour according to FNE in self (A) conditions, but not in other (B) conditions. C) Simulated data for different parameter values of the α_{self} parameter from the self/other belief-update model. Parameter values differentiate behaviour according to FNE in self (C) conditions, but not in other (D) conditions

Correspondence between associative learning and belief-update models

In order to determine the stability of the $\lambda_{selfneg}$, α_{self} correlation across models, we computed the correlation for each nested model variant within the best-fitting model family. Figure S4.2A displays the correlations between the self-negative learning rate parameter from variants of the self/other valence models and the α_{self} parameter from the best-fitting self/other belief-update models. All parameter correlations were significant and have similar values across model variants, with all displaying negative Spearman's correlations ranging from $r = -0.48$ to $r = -0.57$. Figure S4.2B displays the converse, thus shows variants of α_{self} across different self/other belief-update model variants correlated with the best-fitting self/other valence model parameter. Again, all parameter correlations were significant and have similar values across model variants, with all displaying negative Spearman's correlations ranging from $r = -0.49$ to $r = -0.57$.

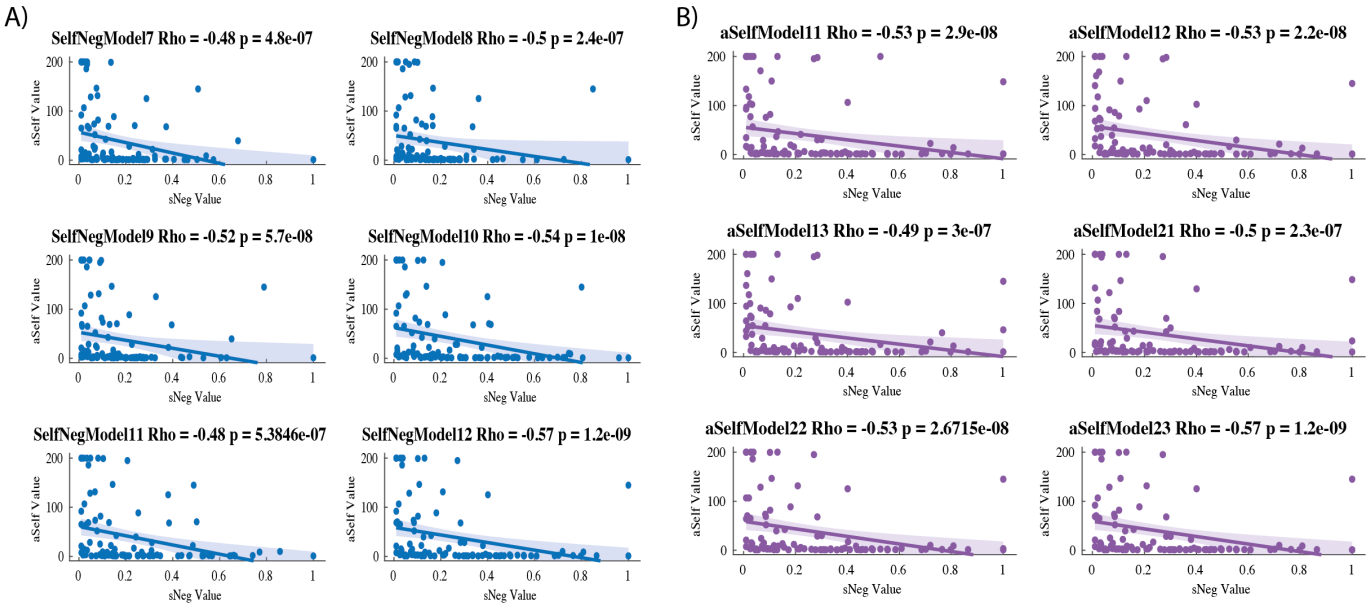

Figure S4.2: A) Correlations between the $\lambda_{selfneg}$ learning rate parameters from variants of the self/other valence models and the α_{self} parameter from the best-fitting belief-update model. B) Correlations between the α_{self} parameters from variants of the self-other belief-update models and the $\lambda_{selfneg}$ learning rate parameter from the best-fitting self/other valence models.

Generative performance of best-fitting model from other model families

Simple models

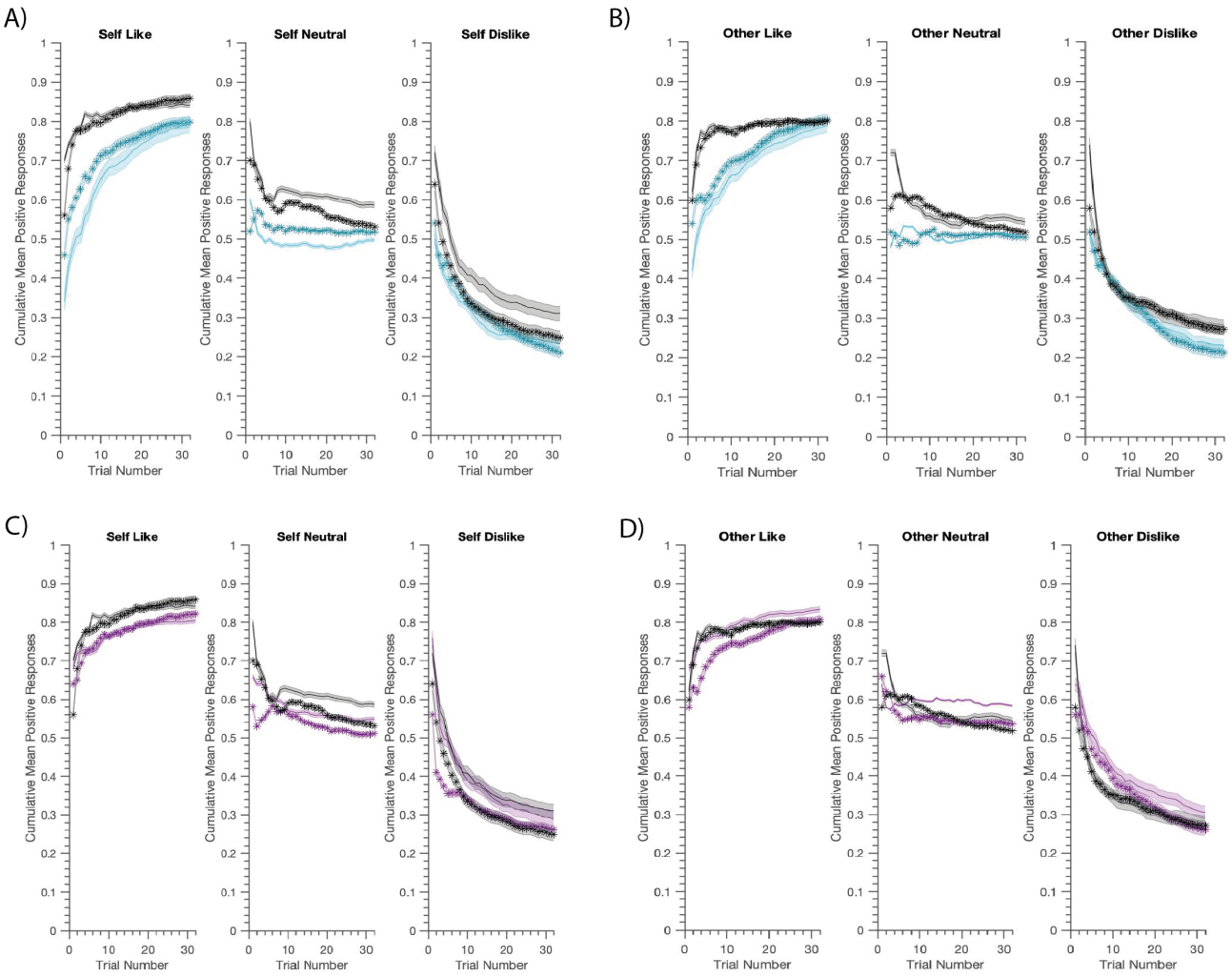

Figure S4.3: A) B) Generative performance for the Associative Learning Valence Model (Model Number 5 Table S4.1); mean cumulative positive words chosen for actual data (in black) vs. data generated from MCMC fits (cyan). The generated data captures the asymmetries in positive vs. negative word selection but not the key high VS low FNE differences. C) D) Generative performance for the Belief-Update Simple Model (Model Number 1 Table S4.2); mean cumulative positive words chosen for actual data (in black) vs. data generated from MCMC fits (cyan). Data is visualised using median-split FNE scores (lighter=lower BFNE) and shaded zones represent $+/-$ SEM. The generated data captures the asymmetries in positive vs. negative word selection but not the key high VS low FNE differences for the self/other distinction.

Table S4.4: Generative statistics for simple models. Both simple models display poor ability to reproduce many of the key statistics, including the 3-way-interaction.

Contrast	AL Mean β [CI]	% of sig samples	BU Mean β [CI]	% of sig samples
Main effect BFNE	-0.08 [-0.07 -0.09]	10	-0.27 [-0.26 -0.28]	98
Main effect self/other	-0.04 [-0.28 0.20]	0	-0.21 [-0.52 0.10]	3
Main effect persona: like	50.67 [50.11 51.23]	100	28.3 [27.55 29.04]	100
Main effect persona: neutral	25.37 [24.55 26.20]	100	15.99 [15.23 16.74]	98
BFNE X self/other	0.01 [-0.82 0.83]	0	0.01 [-0.01 0.01]	2
BFNE X persona: like	0.17 [0.16 0.18]	30	0.31 [0.29 0.32]	87
BFNE X persona: neutral	0.08 [0.06 0.10]	9	0.13 [0.11 0.15]	22
BFNE X self/other X persona	-0.01 [-0.02 0.02]	1	-0.01 [-0.02 0.01]	1

Full models

Table S4.5: Generative performance statistics from full models.

	Associative learning model		Belief-update model	
Contrast	Mean β coefficient	% of sig samples	Mean β coefficient	% of sig samples
Main effect BFNE	-0.66 [-0.67 -0.65]	100	-0.73 [-0.74 -0.72]	100
Main effect self/other	-10.76 [-11.04 -10.49]	100	-13.45 [-13.76 -13.14]	100
Main effect persona: like	21.68 [21.09 22.27]	100	25.61 [25.08 26.13]	100
Main effect persona: neutral	16.86 [16.16 17.56]	98	15.61 [14.80 16.41]	92
BFNE X self/other	0.27 [0.26 0.28]	100	0.27 [0.27 0.28]	100
BFNE X persona: like	0.63 [0.62 0.64]	100	0.72 [0.71 0.73]	100
BFNE X persona: neutral	0.16 [0.14 0.17]	22	0.28 [0.26 0.30]	64
BFNE X self/other X persona	-0.22 [-0.22 -0.23]	98	-0.23 [-0.24 0.22]	94

[a]Note: [Lower CI Upper CI 95%]

Table S4.6: Parameter weights on FNE from the full models, derived from clinically informed model-fitting. The key FNE parameter differences were the same as for the selected models.

Associative Learning parameter	Mean w [Lower CI − Upper CI 95%]	Belief-update parameter	Mean w [Lower CI − Upper CI 95%]
$\lambda_{self,+ve}$	$0.01\ [-0.09\ 0.10]$	α_{self}	$\mathbf{-0.44}\ [\mathbf{-0.80\ -0.04}]$
$\lambda_{self,-ve}$	$\mathbf{0.11}\ [\mathbf{0.01\ 0.22}]$	β_{self}	$-0.04\ [-0.73\ 0.81]$
$\lambda_{other,+ve}$	$-0.06\ [-0.14\ 0.03]$	α_{other}	$-0.05\ [-0.18\ 0.09]$
$\lambda_{other,-ve}$	$-0.01\ [-0.08\ 0.07]$	β_{other}	$0.14\ [-0.24\ 0.52]$
τ	$-0.01\ [-0.08\ 0.07]$	η_{self}	$-0.33\ [-0.79\ 0.13]$
Initial bias	$-0.58\ [-1.22\ 0.07]$	η_{other}	$-0.10\ [-0.47\ 0.26]$
Pos. bias	$-0.04\ [-0.19\ 0.10]$	τ	$-0.09\ [-0.25\ 0.06]$
		$\alpha_{initial}$	$-0.14\ [-0.32\ 0.04]$
		$\beta_{initial}$	$-0.09\ [-0.04\ 0.22]$

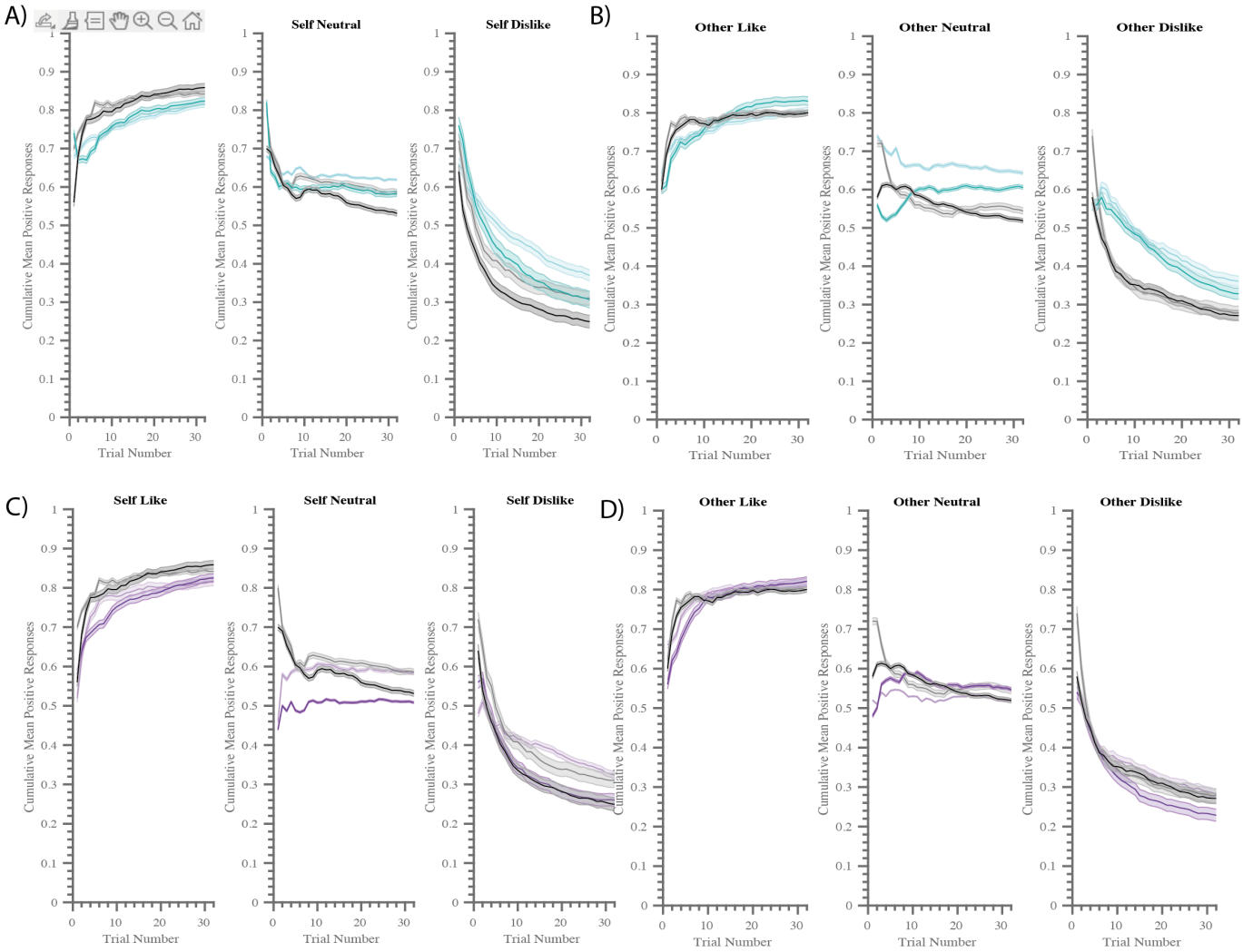

Figure S4.4: A) B) Generative performance for the Associative Learning S/O Valence Model (Model Number 12 Table S4.1); mean cumulative positive words chosen for actual data (in black) vs. data generated from MCMC fits (cyan). The generated data captures the asymmetries in positive vs. negative word selection and the key high VS low FNE differences, similarly to Model 18. Generative performance for the Belief-update self-other full model (Model Number 24 Table S4.2); mean cumulative positive words chosen for actual data (in black) vs. data generated from MCMC fits (mauve). Data is visualised using median-split FNE scores (lighter=lower BFNE) and shaded zones represent $+/-$ SEM. The generated data captures the asymmetries in positive vs. negative word selection and the key high VS low FNE differences, similarly to Model 14

5 Chapter 5: Computational Phenotyping of Anxiety

5.1 Acknowledgements

For providing guidance on the development of this experiment, I would like to give my thanks to Robb Rutledge and Tobias Hauser. I would also like to thank Corey White for kindly providing the exact stimulus set and instructions from his original study so that precise experimental conditions could be better replicated and to Toby Wise for collaborating on this project and providing helpful advice and motivation. This work is currently being written up in preparation for submission as a research article.

The model fitting for the Aversive Learning 'SpaceShip' Task was performed by my collaborator and co-author Dr Toby Wise, the model and method of which was reported in his published paper Wise and Dolan (2020), but is included here briefly. The model parameters and model-derived quantities were provided by him, but subsequent analyses presented in this thesis were performed by myself.

5.2 Abstract

This study aimed to develop a computational phenotype of anxiety, by relating computational parameters that capture anxiety-related mechanisms to anxiety symptomatology. We sought to investigate four anxiety-related concepts: threat bias, negative-self bias, aversive learning and avoidance as well as measures of uncertainty from all tasks. Using parameters derived from computational modelling, we found that trait anxiety was related to our pre-selected threat bias and negative self-bias parameters, but was unrelated to threat learning or avoidance parameters. Cognitive, social and somatic symptom dimensions further revealed reduced threat learning for somatic anxiety and specified the negative self-bias to cognitive and social anxiety, but not somatic. Threat bias parameters were the only parameters consistently found

across these three anxiety subtypes. We further identified a general distress component which comprised of high anxiety, mood and stress symptoms and was related to greater threat bias, greater negative-self bias and reduced threat learning. Contrary to our expectations, none of our measures of uncertainty were related to anxiety. This work forms the first attempt to create a computational phenotype of anxiety.

5.3 Introduction

The introductory chapter of this thesis detailed key computational studies that investigated anxiety related biases. Theoretical work has aimed to understand how this research may fit together (Raymond et al., 2017; Bishop and Gagne, 2018), however to date, there has been no empirical investigation of the relationships between these different computational mechanisms, nor their relationship to different anxiety dimensions. Developing a computational profile, or phenotype (Patzelt et al., 2018) of anxiety is important as it can illuminate mechanisms that cut across different anxiety disorder categories, moving towards a dimensional approach to mental health (Insel et al., 2010; Kotov et al., 2018). This approach aids in understanding the heterogeneity within traditional DSM anxiety disorder diagnostic categories and the shared overlap between them. Moreover, understanding the computational parameter space can help provide target mechanisms for therapeutical interventions that are more tailored to the individual presentation, one of the aims of precision psychiatry (Fernandes et al., 2017). To this aim, we explored the relationship between four cognitive processes theorised to play an important role in different forms of anxiety (Raymond et al., 2017; Bishop and Gagne, 2018), namely: threat bias, negative self-bias in social evaluation, aversive learning and avoidance and we introduce each of these in turn.

A high bias to threat in people with high anxiety has been demonstrated in a number of different ways, most often making use of reaction times to threat as a measure of attentional bias, such as in dot-probe paradigms where participants that respond more quickly to threat related cues are deemed to hold a threat bias (Wieser and Keil, 2020; MacLeod et al., 1986). However tasks that rely solely on RTs as have been criticised as being difficult to interpret and having poor reliability (Rodebaugh et al., 2016). Lexical decision-making tasks, in which participants are required to classify words as either threatening or safe, offer an additional metric on top of reaction times, the threat classification itself, which can aid interpretation of RTs and allow analysis using signal detection theory (SDT) and drift-diffusion modelling (DDM). A number of studies (White et al., 2010a,b, 2016) have demonstrated that greater bias towards threat in people with high anxiety results from a weaker criterion parameter, which signifies a lower threshold for classifying words as threatening as compared to safe. They further showed that highly anxious people have a larger threat drift rate, and a lower neutral drift rate, rendering evidence accumulation for threatening words faster and neutral words slower. We will aim to replicate the results of White et al. (2016) and investigate how threat bias computational measures relate to other computational parameters and to different anxiety symptoms.

Although differences in processing social information is often associated with social anxiety in particular, people with trait anxiety have shown differences in social judgement making and selective attention to social threat tasks. High anxiety groups report less trustworthiness and greater hostility (Willis et al., 2013) from faces and display a selective attention to negative social-evaluative words (Mansell et al., 2002; Taylor and Alden, 2010) compared to positive ones. Within social anxiety, there are numerous studies that have demonstrated

a greater negative bias during social evaluation, where fewer positive words are endorsed (Button et al., 2012, 2015; Koban et al., 2017). The computational mechanisms of this self-evaluation process and the negative self-bias have been investigated in subclinical (**Chapter 4**) and clinical social anxiety (Koban et al., 2017), manifesting computationally in higher self-negative learning rates and a lower trait positive self belief (presented in **Chapter 4**). We aimed to replicate these computational results using our previously developed computational models of social evaluation learning and further aimed to investigate whether the negative self-bias is observed in just social anxiety, or can also be found in other anxiety types as previous results might suggest.

In contrast to the more unconscious processing that underlies the threat bias, cognitive processes such as learning from threat/punishment and safety/reward have also been shown to be altered in people with high anxiety (Wise and Dolan, 2020; Browning et al., 2015). In Browning et al. (2015), people with high trait anxiety were less able to adjust to volatile reward schedules, impacting learning of stimulus-outcome contingencies and manifesting computationally in a lower learning rate difference between stable to volatile blocks. In Wise and Dolan (2020), cognitive anxiety was related to greater updating to aversive outcomes and a lower estimated safety probability. Curiously, this association was reversed for people with somatic anxiety, who generally showed a 'positive bias' for aversive learning, updating less to aversive outcomes and having higher estimates of safety. The underlying reasons for this distinction are unclear, however. We aimed to investigate the aversive learning process using a greater range of specifically tailored anxiety symptom questionnaires in order to elucidate the relationship between updating to threat and safety outcomes in the same paradigm (Wise

and Dolan, 2020).

The process by which people with anxiety learn about rewards and punishments has been explored using paradigms that require the choice between a low and high reward stimulus. However, in real life, individuals who come to expect negative outcomes might begin to avoid that stimulus in order to escape aversive experiences (Bach, 2015). The presence of negative outcomes is thought be especially important if they occur early on in an environment where little is known about the probability of punishment and may precipitate greater avoidance, even if the environment overall is reward-rich (Meacham and Bergstrom, 2016), however this idea has not been directly tested computationally. There are numerous studies that suggest individuals with high levels of anxiety are more predisposed towards avoidance and that this is heightened in the presence of uncertainty, such as when learning is required (Andreatta et al., 2017; Mkrtchian et al., 2017; Sege et al., 2018; Dymond and Roche, 2009). It is important to note that avoidance can be adaptive if the environment is punishment-rich, but becomes suboptimal in environments where the overall reward is higher, as there is an opportunity cost to missing out on rewards lost due to avoidance. In order to investigate the role of early negative outcomes on avoidance behaviour, we created a novel task and series of computational models to investigate our hypothesis that highly anxious individuals would avoid more in environments with early negative outcomes, even if they are overall reward-rich.

An important element in these studies is the role of uncertainty, which has been proposed to play an important role in the pathology of anxiety (Pulcu and Browning, 2019) and can be directly quantified using computational modelling (Mathys et al., 2014; de Berker et al.,

2016; Wise and Dolan, 2020). Using a leaky beta model, Wise and Dolan (2020) was able to show that people with cognitive anxiety overestimate safety uncertainty, whereas people with somatic anxiety underestimate it. The relationship between different forms of uncertainty has not been computationally established and here, we aimed to determine the relationship between uncertainty and anxiety symptoms.

Aims This study firstly aimed to provide a replication of studies that utilise computational modelling in order to investigate four cognitive biases found in people with high anxiety. Secondly, this study aimed to characterise how the crucial computational parameters derived from these four tasks related to anxiety symptomatology measured by psychiatric questionnaires. Lastly, this study had the goal to determine the relationship between uncertainty measures quantified by computational modelling from each task and anxiety symptomatology.

5.4 Methods

5.4.1 Participants.

A total of 451 participants were recruited using the online research platform Prolific (www.prolific.co). We aimed to collect as many participants as resources allowed, however specified a minimum of 440 to allow for the replication of results from the weakest anticipated effect size. The weakest effect size of the computational parameters of interest was the drift-rate to neutral words, which had a Cohen's D of .41. Thus to provide 99% power to detect this statistical effect, the minimum number of 440 was required and reached. The mean age of the sample was 31.16, *SD* (10.94) and there were 167 self-identifying men, 275 self-identifying women and 9 self-identifying non-binary participants. The ethnicity of the sample participants consisted of 370 white, 30 asian, 20 black, 27 mixed-race and 7 other.

The inclusion criteria for the study were people of any gender, aged 18-65 years old, fluent English speaking and a current UK resident. Exclusion criteria were self-reported current treatment for active psychiatric or neurological disorder, self-reported likely brain injury or dysfunction, including likely dysfunction due to COVID-19 illness and a learning disability more severe than 'mild learning difficulty'. We obtained ethical approval for this study from the local University College London Research Ethics Committee (REF 16993/001).

We acquired informed consent from participants in the form of an online checkbox which was obtained before the experiment. All participants received a payment of £8.22 per hour for their time and received a further performance dependent study bonus of £1-2 pounds. The total duration of the study was 2 - 2 and a half hours. We enriched the sample with high anxiety participants, using the trait subscale of the STAI as a measure of trait anxiety. Therefore, 80% of the sample was recruited from the population as usual and a further 20% of the sample were taken from a high trait anxiety subgroup, targeting those scoring equal to or above the top 80% (cutoff of 64) trait anxiety score. Only 1 subject failed more than 1 of our attention checks and was therefore excluded from all analyses. Data failed to save properly for 1 subject during the threat bias task, for 6 subjects during the social evaluation task, for 5 subjects during the early experience avoidance task and for a further 30 for the Aversive Learning 'SpaceShip' Task. A further 4 participants were excluded from the threat bias task for having more than 50% of their trials as fast trials (< 200ms). Therefore, all analyses that consist of a combination of task parameters have a sample size of $n = 404$.

5.4.2 Cognitive Task Battery

We included 4 cognitive tasks in the battery, namely the Social Evaluation Learning Task, the Threat Bias Lexical Decision-Making Task, the Aversive Learning 'SpaceShip' Task and the Early Experience Avoidance Task. The details of the Social Evaluation Learning Task are given in **Chapter 4** and are therefore not repeated here. The order of the tasks were pseudorandomised, with the most enagaging task, the Aversive Learning task always presented last and the other three presented in a counterbalanced order.

5.4.3 Threat Bias Lexical Decision-Making Task

In order to assay Threat Bias, we used a Lexical Decision Making task, published in White et al. (2016) and the same stimuli and instructions were used in our version of this task. In this task, a series of English words were presented in the middle of the screen, one at a time. The words were preceded by a fixation cross (500ms) and the word was displayed on the screen for 2500ms unless a button (Z or M) was pressed to indicate a response (Figure 5.1A). The words were either drawn from a list of 120 threatening words (e.g. death, upset), or 120 non-threatening words (e.g. fringe, leaf), matched for letter length and word use frequency. The threatening words were made up of social threats (e.g. embarrassed, unpopular) or physical threats (e.g. suicide, cancer). The task was divided into 4 blocks of 60 trials each, with 240 trials total. Participants are instructed to classify English words as either threatening or safe, with instructions to choose threatening if the words representing something that was upsetting or worrying. Trials that ended before a response was given were classed as missed trials. Threatening words classified as threatening were marked as hits and if they were classified as safe they were marked as a miss. Non-threatening words classified as

threatening were marked as false-alarms and true-rejections if they were classified as safe.

5.4.4 Aversive Learning 'SpaceShip' Task

In order to assay Aversive Learning, we used the Aversive Learning 'SpaceShip' Task, reported in Wise and Dolan (2020). The task developed to investigate aversive learning was a gamified task, where participants needed to pilot a spaceship by moving up and down the Y-axis in order to avoid asteroid belts that depleted health when hit (Figure 5.1C). Importantly, the asteroid belts had a gap in the rocks that consisted of a safety zone, in which participants could pilot their spaceship through in order to avoid potentially losing points. Although the task was a continuous game experience, it was broken down into 240 trials with each trial consisted of a new asteroid belt and safety location. Safety locations could appear either at the top or bottom of the screen, and the exact position of each safety zone varied according to a safety probability. The probabilities determining the position of the two safety zones were independent and changed over time, meaning the safety position could either be at the top, bottom or both places at the same time. The task for participants was to learn the location of the safety zones so that they could successfully move through the asteroid belts without being hit. Expectations about safety location allowed participants to move their ship towards the expected correct location, which would improve performance. Participants were told to accumulate as many points as possible, indicated by the number of points shown on the screen that was slowly accumulating throughout the task. The only way for participants to lose points was to be hit by the asteroids enough to reduce the ships integrity (indicated by a health bar) to zero, which resulted in the game restarting and points starting again from zero (but keeping the same number of remaining trials). Participants moved the spaceship

using the up and down arrows keys and the task took around 30 minutes to complete.

5.4.5 Early Experience Avoidance Task

In order to investigate the influence of early negative experiences on learning and avoidance behaviour, we developed a new task for use on online platforms. In this task, participants had to decide whether to go for a walk in a novel environment (forage), or to avoid going for a walk and staying home. The stimuli consisted of 120 novel cartoon town environments taken from https://www.freepik.com/. For each environment, participants had to make a number of choices, either 4, 8, 12 or 16, indicated by the horizon shown by footprints at the bottom of the screen (Figure 5.1D). If participants chose the forage option, they received feedback in the form of either a cake (reward) or a virus (punishment). If they selected to stay home and avoid going outside, they received no feedback as to whether there was a reward or punishment and thus gained no information. Participants were instructed to gain as many cakes as possible, as thus would result in more points, and avoid the viruses, which would lose them points. Therefore, participants had to determine which environments were the reward-rich environments and explore those environments more, whilst avoiding the punishment-rich environments. Participants were informed prior to the task that reward-rich environments contained more 75% rewards and 25% punishments and vice versa for the punishment-rich environment. Importantly, we manipulated the distribution of outcomes in each environment, such that the distribution of punishments were either early on in the environment (early-negative), evenly distributed (middle-negative), or later in the environment (late-negative). Thus, in an early-negative environment, the outcomes of the first few trials are more likely to deliver a punishment, despite the overall probability of punishment being either 25 or 75%. This ma-

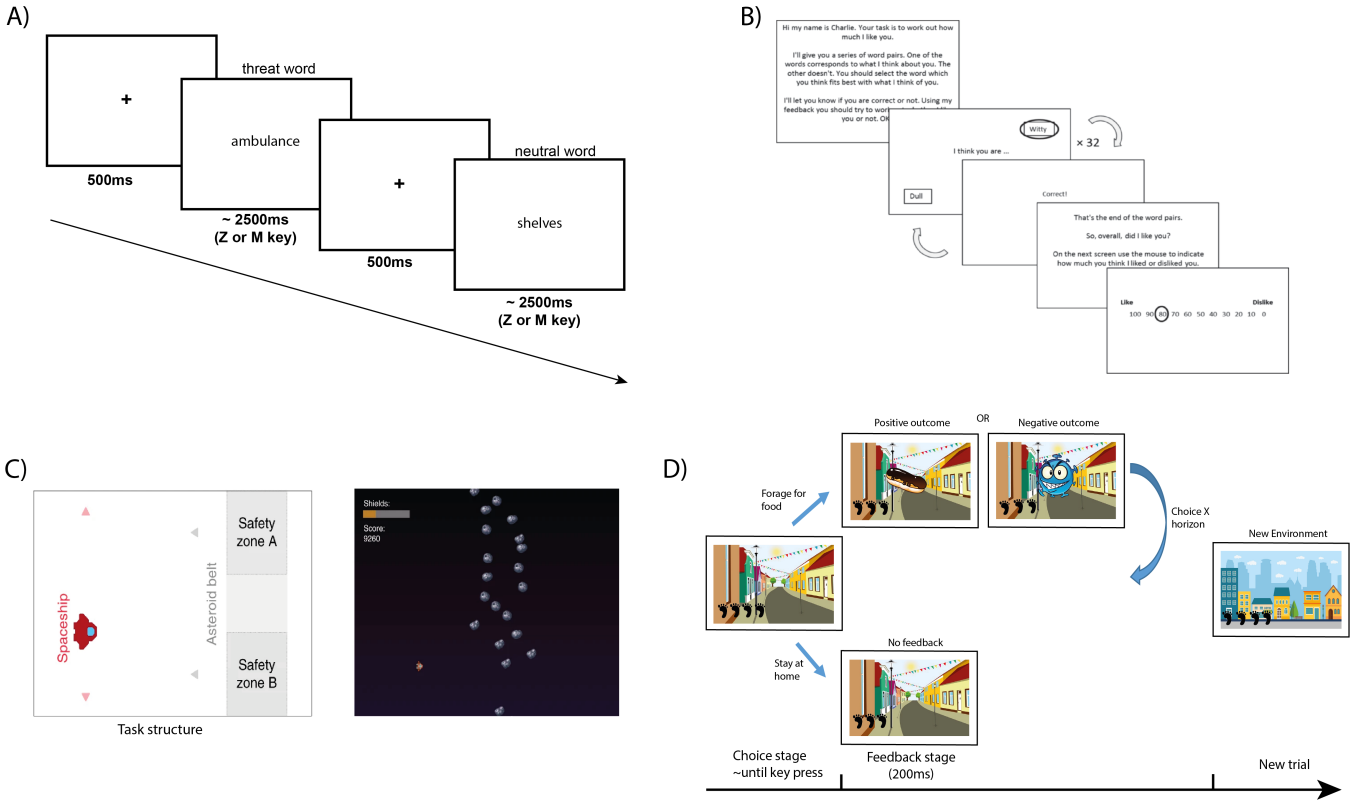

Figure 5.1: A)Threat Bias Lexical Decision-Making Task B) Social Evaluation Learning Task, C) Aversive Learning 'SpaceShip' Task D) Early Experience Avoidance Task

nipulation allowed us to test our main hypothesis that early negative experiences in particular will promote avoidance even in environments that are overall reward-rich. Participants were informed that each novel environment is independent and unrelated to previous environments, thus they had to learn the reward/punishment contingencies each trial. There were 120 trials in total, split into 4 blocks with an even distribution of good and bad environments in each block.

5.4.6 Questionnaire battery

As the main aims of this experiment were to investigate the relationship between different computational parameters and anxiety symptomatology, we included a range of anxiety questionnaires to cover a diverse range of anxiety presentation. To measure state and trait

anxiety, we included the STAI (Spielberger et al., 1983), to measure cognitive and somatic anxiety, we included the STICSA (Grös et al., 2007), to measure social anxiety, we included the the brief fear of negative evaluation scale (BFNE) (Leary, 1983), to measure intolerance of uncertainty, we included the intolerance of uncertainty scale (IUS) (Boswell et al., 2013) and to measure anxiety sensitivity, we included the Anxiety Sensitivity Scale-3 (ASI) (Taylor and Montgomery, 2007). As anxiety is largely comorbid with depression, we also included the BDI-II (Beck et al., 1961). Sub-aims of our study were to investigate anxiety in relation to real world measures of stress and uncertainty, therefore we also included measures of education level, household income and employment status, as well as the Perceived Stress Scale (PSS) (Cohen et al., 1983). We further included questionnaires to measure demographic variables including gender, age and ethnicity.

5.4.7 Preregistration and analysis plan

This study was pre-registered on the Open Science Framework [https://osf.io/z4rgu/], which contains details for the main hypotheses of this study, as well as the proposed analysis plan. The analysis plan consisted of first performing a replication analysis for each task to attempt reproduce the original reported behavioural and computational results. Alongside traditional statistics, we also report Bayes Factor 10 scores (BF) where possible, as these scores give information about the strength of evidence for both the null and alternative hypothesis (Jeffreys, 1998). BF scores are positive numbers, with numbers above 1 numbers indicating support for the alternative hypothesis and numbers below 1 indicating support for the null hypothesis. Following the replication analysis, we aimed to determine the relationships between parameters both across tasks and from tasks to symptoms using a combination of factor analysis and PLS regression, detailed below.

5.4.8 Computational Modelling and Key Parameters

For each task, we fit the best-fitting computational models reported in the previously published work. To investigate threat bias, we fit the SDT model and DDM reported in White et al. (2016). To investigate negative self-bias, we fit the best-fitting associative learning and belief-update models (described in detail in **Chapter 4** and therefore omitted in the methods section here). To investigate aversive learning, we fit the leaky beta model reported in Wise and Dolan (2020). To investigate early negative experience and avoidance, we fit a similar leaky beta model, with full description given in text and full model space in the Supplementary Information.

Signal Detection Model The SDT model calculates measures of stimulus discriminability and bias, labelled as d' and criterion respectively. Larger values of d' indicate a greater ability to discriminate between neutral and threatening stimuli. The key criterion measure, c, which indexes bias, was calculated as $-0.5 * (z[hit\ rate] + z[false\ alarm\ rate])$. Larger values of c indicate a higher threshold for which stimuli are labelled as threatening (lower bias towards threat).

Drift Diffusion Model The DDM was fit using the fast-DM package (Voss and Voss, 2007). The data entered into the fitting procedure were participant RTs (measured in seconds) and response accuracy (correct vs incorrect). The DDM consisted of 8 parameters in total, with 5 parameters representing the key decision making variables which were a = boundary separation, z = starting point, Ter = non-decision time, v_t = drift rate for threatening words, v_n = drift rate for neutral words. There were additionally 3 parameters that capture variability in the decision components, s_t = variability in non-decision time, s_z = variability in starting

point, *eta* = variability in drift rate. Model fit was assessed using Chi-Square (χ^2) estimation. Full descriptions of the model and the parameters are given in the SI.

Leaky beta models Both the Aversive Learning and the Avoidance Learning tasks were modelled using leaky beta models (Wise et al., 2019; Wise and Dolan, 2020). For the Aversive Learning task, we used the best-fitting leaky beta model reported in Wise and Dolan (2020) and adapted this model for the Avoidance Task. Importantly, similar to the belief-update models reported in **Chapter 4**, these models allow for the quantification of uncertainty, as they not only give point estimates of value, the beta parameters quantify an entire belief distribution, allowing the mean and variance of the distribution to be quantified trial-by-trial. They also allow asymmetrical value updating, through the separation of updating parameters, τ, for threat/safety (τ_{threat}, τ_{safety}) or gains/losses (τ_{gain}, τ_{loss}).

The final leaky beta model fit to the Aversive Learning task was an asymmetric leaky beta model. This model updates the α and β parameters of a beta distribution (here A and B) separately for each safety zone, X. The asymmetry of the model is reflected in the separation of the update τ parameters, for threat and safety (τ_{threat}, τ_{safety}, see eq 44). In the context of this task, threat represents hitting an asteroid and safety represents safely passing through the asteroid belt. The λ parameter is a forgetting rate, which reduces the A and B parameters and the W parameter is a weight parameter, which downweights the unchosen option and is set to 1 for the chosen option.

$$
\begin{aligned}
A^X_{t+1} &= (1-\lambda)\cdot A_t + \tau^{safety}\cdot outcome^X_t \cdot W \\
B^X_{t+1} &= (1-\lambda)\cdot B_t + \tau^{threat}\cdot (1-outcome^X_t)\cdot W
\end{aligned} \tag{44}
$$

The model also incorporated a stickiness parameter, $0 < S < 1$, which increased the value of the previously chosen option to make the same choice more likely (eq. 45).

$$
\begin{aligned}
P^X_{t+1} = P^X_{t+1}{}^S if chosen \\
P^X_{t+1} if unchosen
\end{aligned} \tag{45}
$$

The leaky beta model for the aversive learning task was similar in form to this model, but with some modifications and conceptual differences. The beta distribution here represents the beliefs that participants hold about how reward-rich the environment is. The evidence for positive and negative outcomes in each environment, X, are represented by α and β parameters respectively, with the value of each being updated every trial, t, according to the feedback received, multiplied by a τ parameter, that weights recent outcomes more or less, depending on its fitted value and split by valence, with τ^{gain} updating positive outcomes (reward) and τ^{loss} updating negative outcomes (loss). To avoid nonsensical distributions of probabilities, such as bimodal distributions that can arise when the α and β parameters drop below 1, we gave a lower bound to these quantities by ensuring they did not drop below 1. We further included an a priori parameter -1 < apriori < 1, which biased the starting Q values for each environment towards the action forage or avoid. This parameter acted as a bias for each environment for the initial trial only.

$$\begin{aligned} \alpha_{t+1}^{X} &= (1-\lambda)\cdot\alpha_t + \tau^{gain}\cdot outcome_t \\ \beta_{t+1}^{X} &= (1-\lambda)\cdot\beta_t + \tau^{loss}\cdot(1-outcome_t) \end{aligned} \tag{46}$$

The expectation of a positive outcome is then computed by calculating the mean of the beta distribution (eq. 46), with the probability of choosing to avoid given as the inverse (eq. 47). Similar to the Aversive learning paradigm, we also included a stickiness parameter, *S*, which was an exponential applied to the probabilities of the actions (eq. 48), allowing for a stickiness parameter for foraging (S^{forage}) and a stickiness parameter for avoiding (S^{avoid}). Smaller values of *S* indicate greater choice stickiness. To account for differences in loss sensitivity, we also allowed *Rloss* to be a free parameter in some models, which allowed the value for loss to vary for individuals, thus, greater *Rloss* values suggest greater sensitivity to losses (eq. 49)

$$P(X) = \alpha^X/(\alpha^X+\beta^X) \tag{47}$$

$$P(\bar{X}) = (1-P(X)) \tag{48}$$

$$Q^X = (P(X)^{Sforage})\cdot Rgain + (P(\bar{X})^{Savoid})\cdot Rloss \tag{49}$$

Finally, the Q values for each environment for each trial were fed into a standard softmax function, with an inverse temperature parameter (eq. 50).

$$p(action) = \frac{1}{1+exp-\frac{(Q(forage)-Q(avoid)}{\tau}} \tag{50}$$

Model Fitting and Estimation The model fitting procedure reported in White et al. (2016) was a custom built model, fit using Matlab and we were not able to replicate this exact procedure due to lack of available analysis code, therefore we fit the drift-diffusion model using the fast-DM package (Voss and Voss, 2007). Fast-DM is used for fitting drift-diffusion models to binary decision tasks, and fits the same parameters as those reported in White et al. (2016), with an additional parameter that measures contaminated trials, which was set to 0. Although the same parameters can be estimated, the bounds on the fast-DM package cannot be altered, therefore the absolute values of the parameters obtained from this fitting procedure cannot be compared to values reported in White et al. (2016). Model fit was assessed using χ^2.

For social evaluation learning, we used the same model fitting procedure reported in **Chapter 4**, first fitting the models using maximum likelihood estimation, using the fmincon function in Matlab and then fitting the models hierarchically using MCMC implemented in rStan. Model fit was assessed using LOO. The Aversive Learning leaky beta model was fit by my coauthor Dr Toby Wise, using the package PyMC3 and using a variational inference procedure. Model fit was assessed using Watanabe-Akaike Information Criterion (WAIC). The Avoidance leaky beta model was fit using MLE implemented using the fmincon function in Matlab. Model fit was assessed by examining AIC, BIC and pseudo-r^2 measures.

Key Parameters Although we obtained estimates for all parameters of the model, for our main analyses, we were interested in the parameters that had associations with anxiety (Table 5.1). For threat bias, this was the criterion parameter, c, derived from the SDT and threat drift rate v_t, neutral drift rate v_n, and starting bias z/a parameters from the DDM. For the negative self-bias, we were interested in both the self-negative learning rate, λ_{self-},

Task	Model	NP	Parameters	Key Parameters
Threat Bias	SDT	2	d', c	c
	DDM	8	a, Ter, z/a, v_t, v_n, var_v, var_z, var_{ter}	v_t, v_n, z/a
Social Evaluation	Associative Learning	6	λ_{self+}, λ_{self-}, λ_{other}, init, p, τ	λ_{self-}
	Belief-Update	8	α_{self}, β_{self}, α_{other}, β_{other}, α_{init}, β_{init}, mem, τ	α_{self}
Aversive Learning (SS)	Leaky beta	6	λ, S, τ_{safety}, τ_{threat}, β, w	τ_{safety}, τ_{threat}
Avoidance	Leaky beta	7	λ, S, τ_{gain}, τ_{loss}, β, $apriori$, $Rloss$	τ_{gain}, τ_{loss}

Table 5.1: Task and computational models, alongside key computational parameters.

from the associative learning model and the trait self-positive belief parameter, α_{self}, from the belief-update model. For aversive learning, we were interested in the safety, τ_{safety} and threat τ_{threat} update parameters from the leaky beta model. As the avoidance task was a novel paradigm and model, we first performed an exploratory analysis to determine which parameters, if any, had a relationship with anxiety, but had an *a priori* focus on the two update (τ) parameters for gain τ_{gain} and loss τ_{loss}.

Uncertainty and model-derived quantities Alongside the aforementioned key parameters, we wanted to investigate the relationship between uncertainty, model-derived quantities such as value and anxiety. For social evaluation, uncertainty for self and other was computed by taking the variance of the beta distribution free parameters for both self and other. Mean approval for both self and other was calculated as the mean of the beta distributions. For the aversive learning task, we computed the mean safety probability and the mean safety uncertainty. For the avoidance task, we computed the mean value for good and bad environments and the mean uncertainty for good and bad environments. We obtained these measures of uncertainty and other model-derived quantities for each task for each participant by simulating data from the aforementioned learning models using the best-fitted parameter values for each participant. As the Threat Bias task is not a learning task, we calculated uncertainty as the variability of the (log) RTs.

5.4.9 Bayesian Regression Models

To investigate whether model parameters and model-derived quantities of interest, such as uncertainty had any relationship to anxiety and mood symptom dimensions, we implemented Bayesian linear regression models, with age and gender entered as covariates (Wise and Dolan, 2020; Patzelt et al., 2019). As our gender variable had three levels, but one of our levels only had 9 participants, we opted to remove the third level from our analyses and treat the gender variable as a binary categorical variable, as there was not enough power to detect any effects for the third category. Bayesian linear regression models are similar to non Bayesian general linear models, but they have the additional advantage of computing a posterior distribution for the parameter estimate, which gives a measure of uncertainty for the beta values. Here, we report the 95% credible intervals for each beta estimate, which give the 0.025 and 0.975 quantiles of the MCMC sample. We implemented the Bayesian regression models using the bayeslm and estimate functions in Matlab. We first constructed a prior model for each of the parameters, using a Diffuse Prior model with uninformative Jeffreys priors, so that the joint prior density over regression weights and variance was simply proportional to the precision. We implemented both independent regression models (where predictors were entered separately in independent regression models) as well as combined regression models (where predictors were entered together into the same model) in order to examine the effect of the shared variance, similar to analyses reported in Gillan et al. (2016); Patzelt et al. (2019).

5.4.10 Factor Analysis on Questionnaires

In order to investigate anxiety and mood symptom subtypes and their relationship to the key computational parameters, we performed a factor analysis on the item-level questionnaire

data. We removed the STAI from this analysis and used the STICSA as the measure of trait anxiety, as it allowed for decomposition of cognitive and somatic subscales, whereas the STAI only allows for a trait/state distinction. We performed the factor analysis in R on 105 questionnaire items, using the factanal() function, using MLE. We used an oblique rotation, which allows for correlations between factors. In order to assess the best number of factors, we defined an empirical elbow using Cattell's criterion (Cattell, 1966), obtaining an empirical estimate of the elbow using the Cattell-Nelson-Gorsuch (CNG) test (Gorsuch and Nelson, 1981), implemented in R using the nCng() function. Based on the difference in slopes of the eigenvalues, the CNG test gave a three-factor solution to our data (Figure 5.8). We discuss the factors in more detail in the results section. In order to determine the relationship between these symptom dimensions and the key computational parameters, we performed Bayesian regression models with each of the dimensions entered into the model independently. We performed a further analysis to assess the specificity of the results when removing the shared variance, by entering the symptom dimensions as predictors together in the same model.

5.4.11 Factor Analysis on Parameters

In order to investigate the latent structure between parameters independently of anxiety symptoms, we performed an exploratory-confirmatory factor analysis (ECFA) approach using our key parameters as dependent variables. In order to prevent issues regarding overfitting, we divided the dataset into a train and test set using 75% and 25% of the data respectively. For the exploratory FA, we assessed the best number of factors again using the factanal() function in R. The maximum number of factors to test in the CFA was determined by Parallel Analysis. For the confirmatory FA, we tested the models estimated from the EFA on the

test dataset using SEM. We used criteria of under-determination, BIC and CFI to compare increasing number of factors and determined whether these model fit statistics improved with an increasing number of factors. We chose the best factor solution based on our fit criteria and the convergence of the SEM.

5.4.12 PLS regression analysis

Following our pre-registered analysis plan, we sought to further understand the relationships between our key parameters and psychopathology using a PLS data reduction approach (Wise and Dolan, 2020). As PLS and other data reduction approaches are susceptible to overfitting to the data, we used a 'k-fold' cross-validation approach, using 10-folds, to ensure our results were generalisable. We first created a train and test set using 75% an 25% of the data respectively. On the training set, we investigated how many components were the best number of components by fitting 1 to 6 components to the data. We evaluated the performance of each component number using the MSE of the predictive accuracy across 10-folds of validation. We then checked the performance of the best component number on the held out data test set. The normalised key parameters were entered into the PLS as dependent variables and all of the questionnaire items were entered as predictor variables. To assess the significance of the parameter and item loadings, we performed a bootstrap analysis by generating 1000 datasets of shuffled values at the subject level. We sampled with replacement to generate the datasets and then computed bootstrap confidence intervals.

	Proportion Threat Response		Median Correct RT (ms)	
Anxiety Group	Neutral Words	Threat Words	Neutral Words	Threat Words
Low	0.10 (0.12)*	0.74 (0.22)**	752 (155.4)	783 (136.37)
High	0.12 (0.14)*	0.81 (0.18)**	782 (158.45)	763 (131.93)

Table 5.2: Proportion threat response for threat and neutral words and median RTs for correct responses split by low (n = 183) and high (n = 179) anxiety groups. * indicates significance <.05, **indicates significance <.01.

5.5 Results

5.5.1 Threat Bias

Behavioural Results Following the analysis reported in White et al. (2016), we split the data according to the STAI trait anxiety scores for the top 60% and bottom 40% of our sample. This resulted in a low anxiety group with n = 183 and a high anxiety group of n = 179. We replicated the effect of anxiety on threat classification (Table 5.2), with the high anxiety group significantly more likely to classify words as threatening. This was true for threat words $t(360) = 3.10, p < .01$, BF = 6.23 and neutral words $t(360) = 2.16, p < 0.05$, BF = 0.54 (5.2A). We also replicated the correlational analysis that showed a positive correlation between trait anxiety and threat response proportion for threat words, Spearman's $rho = 0.17, p < .001$ as well as for neutral words, Spearman's $rho = 0.13, p < .01$. A RM-ANOVA on the median correct RTs revealed no main effect of anxiety on correct RTs $F(1,360) = .14, p = .71, BF = 0.06$ and no main effect of threat on correct RTs $F(1,360) = .83, p = .36, BF = 0.08$. However, we did replicate the significant interaction between anxiety and threat level RTs $F(1,360) = 13.80, p < .001, BF = 47.60$, with low anxiety responding more quickly to neutral words and more slowly to threat words, compared to their high anxiety counterparts (see Table 5.2 for median RTs).

	SDT					DDM					
Anxiety Group	d'	c	a	Ter	z/a	v_t	v_n	var_v	var_z	var_{Ter}	Fit χ^2
Low	2.3 (0.79)	0.38(0.61)***	0.47 (0.31)	515.18 (112.56)	0.48 (0.14)	1.9(2.54)**	3.3 (2.65)	1.88 (2.18)	0.18 (0.26)	0.22 (0.12)	17.79 (10.30)
High	2.5 (0.82)	0.19 (0.54)	0.49 (0.33)	520.60 (110.02)	0.49 (0.14)	2.8 (3.54)	3.2 (3.62)	1.93 (2.08)	0.24 (0.29)	0.23 (0.15)	16.92 (11.21)

Table 5.3: Computational parameter differences from both the SDT and DDM models, split by low and high anxiety groups. * indicates significance <.05, ** indicates significance <.01, ***indicates significance <.001.

SDT and DDM modelling results We replicated the effect of trait anxiety on the criterion, *C* parameter, with high anxiety participants showing a significantly weaker criterion parameter than low anxiety participants $t(360) = -3.16, p < .01, BF = 7.38$. We also replicated the correlational analysis $r = -0.17, p < .001$ that further showed a negative relationship between trait anxiety and criterion. We replicated the null result for the d' parameter, with no difference between anxiety groups $t(360) = 1.05, p = 0.29, BF = 0.09$, as well as replicating the lack of correlation between trait anxiety and d' $r = 0.01, p = .28$. Overall there is strong evidence that trait anxiety participants display a weaker criterion for classifying words as threatening and no evidence to suggest that they have a greater ability to discriminate between neutral and threatening stimuli.

We replicated the effect of trait anxiety on threat drift rate, with high anxiety participants showing a significantly greater threat drift rate $t(360) = 2.69, p < .01, BF = 1.93$ than low anxiety participants. Correlational analyses further supported this association $r = 0.13, p <$ .01. We did not replicate the difference between the anxiety groups for either the neutral drift rate $t(360) = -0.23, p = .82, BF = 0.05$ or starting bias $t(360) = -0.53, p = .59, BF = 0.06$ parameters and correlation analyses further supported these null results ($r = -0.02, p = .73$ and $r = 0.002, p = .96$ respectively).

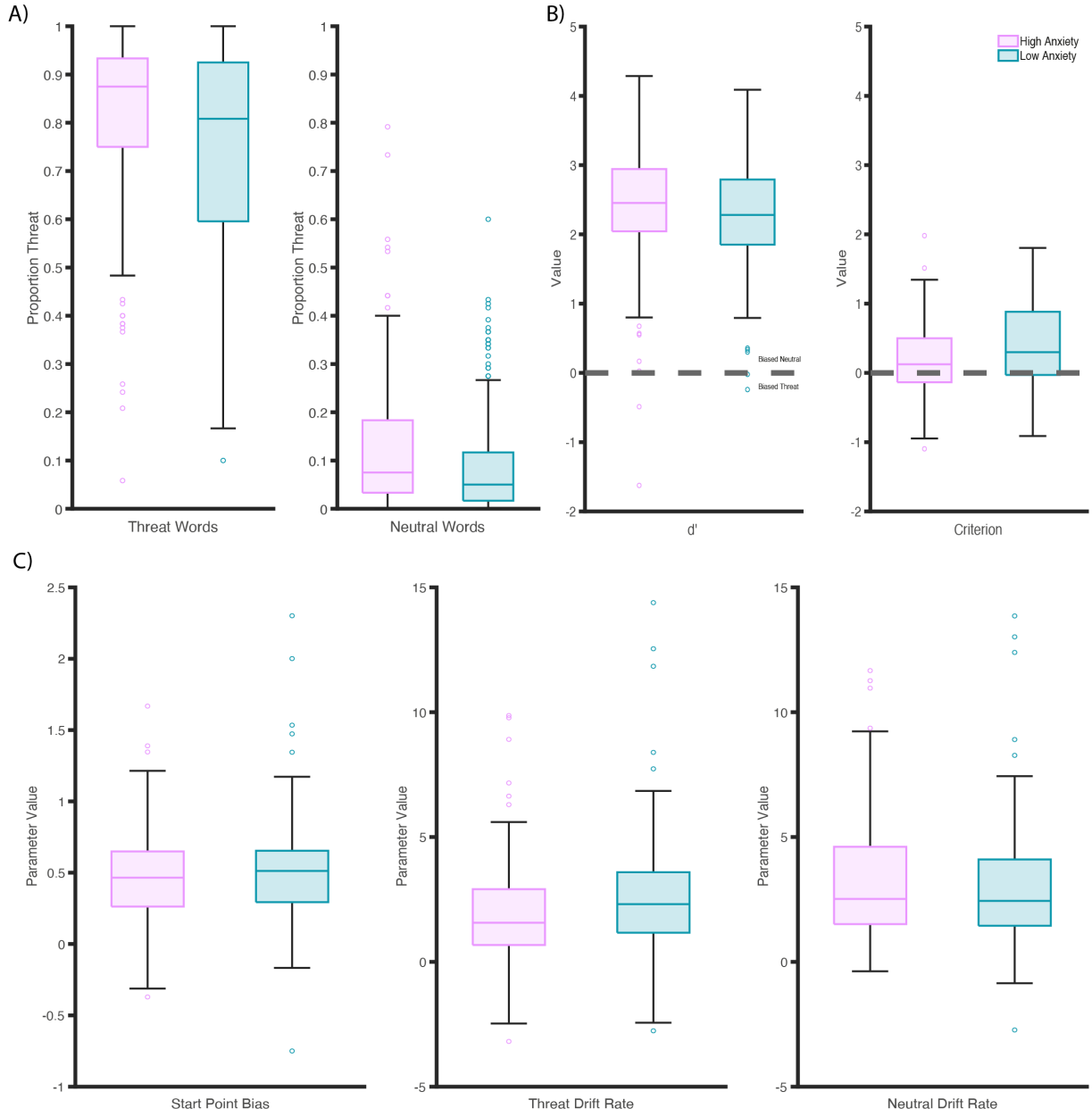

Figure 5.2: A) Proportion of threat responses given for both true threat and true neutral words. B) Boxplots displaying the SDT model parameters, d' and criterion, c, split by high and low anxiety participants. C) Boxplots displaying the key DDM parameters, starting bias, threat drift rate and neutral drift rate, split by high and low anxiety participants.

Exploratory Bayesian Regressions Although we aimed to explore the key anxiety-related parameters highlighted in White et al. (2016), we also wanted to determine the relationship between anxiety and all of the model parameter results in our sample. We

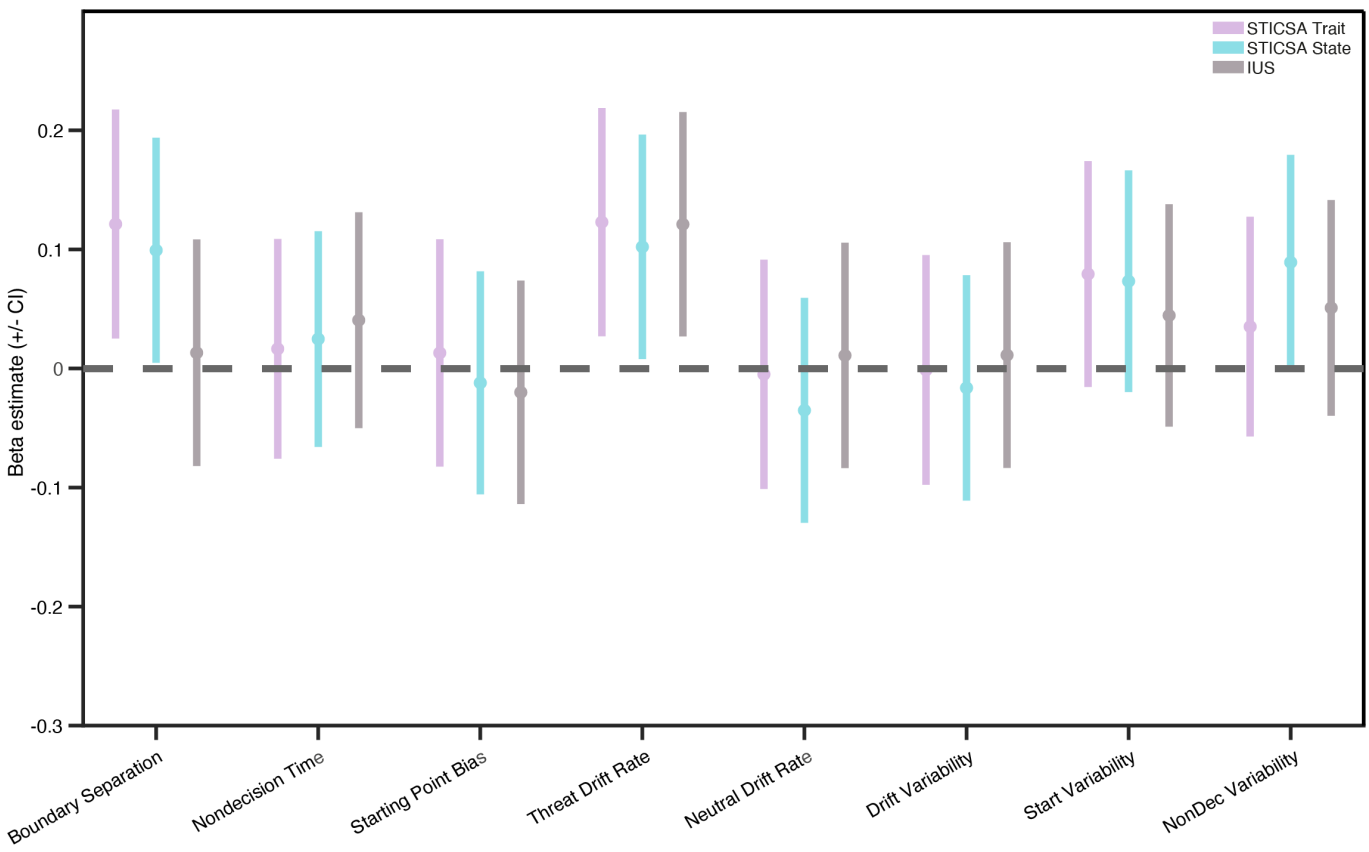

Figure 5.3: Results of the Bayesian regression of trait and state anxiety and IUS as predictors for each of the model parameters. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

performed an exploratory analyses of all of the model parameters and their associations with trait and state anxiety and covariate demographics such as age and gender using Bayesian regression. Trait and state anxiety (as measured by the STICSA) were related to greater boundary separation (M = 0.12, CI [0.03 0.22], M = 0.10 CI [0.001 0.19] respectively), meaning they tended towards response caution. Trait, state and IUS measures were related to higher threat drift rate, further supporting the results of the aforementioned results (M = 0.13, CI [0.03 0.22], M = 0.10 CI [0.01 0.20], M = 0.12, CI [0.03 0.22] respectively).

5.5.2 Aversive Learning

We performed Bayesian regression in order to determine whether the key model parameters, as well as uncertainty and mean safety probability had a relationship to anxiety. We fit the models using the 'estimate' function in Matlab, which performs MCMC using HMC estimation, with 2000 burn-in samples and 8000 draws. None of the key variables of interest were significantly associated with age or gender demographic variables (Supplementary Information) and we included these variables as covariates in subsequent analyses. The results of the Bayesian regression analyses revealed a significant effect of trait ($m = -0.10$, CI[-0.20 0.00]) and state anxiety ($m = -0.10$, CI[-0.21 -0.01]) on threat updating, with greater anxiety associated with reduced threat updating shown in Figure 5.4A. There were no further associations with the safety update parameter, the mean safety uncertainty or mean safety probability quantities. The main results of Wise and Dolan (2020), surrounded the split between Cognitive and Somatic anxiety, however we failed to replicate this distinction. The results of the subscale analyses displayed in Figure 5.4B, reveal a similar pattern across the Cognitive and Somatic subscales, with significant associations with the threat update parameter for the Trait Somatic (M = -0.10 CI [-0.20 -0.001]), State Cognitive (M = -0.08 CI [-0.21 -0.01]) and State Somatic subscales (M = -0.10 CI [-0.20 -0.001]).

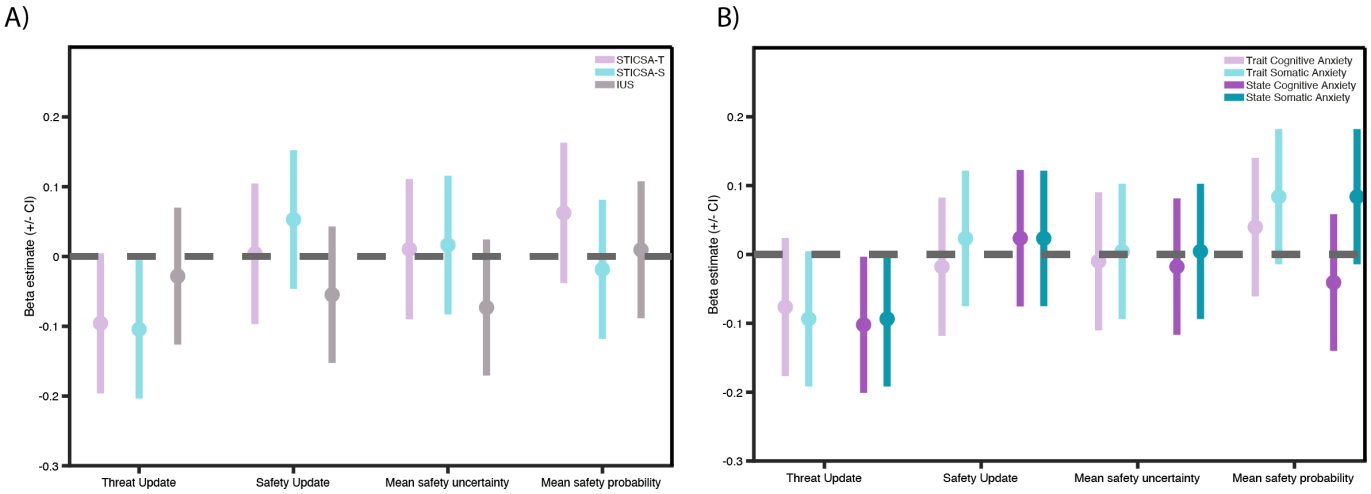

Figure 5.4: Results of the Bayesian regression of predictors A) STICSA trait, state and intolerance of uncertainty and B) STICSA cognitive and somatic subscales for key variables of interest. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

5.5.3 Avoidance Learning

Behavioural Results To test for differences between high and low anxiety, we split the data according to the STAI trait anxiety scores for the top 60% and bottom 40% of our sample. This resulted in a low anxiety group with n = 185 and a high anxiety group of n = 188. To test the hypothesis that early negative experiences are associated with increased avoidance, we performed a mixed ANOVA, with the probability of choosing to avoid as the measurement variable and environment type (good vs bad), outcome distribution (late negative, middle negative and early negative) as within-subjects factors and anxiety group as between-subjects factor and age and gender as covariates. As expected, there was a significant main effect of environment type, with people choosing to avoid more in the bad environments in general, $F(1,356) = 71.18, p < .001$ (Figure 5.5A). As hypothesised, there was also a significant main effect of outcome distribution, $F(1.55,551.46) = 39.79, p < .001$ (Greenhouse-Geisser corrected) with people choosing to avoid more for early negative environments and least for late negative environments. There was also a significant environment X outcome distribution

interaction $F(1.99, 708.47) = 10.81, p < .001$. Contrary to our hypothesis, there was no main effect of anxiety group on the probability to avoid $F(1, 356) = 1.30, p = .26$ or any further interactions between anxiety and environment or distribution.

We also analysed log RTs in relation to environment type (Figure 5.5B), outcome distribution and anxiety group using a mixed ANOVA with age and gender as covariates and high and low anxiety group as a between subjects factor. There was no significant effect of environment type $F(1, 369) = 2.49, p = .12$ or distribution $F(1.96, 724.49) = .68, p = .51$(Greenhouse-Geisser corrected), but there was a significant environment X distribution interaction $F(2, 737.33) = 3.89, p < .05$, with middle and early negative outcome distributions displaying faster RTs for bad environments. There was also an overall main effect of anxiety, with the highly anxious group significantly slower in RTs $F(1, 369) = 4.88, p < .05$.

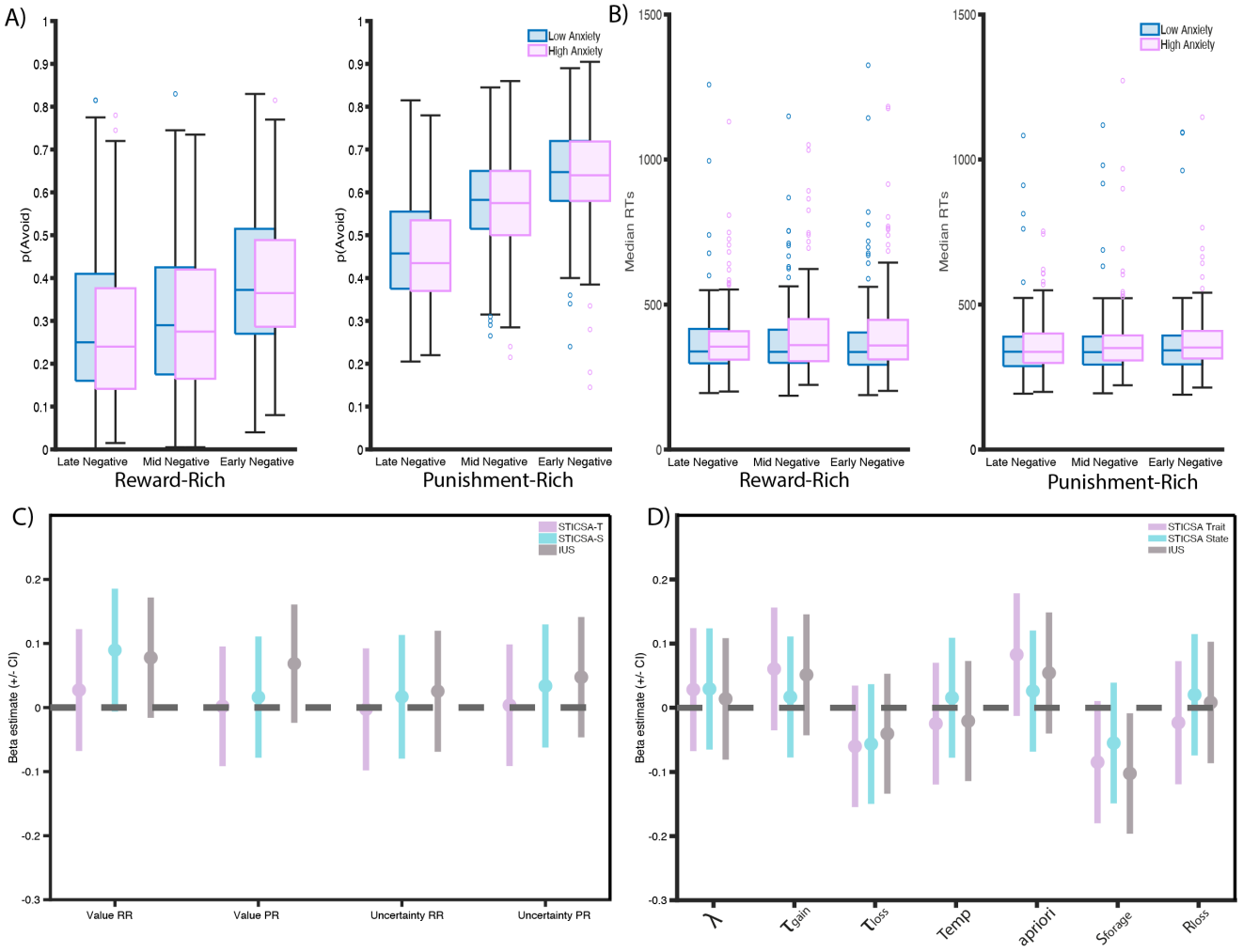

Figure 5.5: A) Probability of choosing to avoid for each environment X distribution type, split by high and low anxiety groups. B) Median RTs for each environment X distribution type, split by high and low anxiety groups. C) Bayesian regressions for model derived quantities, mean value for reward-rich (RR) and punishment-rich (PR) environments and mean uncertainty for RR and PR environments. D) Bayesian regressions for best-fitting model parameters with trait and state anxiety and intolerance of uncertainty as predictors. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

Computational Modelling of Avoidance To understand the computational mechanisms of avoidance learning, we fit a number of computational models $(n = 12)$ using Maximum Likelihood Estimation (MLE), implemented in Matlab, using the fmincon function. We fit both AL and leaky beta models as we wanted to determine if models that explicitly capture uncertainty better capture the avoidance learning process. The best fitting model according to BIC (Table 5.4) was a leaky beta model with an initial bias parameter, separate Tau parameters for positive and negative outcomes (τ^{gain}, τ^{loss}), a stickiness parameter for the forage option (S^{forage}), a memory parameter (λ) and a temperature parameter.

The advantage that the leaky-beta models had over the AL models suggests that a model that explicitly described uncertainty was needed in order to capture the learning process. The best-fitting leaky beta model suggested that separate update parameters for positive and negative outcomes were required. Moreover, the model comparison gives (marginal) evidence against separate stickiness parameters and suggests that repeat decisions were required for choosing to forage rather than for choosing to avoid.

We performed Bayesian regression in order to determine whether any model parameters had a relationship to anxiety. We fit the models using MCMC, with 2000 burn-in samples. Contrary to our hypothesis, there was no relationship between trait and state anxiety or intolerance of uncertainty and the threat update τ_{loss}, parameter. The only parameter to

Model Family	Name	NP	LL	AIC	BIC	Pseudo r^2
RW	Simple	4	-274440.60	552449.21	561529.91	0.26
RW	Valence	5	-272350.84	549161.67	560512.54	0.27
RW	Valence + sticky	6	-253799.48	512950.96	526572.00	0.32
RW	Valence + 2sticky	7	-253328.86	512901.72	528792.94	0.32
Beta	Simple	5	-284107.08	572674.16	584025.03	0.23
Beta	Valence	6	-261023.86	527399.72	541020.77	0.30
Beta	**Valence + sticky**	**7**	-247845.36	**501934.72**	**517825.94**	**0.33**
Beta	Valence + 2sticky	8	-247568.30	502272.61	520434.00	0.33

Table 5.4: Fit statistics from MLE for the 4 associative learning models and 4 of the beta models that included the *Rloss* parameter. The best fitting model is indicated in bold.

display a symptom relationship was the stickiness, S^{forage}, parameter, with individuals who had greater intolerance of uncertainty displaying greater choice stickiness (lower values represent higher stickiness) (Figure 5.5D), (M = -0.11, CI [-0.21 -0.01]), representing a greater tendency to repeat forage choices, possibly reflecting an information gathering tendency.

5.5.4 Social Evaluation

Behavioural Results The results of the linear mixed model with only main effect terms revealed a significant replication effect of rule (like vs neutral vs dislike) on positive responses (dislike as the reference category), with the like and neutral conditions associated with significantly greater positive responses ($\beta = 52$, 95% CI [50.41 53.59] $p < 0.001$ and $\beta = 25.86$, 95% CI [24.26, 27.45], $p < 0.001$ respectively). There was no main effect of self/other condition $\beta = -0.42$, 95% CI [-1.72 0.88], $p = 0.53$ or of FNE $\beta = -0.03$, 95% CI [-0.10 0.04], $p = 0.43$, contrary to expectations.

The results of the linear mixed model with interaction terms again revealed a significant replication effect of rule (like vs neutral vs dislike) on positive responses (dislike as the reference category), with the like and neutral conditions associated with significantly greater positive responses ($\beta = 50.18$, 95% CI [33.25 67.12], $p < 0.001$ and $\beta = 19.85$, 95% CI [2.92, 26.79], $p < 0.05$ respectively). We failed to replicate the other main effects and interaction effects of interest, namely there was no main effect of FNE ($\beta = -0.09$, 95% CI [-0.37 0.19], $p = 0.54$), no self/other X FNE interaction ($\beta = 0.007$, 95% CI [-0.17 0.18], $p = 0.93$), and no three-way self/other X rule X FNE interaction ($\beta = -0.04$, 95% CI [-0.21 0.28], $p = 0.77$) (although the original result was only at trend level). Overall, we failed to replicate most of the results relating to FNE in this sample. However, the forthcoming

analyses suggests that social evaluation learning in this sample was more strongly related to a range of anxiety symptoms, rather than FNE specifically .

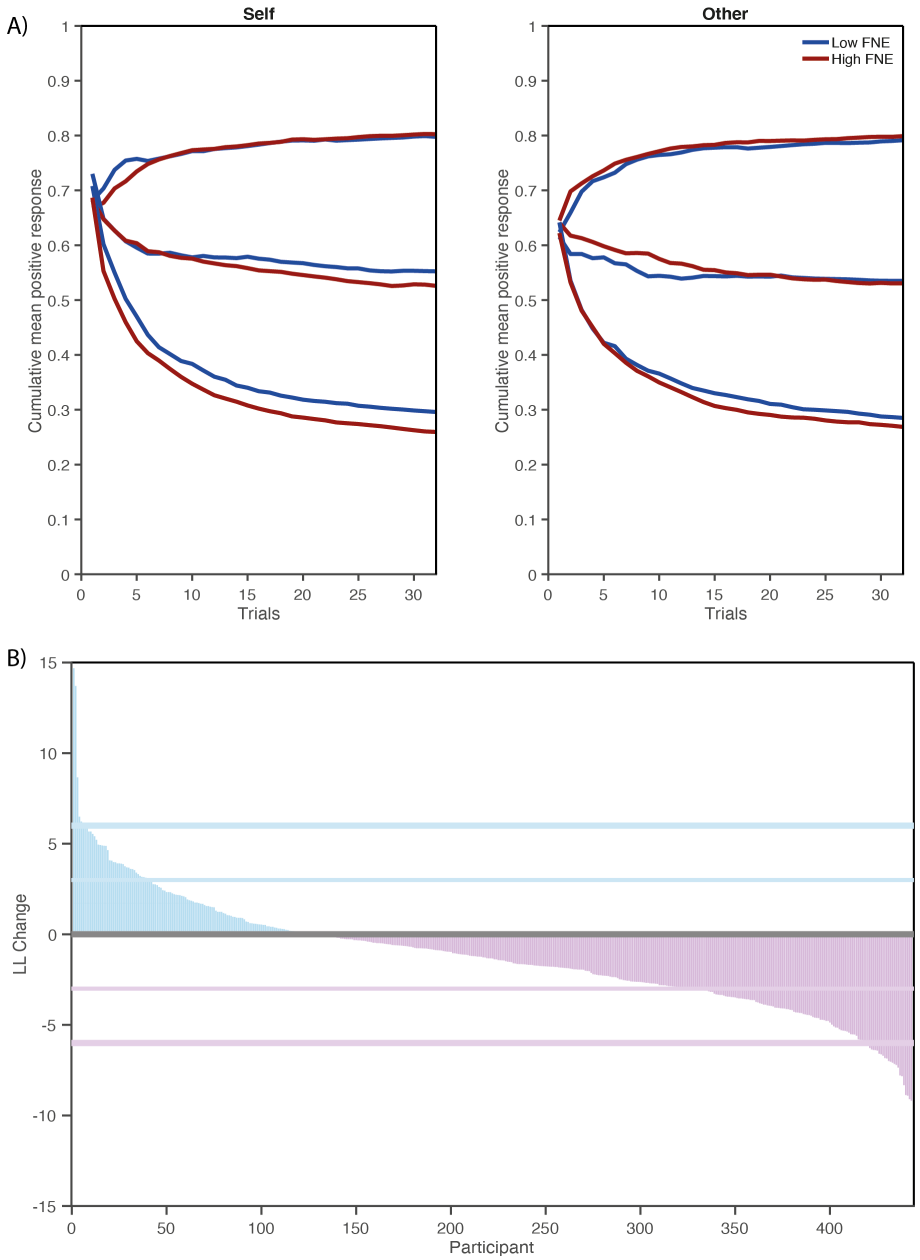

Figure 5.6: A) Cumulative mean positive responses visualised by a median split on BFNE. B) The difference in individual log likelihoods given by MLE fits for the associative and belief-update models. Positive (blue) values represented evidence for associative and negative (purple) represents evidence for belief-update model. The horizontal lines represent mild and strong evidence. The majority of participants (72% of sample) were best represented by the belief-update model.

Computational Modelling Results We first fit the associative and belief-update models using MLE in order to examine whether one model better captured the social evaluation process at the population level. As seen in Figure 5.6B, the belief-update model fit a larger number of individuals (72%), replicating previous results. Similar to previous results (**Chapter 4**), we also found no association, $r = -0.02, p = 0.68$, between social anxiety and the log likelihood difference between models, suggesting socially anxious people can be described by both associative and belief-update models. We then fit both models using the clinically informed model fitting procedure described in **Chapter 4**. Contrary to our hypothesis, both the $\lambda_{selfneg}$ learning rate $m = 0.02, CI[-0.03, 0.06]$ and the alpha-self parameter $m = -0.03, CI[-0.17, 0.10]$ had credible intervals overlapping zero, suggesting neither of these parameters had any associations with social anxiety (BFNE). We did however replicate the strong correlation between the self-negative learning rate and the trait alpha-self parameter, Spearman's Rho $= -0.75, p. < .001$, again suggesting people who learn faster from negative outcomes also have a reduced trait self-positive belief.

5.5.5 Relationships between key computational parameters

The bivariate correlations between key (transformed) parameters show correlations mainly within task and few between tasks (Figure 5.7A). The strongest correlation was the aforementioned correlation between $\lambda_{self-neg}$ and α_{self} parameters of the social evaluation learning task, $r = -0.75, p. < .001$. The two aversive learning parameters were strongly inversely correlated, suggesting greater learning from threat is paired with reduced learning from safety $r = -0.58, p. < .001$. The criterion parameter was also strongly negatively correlated with the threat drift rate $r = -0.47, p. < .001$ and positively correlated with the neutral drift rate $r = 0.42, p. < .001$ parameters, suggesting individuals who had a stronger tendency to

classify words as threatening also have faster evidence accumulation for threatening words and slower evidence accumulation to neutral words. The two drift rate parameters were also positively correlated $r = 0.33, p. < .01$, suggesting people who have faster evidence accumulation in one condition also tend to have faster evidence accumulation in the other. The two avoidance learning update parameters were positively correlated $r = 0.3, p. < .01$, suggesting people had an overall tendency to update more or less across valences. Across tasks, there was a weak, but significant negative correlation between the starting point bias and the safety update aversive learning parameter $r = -0.13, p. < .05$.

We further explored the relationships between our selected computational parameters by performing an exploratory-confirmatory factor (ECFA) analysis using 75% of the data on the exploratory FA and 25% on the confirmatory FA. The results of the exploratory FA on the training data, as assessed by Cattell's Criterion (Cattell, 1966) suggested that a 7-factor solution fit the data best. Taking 7 factors as a starting point, 7 factors were tested against a lower number of factors in the CFA on the test data, the results of which suggested that a 5-factor solution fit the data best, according to our fit indices (Figure 5.7B). Smaller factor solutions resulted in non convergence of the SEM models during CFA. In order to investigate whether there was a structure to the parameters that existed beyond the within-task factors, we further tested an ECFA model with orthogonalised variables, however this resulted in ultra-heywood cases in the SEM model. The 5-factor structure reflects the within-task correlations and the inability of the model to fit a factor structure once these within-task correlations were removed suggests that these combination of parameters are largely measuring different constructs.

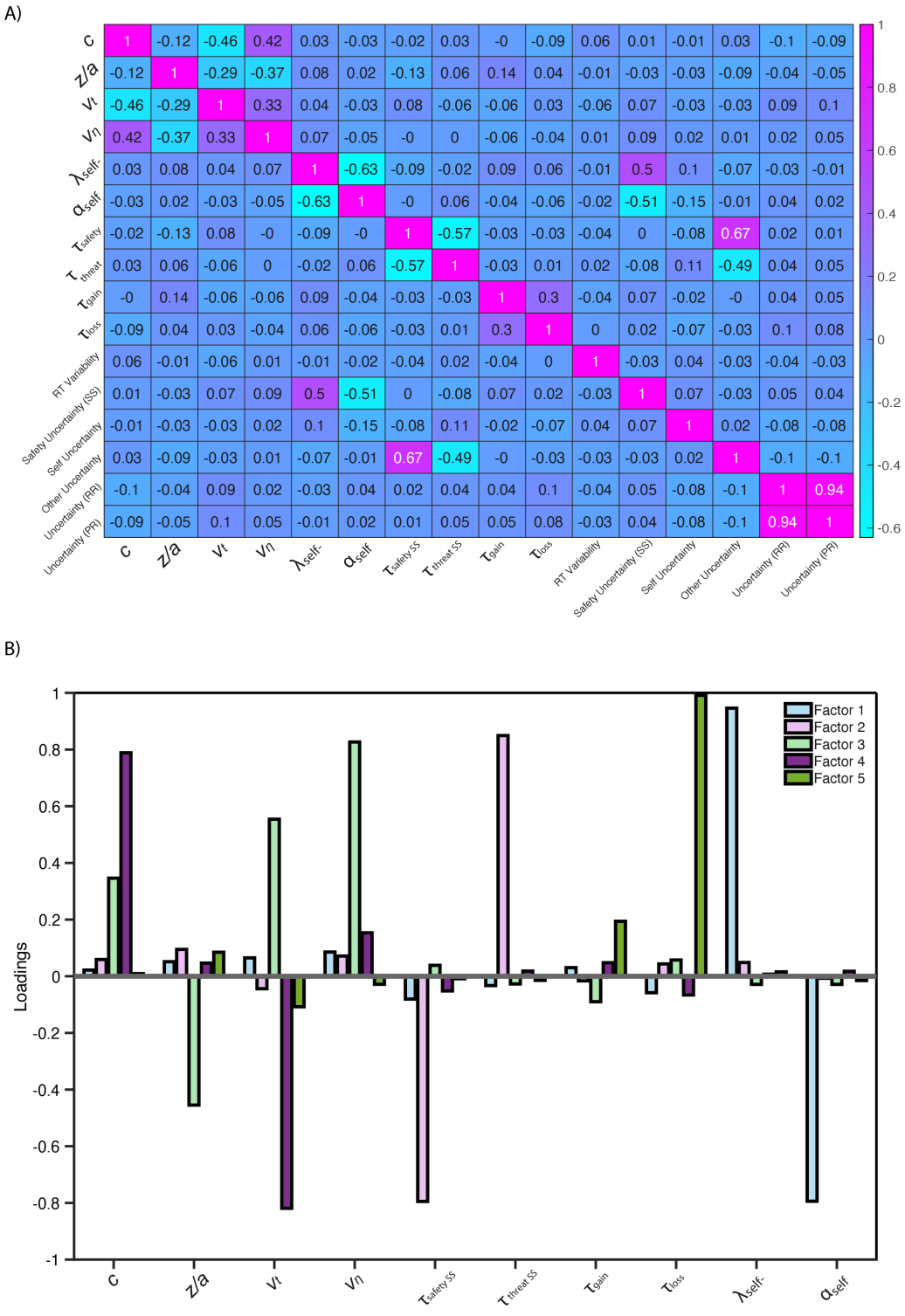

Figure 5.7: A) Bivariate correlations between parameters and uncertainty quantities. Correlations are largely between parameters from the same task. B) Parameter loadings from the 5-factor FA. Factors group together same task parameters. Threat bias task splits into two factors (3 and 4), one replicating the trait anxiety associations and the other representing an overall faster evidence accumulation and lower starting bias.

5.5.6 Key parameter-anxiety relationships

We first examined the relationship between the key model parameters outlined in our preregistration document and trait anxiety. As we included two similar measures of trait anxiety, the STAI and the STICSA, we computed the z-score average of the two measures to produce a trait anxiety metric for the subsequent analyses. We performed Bayesian regression analyses on the key parameters with trait anxiety as a predictor variable and age and gender as covariate demographic variables. Figure 5.9A displays the key relationships. As expected, high trait anxiety was associated with a lower criterion parameter, a greater drift rate to threat, a greater self-negative learning rate ($\lambda_{self-neg}$) and a lower trait positivity (α_{self}) parameter. Contrary to our expectations, there was no association of trait anxiety with the starting bias parameter, drift rate to neutral words, threat and safety learning, and either of the τ parameters from the Avoidance task. We then performed a similar analysis on the uncertainty parameters pre-selected in our preregistration document, with trait anxiety as a predictor variable and age and gender as covariates. Contrary to our hypothesis, none of the uncertainty parameters were associated with trait anxiety.

5.5.7 Symptom-parameter associations

The results of the FA on anxiety and mood symptomatolgy showed evidence for three latent factors in our data, termed 'Cognitive Anxious-Depression', 'Social Anxiety' and 'Somatic Anxiety' dimensions. The factors were named on the basis of the key subscale constructs which had strong loadings for each factor (Table 5.5).

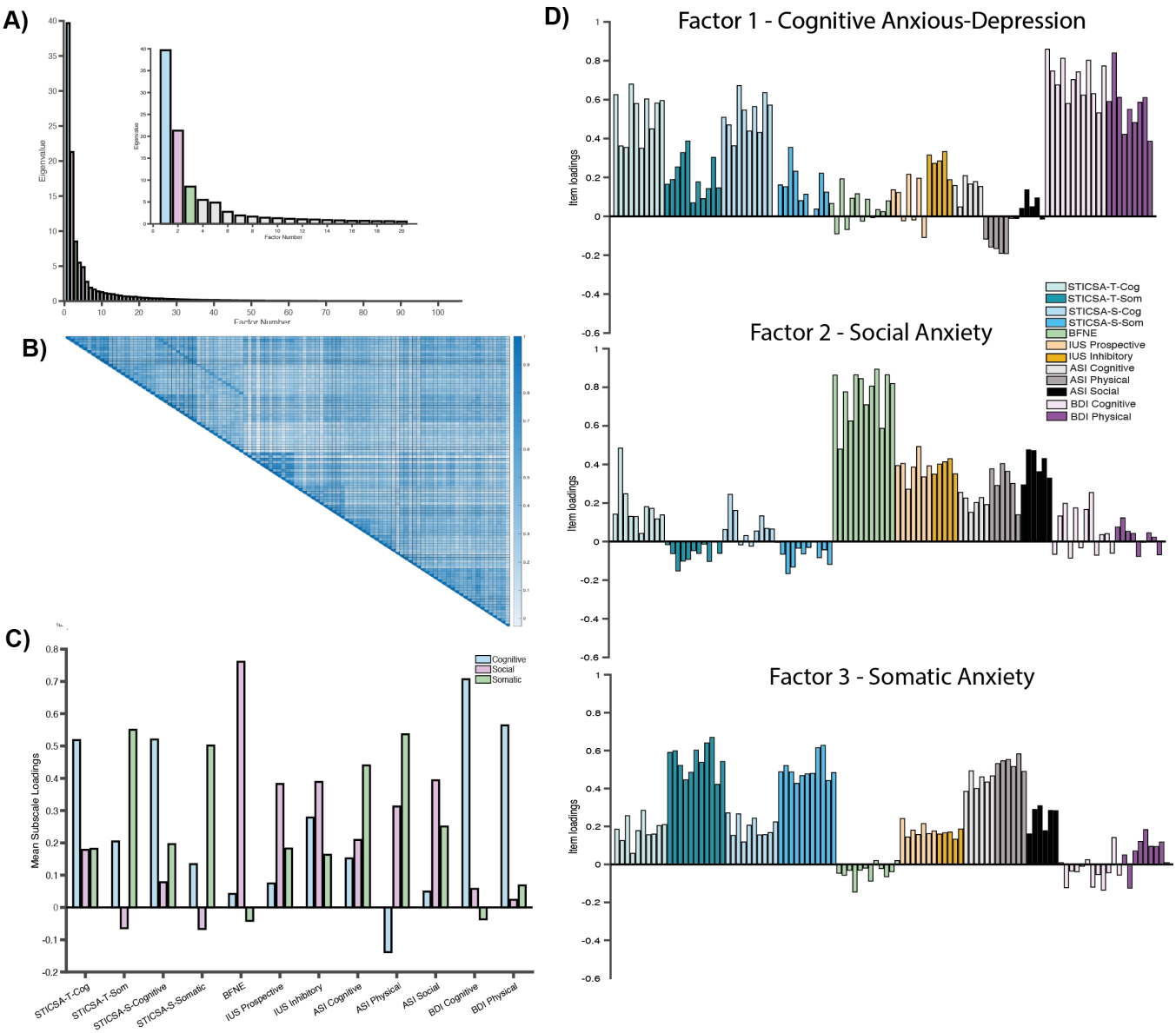

Figure 5.8: A) Eigenvalues for the first 100 and (inset) 20 factors, with three eigenvalues selected according to Cattells criterion highlighted in blue, pink and green. B) Correlation matrix of the questionnaires. C) Mean factor loadings for each subscale. D) Individual item loadings (n = 155) for each of the three factors and coloured according to their associated subscale.

Demonstrated in Figure 5.8, and in Table 5.5, the cognitive anxious-depression (cognitive) factor consisted predominately of the STICSA cognitive subscales, the intolerance of uncertainty inhibitory subscale and both of the depression (BDI) subscales representing cognitive and physical depressive symptoms. The social factor consisted predominantely of social anxiety (BFNE), intolerance of uncertainty (both of the IUS subscales) and anxiety sensitivity (ASI physical and social subscales). Finally, the somatic factor consisted predominately of the STICSA trait and state somatic subscales and all of the anxiety sensitivity (ASI) subscales. Although we determined that these factor names represent meaningful distinctions in the symptom dimensions, we acknowledge that there is overlap between some of the dimensions, with physical subscales present in the cognitive and social dimensions. The factor names were partially guided by research that has supported the computational distinction between cognitive and somatic anxiety (Wise and Dolan, 2020) and research that has investigated generalised anxiety and social anxiety seperately (Khdour et al., 2016).

	Cognitive	**Social**	**Somatic**
	(Factor 1)	(Factor 2)	(Factor 3)
STICSA-T-Cognitive	**0.52 (0.13)**	0.18 (0.12)	0.18 (0.06)
STICSA-T-Somatic	0.21 (0.10)	-0.06 (0.05)	**0.55 (0.08)**
STICSA-S-Cognitive	**0.52 (0.10)**	0.08 (0.08)	0.20 (0.05)
STICSA-S-Somatic	0.13 (0.11)	-0.07 (0.05)	**0.50 (0.06)**
BFNE	0.05 (0.08)	**0.76 (0.13)**	-0.04 (0.05)
IUS Prospective	0.07 (0.12)	**0.38 (0.07)**	0.18 (0.03)
IUS Inhibitory	**0.28 (0.06)**	**0.39 (0.04)**	0.16 (0.20)
ASI Cognitive	0.15 (0.05)	0.21 (0.04)	**0.44 (0.04)**
ASI Physical	-0.14 (0.07)	**0.31 (0.10)**	**0.54 (0.03)**
ASI Social	0.05 (0.06)	**0.40 (0.08)**	**0.25 (0.06)**
BDI Cognitive	**0.71 (0.10)**	0.06 (0.12)	-0.04 (0.08)
BDI Physical	**0.56 (0.13)**	0.02 (0.07)	0.07 (0.09)

Table 5.5: Mean (SD) symptom factor analysis loadings for each of the three factors. Indicated in bold are factor loadings over 0.25 that suggest key constructs for each factor.

	C	z/a	Vt	Vn	τ_{safety}	τ_{threat}	τ_{gain}	τ_{loss}	λ_{self-}	α_{self}
Independent Models										
Cognitive Anxious-Depression (Factor 1)	-0.12 [0.22 -0.03]	0.004 [-0.09 0.10]	0.11 [0.01 0.20]	-0.02 [-0.11 0.08]	-0.002 [-0.10 0.10]	-0.06 [-0.15 0.04]	-0.03 [-0.13 0.06]	-0.07 [-0.16 0.02]	0.14 [0.05 0.23]	-0.19 [-0.21 0.02]
Social Anxiety (Factor 2)	-0.12 [-0.22 -0.04]	-0.03 [-0.13 0.06]	0.17 [0.08 0.27]	0.05 [-0.04 0.15]	-0.04 [-0.14 0.06]	-0.05 [-0.14 0.05]	-0.06 [-0.16 0.03]	-0.07 [-0.17 0.03]	0.13 [0.03 0.22]	-0.09 [-0.19 0.002]
Somatic Anxiety (Factor 3)	-0.12 [-0.24 -0.06]	-0.02 [-0.12 0.07]	0.09 [-0.003 0.18]	-0.03 [-0.13 0.06]	0.03 [-0.07 0.12]	-0.11 [-0.21 -0.02]	-0.01 [-0.11 0.08]	-0.04 [-0.13 0.06]	0.01 [-0.09 0.10]	-0.001 [-0.10 0.09]
Covariate Model										
Cognitive Anxious-Depression (Factor 1)	-0.01 [-0.14 0.12]	0.05 [-0.08 0.19]	-0.002 [-0.13 0.13]	-0.04 [-0.17 0.09]	0.001 [-0.14 0.14]	0.02 [-0.11 0.16]	-0.004 [-0.14 0.13]	-0.05 [-0.18 0.08]	0.17 [0.04 0.30]	-0.16 [-0.29 -0.02]
Social Anxiety (Factor 2)	-0.08 [-0.19 0.03]	-0.05 [-0.17 0.06]	0.16 [0.05 0.28]	0.09 [-0.02 0.21]	-0.06 [-0.17 0.06]	-0.01 [-0.13 0.10]	-0.07 [-0.18 0.05]	-0.05 [-0.16 0.07]	0.08 [-0.04 0.19]	-0.05 [-0.16 0.07]
Somatic Anxiety (Factor 3)	-0.11 [-0.23 0.001]	-0.03 [-0.15 0.08]	0.03 [-0.09 0.14]	-0.05 [-0.16 0.07]	0.05 [-0.08 0.17]	-0.12 [-0.24 -0.0001]	0.01 [-0.10 0.13]	0.01 [-0.11 0.13]	-0.12 [-0.24 -0.001]	0.11 [-0.008 0.23]

Table 5.6: Bayesian regression model results given as mean [Lower CI Upper CI] for both independent models, with predictors entered separately and covariate models with predictors entered into the same model.

The results of the Bayesian regression with independent predictors (Figure 5.9A and Table 5.6) showed dissociations between symptom dimensions for some of our key parameters. Namely, the threat drift rate, Vt was associated with cognitive and social anxiety but not somatic. The aversive threat learning τ^{threat} parameter was associated with somatic anxiety but not other dimensions and the social evaluation parameters were associated with cognitive and social but not somatic dimensions. The only parameter to show significant associations across the three factors was the criterion, C parameter, suggesting cognitive, social and somatic anxiety all share the increased tendency to classify non-threatening words as threatening. When removing the shared variance between the factors in a model that included all three dimensions together, the criterion parameter was now associated only with the somatic dimension. Interestingly, the somatic dimension was also associated with a lower self-negative learning rate for social evaluation, and a reduced threat update parameter, suggesting a more positive bias overall amongst negative outcome learning parameters. The only parameter that was associated with the social dimension was the threat drift rate. Interestingly, the parameters of the social evaluation task were not associated with the social dimension when accounting for shared variance but our cognitive dimension had higher self-negative learning rates as well as lower trait self-positivity.

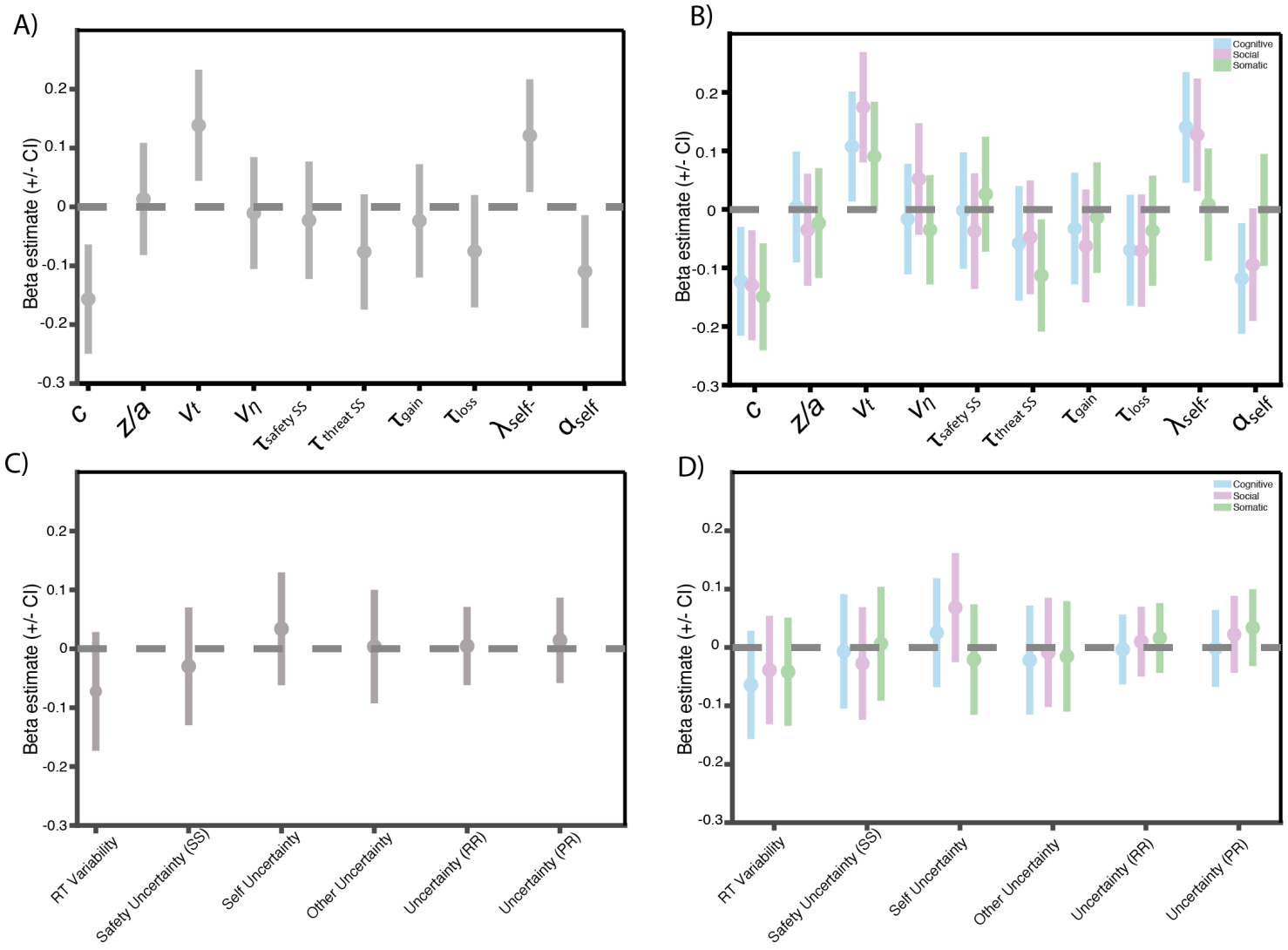

Figure 5.9: A) Results of the Bayesian regression models for key model parameters. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval. B) Bayesian Regression Models for each key parameter with dimension predictors entered separately into each model, with age and gender as covariates. C) Bayesian Regression Models for uncertainty and model derived quantities with dimension predictors entered separately into each model, with age and gender as covariates D) Bayesian Regression Models for uncertainty and model derived quantities with dimension predictors entered together into the same model, with age and gender as covariates.

Contrary to our hypothesis that overestimation of uncertainty would relate to higher anxiety, the results of the Bayesian regression with the uncertainty and model-derived quantities with independent predictors (Figure 5.9C and D) showed no associations between uncertainty and any of the dimensions. When removing the shared variance between the factors in a model that included all three dimensions together, there were no associations amongst any of the uncertainty parameters and symptom dimensions.

5.5.8 PLS on parameter-symptoms

The results of the PLS analysis suggest one component fit the data best. An examination of the questionnaire item loadings displayed in Figure 5.10 shows that almost all items load positively with this component, suggesting it can be thought of as a 'general distress' component, akin to the p-factor (Caspi et al., 2014). The highest loading items are from a range of different questionnaires, but feature a number of symptoms from the somatic subscales of the questionnaires, suggesting somatic symptoms are a prominent feature of the general distress factor. Examination of the parameter loadings displayed in Figure 5.10, reinforces and extends the results presented in the Bayesian regression analyses, with the criterion, danger learning and trait-self parameters loading negatively and the threat drift rate, and self-negative learning rates loading positively. This suggests that people high in general distress have a greater threat bias, reduced learning from danger, as well as a negative self-evaluation bias. The other parameter loadings, as indicated by 95% CIs, were not significantly different from zero (Figure 5.10). The significant parameters in the PLS are the same parameters that have a relationship to trait anxiety, however the questionnaire items span a range of symptomatolgy, suggesting these parameters may index processes common to a range of mental health assays, such as anxiety, low mood and intolerance of uncertainty.

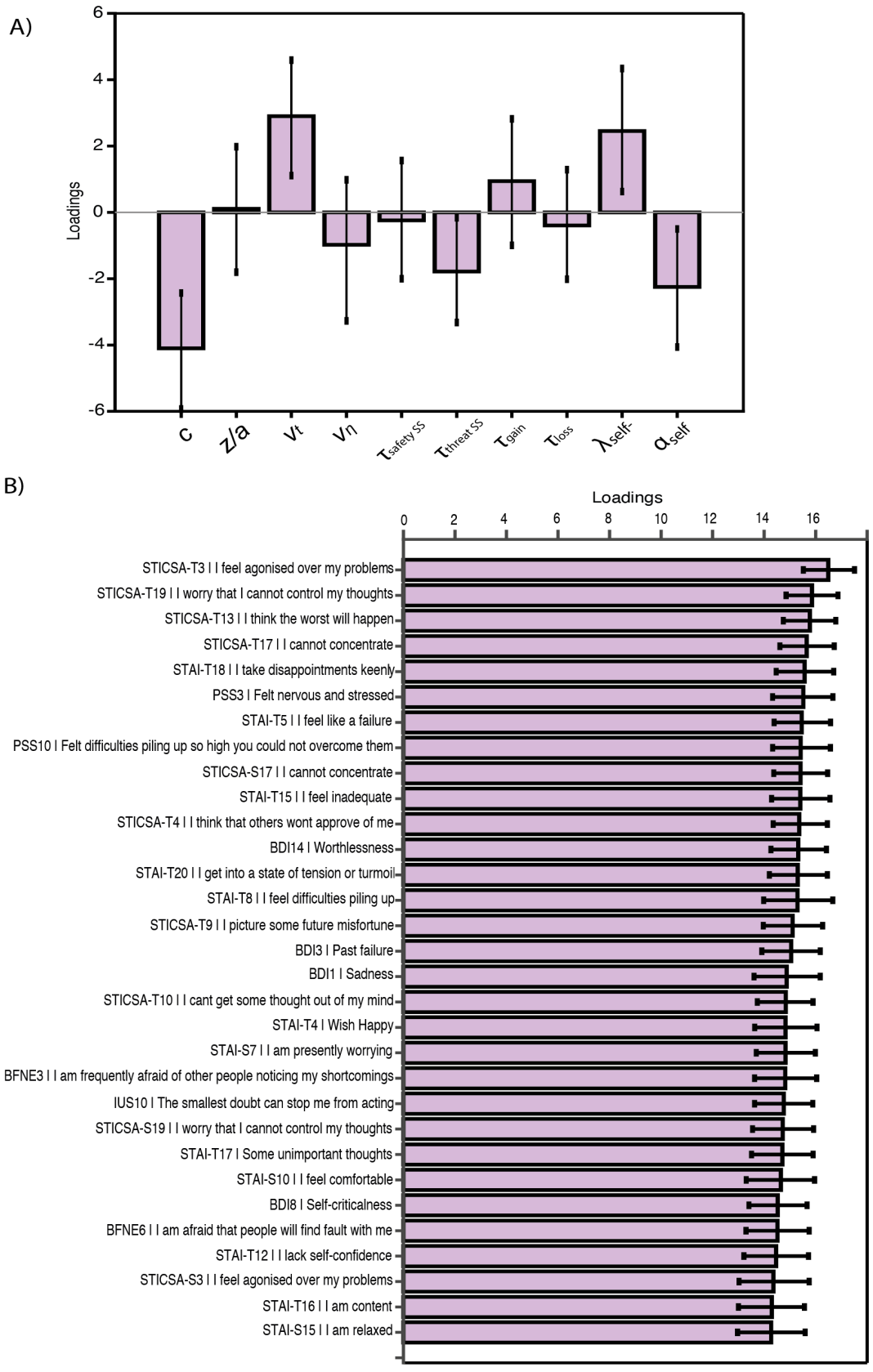

Figure 5.10: A) One component PLS regression loadings for the key parameters from the bootstrapped samples. Error bars are bootstrap confidence intervals. B) One component PLS regression loadings for the top 80% of items. Error bars are bootstrap confidence intervals

5.6 Discussion

This study aimed to determine the relationships between four computational processes theorised to be important for anxiety (Raymond et al., 2017), namely: threat bias, negative self-bias, aversive learning, and avoidance. Using computational modelling to derive previously identified (Button et al., 2015; White et al., 2016; Wise and Dolan, 2020) and novel anxiety-related computational parameters, including avoidance and uncertainty estimation, we examined how these parameters related to each other and to different parts of the spectrum of anxiety symptomatology. Overall, our results showed that trait anxiety and a data-driven 'general distress' factor were associated with a high bias towards threat in a non-learning context, a greater negative self-bias during social evaluation, and reduced learning from threat in aversive learning contexts. Symptom dimension analysis revealed distinctions between anxiety subtypes, with cognitive and social dimensions showing high bias towards threat and a negative self-evaluation bias, whereas somatic anxiety showed only a weaker criterion alongside reduced learning from threat. Contrary to our expectations, none of the uncertainty parameters were associated with trait anxiety or any of our anxiety dimensions. We discuss the results in turn.

The first part of the analysis focused on the replication of the key results for each of the anxiety tasks. The strongest results overall were found within the threat bias task, with BF scores suggesting strong evidence for our hypothesis that people with high trait anxiety will display a weaker criterion and larger threat drift rate. However, we found no evidence for the previously reported (White et al., 2016) relationship between trait anxiety and reduced drift rates to neutral words, or a starting bias towards threat. We partially replicated the results of the aversive learning task (Wise and Dolan, 2020) by also finding an association between trait

anxiety and reduced threat learning, however we could not replicate the distinction between cognitive and somatic subtypes when using the STICSA subscales. Looking at the distributions across subscales suggests that our sample had a greater proportion of people scoring highly on the cognitive subscales and somatic subscales had a positive skew, suggesting there might not have been enough variance on the somatic scales to fully distinguish these effects. However, with the increased sensitivity of the factor analysis on our questionnaire data, we were able to pinpoint the association to within the somatic anxiety dimension. We were not able to replicate any associations with the safety learning parameter. The exact reason for the counterintuitive reduced learning from threat in anxiety is unclear, but has been observed in previously published work (Wise et al., 2019; Wise and Dolan, 2020) and could be a result of a reluctance to leave areas previously found to be safe (Wise and Dolan, 2020), however a direct test of this hypothesis needs to be performed.

Perhaps the biggest differences between the results presented here and previous results (Button et al., 2015) (**Chapter 4**) were observed in the social evaluation task, with our sample showing no relationship between social anxiety when assessed using BFNE and negative social evaluation style, but a significant association with trait anxiety and our 'general distress' component. However, when assessed using our factor analysis symptom subtypes, there was a significant association to our key parameters and the cognitive anxious-depression dimension as well as our social anxiety dimension, suggesting the analysis benefited from the inclusion of the intolerance of uncertainty scales. The replication of results from **Chapter 4** were also extended in our sample, as our exploratory analyses revealed high trait anxiety was associated with both a higher self-negative learning rate, as previously observed and a lower self-positive learning rate, as well as a lower initial bias. Furthermore, our belief-update

model showed the previously observed reduced trait-self positive parameter and additionally a greater trait-self negative parameter. Taken together, these results may suggest that high anxiety is related to changes in both positive and negative learning, as well as a starting bias towards negativity. It should be noted that the two samples were enriched according to different criteria, with the original sample reported in Button et al. (2015) selected for a third of the sample to be 'high', 'medium' and 'low' social anxiety scores respectively, whereas our sample was specifically enriched with the top 80% of trait anxiety scores. Although our sample was not targeting recruitment at high social anxiety participants, there was a large proportion of our sample that scored above the indicator (Carleton et al., 2011) of possible clinical interest and the distribution was negatively skewed, suggesting our sample had an over-representation of high social anxiety participants. This over-representation of highly socially anxious participants has been observed in the use of online platforms for recruitment (Arditte et al., 2016), but it is unclear how this may influence behaviour. It is conceivable that the high presence of socially anxious individuals is a result of being in an environment in which they feel comfortable, with no need for social interaction with other people, rather that the lab environment in which social interaction is necessary.

One of the aims of this experiment was to create a novel task to assess how early negative experience impacts learning and avoidance. The task was able to behaviourally assay this concept, evidenced by the greater avoidance observed in environments with early negative outcomes, however, there was no hypothesised relationship to anxiety. Greater avoidance has been well documented in anxiety both clinically, behaviourally and computationally (Mkrtchian et al., 2017). In our task, participants had to choose between foraging and therefore reducing uncertainty about the distribution of outcomes and choosing to stay home,

which results in the avoidance of potential negative outcomes but also does not reduce uncertainty. Therefore, it could be that the trade-off between these two concepts, both thought to be related to anxiety, may obscure our ability to measure avoidance. This explanation is also supported by the relationship between individuals who are highly intolerant of uncertainty and greater choice stickiness for the foraging option, meaning they tended to repeat choices that reduce uncertainty. However, our model derived measures of uncertainty were unrelated to any anxiety symptomatology, therefore the exact explanation for our pattern of results is unclear. In our model formulation, the value for each environment is informed only by the mean of the beta distribution parameters. Model formulations that directly incorporate uncertainty into the value update may be informative for investigating these effects.

Using factor analysis, we were able to decompose our anxiety measures into three symptom domains that we termed cognitive anxious depression, social anxiety and somatic anxiety. The use of symptom dimensions to assess specific contributions of subtypes has been used to show specific contributions of compulsivity to model-based learning (Gillan et al., 2016). Here, we were able to show computational distinctions between these subtypes. Complementary to these results, the PLS between our key computational parameters and psychopathology indicated that one component was the best explanation for our data. We termed this component a 'general distress' component, as it loaded positively on the range of our questionnaires assessing different forms of anxiety, stress and low mood. This component therefore cuts across the whole of our anxiety/mood symptom spectrum and suggests that even though the symptom range can be decomposed into three dimensions, broadly representing cognitive, social and somatic anxiety, the shared variance between these dimensions

more strongly contributes to the computational parameter profile.

Interestingly, examining the correlations between our key parameters without anxiety symptoms suggested that parameters were largely uncorrelated across different tasks, but showed correlations within task. The fact that there were few correlations between our key parameters suggests that they are capturing largely distinct processes. Even between the learning tasks, there were few correlations, suggesting that social learning, aversive and avoidance learning are assaying distinct computational mechanisms. This was further supported by our factor analysis over parameters, that indicated a 5-factor solution to our data, with each factor representing groupings of parameters from within the same task. Recent research (Moutoussis et al., 2021) has highlighted a 'decision-acuity' factor, which suggests people who are good at decision-making on one task, will be good at decision-making in others. We do not observe such a general task factor here, however fewer tasks were included in our task battery.

One of the main motivations for this study was the investigation of the relationship between anxiety and different measures of uncertainty. Surprisingly, we found no evidence of a relationship between trait anxiety and any of our uncertainty parameters, nor did we observe any relationship using our anxiety symptom dimensions. Overestimation of and difficulties in learning under uncertainty has been previously documented (Wise and Dolan, 2020; Browning et al., 2015) and has been theorised to be an important mechanism for understanding anxiety (Raymond et al., 2017; Bishop and Gagne, 2018; Bach and Dolan, 2012), therefore it is surprising that this relationship was not observed here across different measures. However, our lack of uncertainty relationship here only suggests that there is no difference in uncertainty at the level of belief formation. To further understand whether

uncertainty may be involved in the action selection process itself, models that explicitly incorporate uncertainty into the value computation should be applied, such as those briefly introduced in **Chapter 4**. Importantly, across all of our learning tasks, the uncertainty based models performed better overall, suggesting that the explicit uncertainty is a necessary component for understanding the learning process, however the benefit of uncertainty-based models in terms of model fit were not associated with anxiety, suggesting there was no difference between the learning styles (associative vs belief based) of people with higher social or trait anxiety vs the rest of the population, in line with previous results **Chapter 4**.

Limitations One limitation of this study and a direction for future research is that we modelled all of our tasks separately and therefore did not utilise the shared information that may occur between tasks within the same individuals. Hierarchical models, which incorporate information shared between tasks that involve similar processes, like learning tasks could provide important information about individuals that can only be understood by considering how different measures may carry shared information. Moreover, this study was conducted during the COVID-19 pandemic, which may have influenced participant behaviour in ways that we could not measure here. The pandemic exacerbated mental health symptoms generally (Pierce et al., 2020), however it is difficult to understand the effect of pandemic related mental health symptoms and performance on the tasks used in this study.

Conclusions This is the first study to assess the relationships between anxiety-related computational parameters and anxiety symptoms within the same individuals. We were able to derive a computational parameter profile for trait anxiety and general distress, consisting of a high threat bias, a self-negative learning style and reduced threat learning. Distinct but

overlapping computational profiles for cognitive, social and somatic anxiety may further orient research towards computational biomarkers.

5.7 Supplementary Information

5.7.1 Sample Information

The anxiety symptom distributions are displayed in Figure S5.1 along with the conventional clinical interest cut-offs for each questionnaire. The state anxiety scores were lower generally than for trait anxiety. For the STAI-Trait, a large percentage of our sample (78%) scored above the 'moderate anxiety' indicator score of 38 (Spielberger, 2010; Ercan et al., 2015). For the STAI-State, 63% scored above this indicator. For the STICSA-Trait, 39% of the sample scored over the clinical cut-off score of 43, and for the STICSA-State, 27% of the sample scored above the cut-off (Van Dam et al., 2013). A further breakdown of the STICSA subscales shows the somatic subscales were generally positively skewed, whereas the cognitive subscales have a more even distribution. There were a large percentage of social anxiety participants in our sample, with 86% of our sample scoring over a clinical cut-off score of 25 on the BFNE (Carleton et al., 2011). For fear of fear, measured by the ASI-3, 29% of our sample scored over the cut-off score of 36, which indicates 'moderate' anxiety (Taylor and Montgomery, 2007). The percentage of our sample displayed high life stress, measured by the PSS was 11% (Cohen et al., 1983). Finally, around a quarter (23%) of our sample displayed symptoms indicative of severe clinical depression, scoring above 30 on the BDI-II (A. T. Beck et al., 1996). Overall, our sample had high proportions of trait anxiety (as assessed by the STAI, which was specifically targeted during recruitment) and social anxiety.

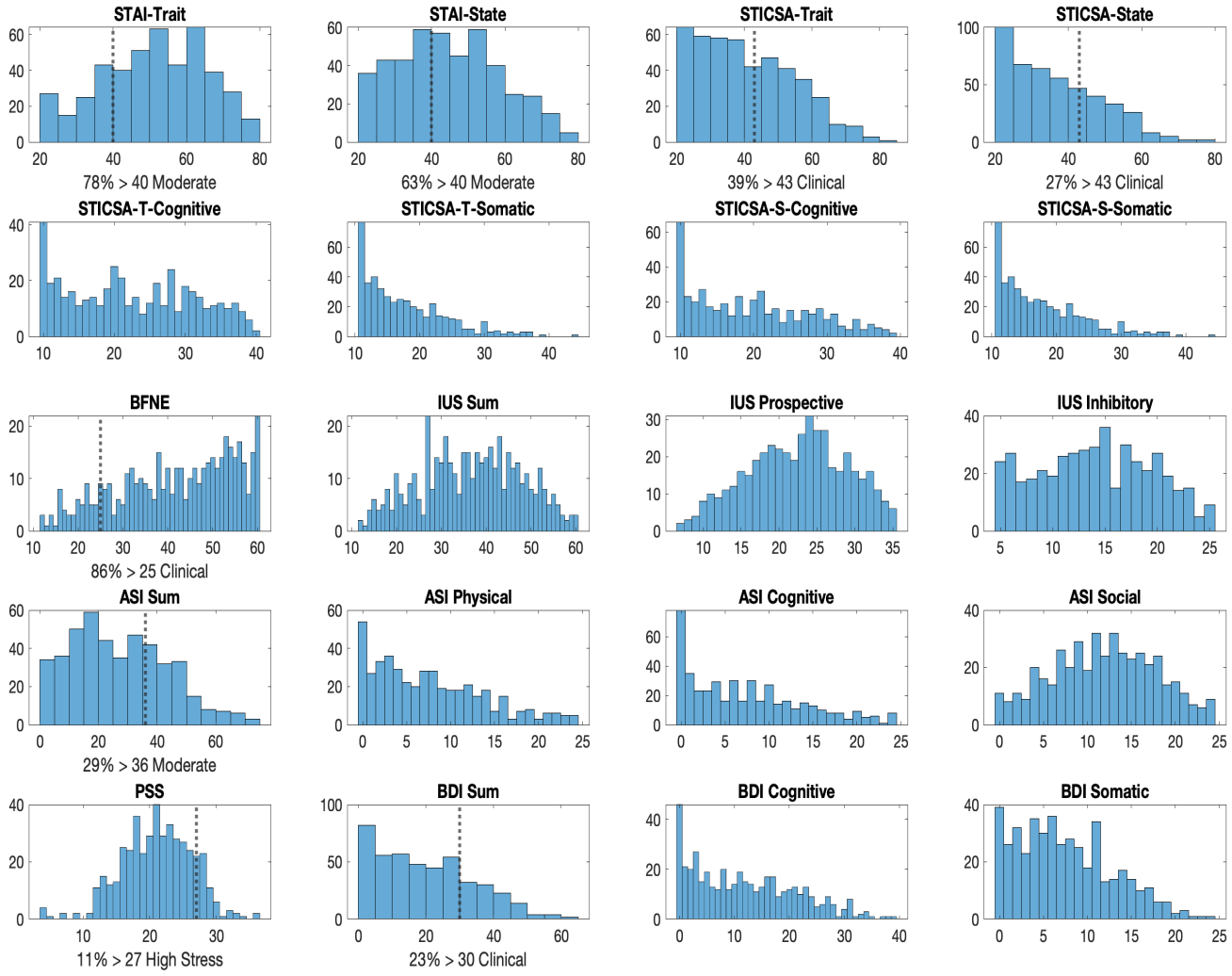

Figure S5.1: Histograms of the anxiety symptomatology in the sample for both questionnaire sum scores and questionnaire subscales. The vertical lines represent the questionnaire guidelines for a clinical interest cut-off and the percentage of the sample that exceeds cut-off is given along the x-axis.

5.7.2 Exploratory parameter relationships

We held no hypotheses about the other model parameters that were not pre-selected, but the exploratory analysis which investigates the relationship between trait anxiety (measured by the z-score average of the STAI and STICSA) and all model parameters is presented for completeness. The exploratory analysis revealed that most parameters had no relationship to trait anxiety. For the threat bias task, we again observe a lower criterion and greater threat

drift rate, but additionally a greater d' parameter. This suggests people with high trait anxiety show a greater ability to discriminate between threat and neutral words. Using this measure of trait anxiety, we no longer see the reduced threat update (τ_{threat}) parameter for the aversive learning task, suggesting those results were more specific to the STICSA, which is thought to be more reflective of anxiety and less so of depression compared to the STAI (Grös et al., 2007). We again observe a smaller stickiness parameter (S_{forage}) for the avoidance task, suggesting highly trait anxious individuals tend to repeat their exploratory choices. For the social evaluation task, we replicate previously reported relationships between the self-negative learning rate (λ_{self-}) and the trait self-positivity parameters (α_{self} **Chapter 4**), but additionally see a reduced self-positive learning rate (λ_{self+}, AL model) and a greater trait-negative parameter (β_{self}, BU model). This extends the findings observed in **Chapter 4**, as it suggests that both positive and negative processing is altered to form a negative self-bias. Moreover, there was also a lower initial bias towards positivity, suggesting anxious individuals start off with negative expectations.

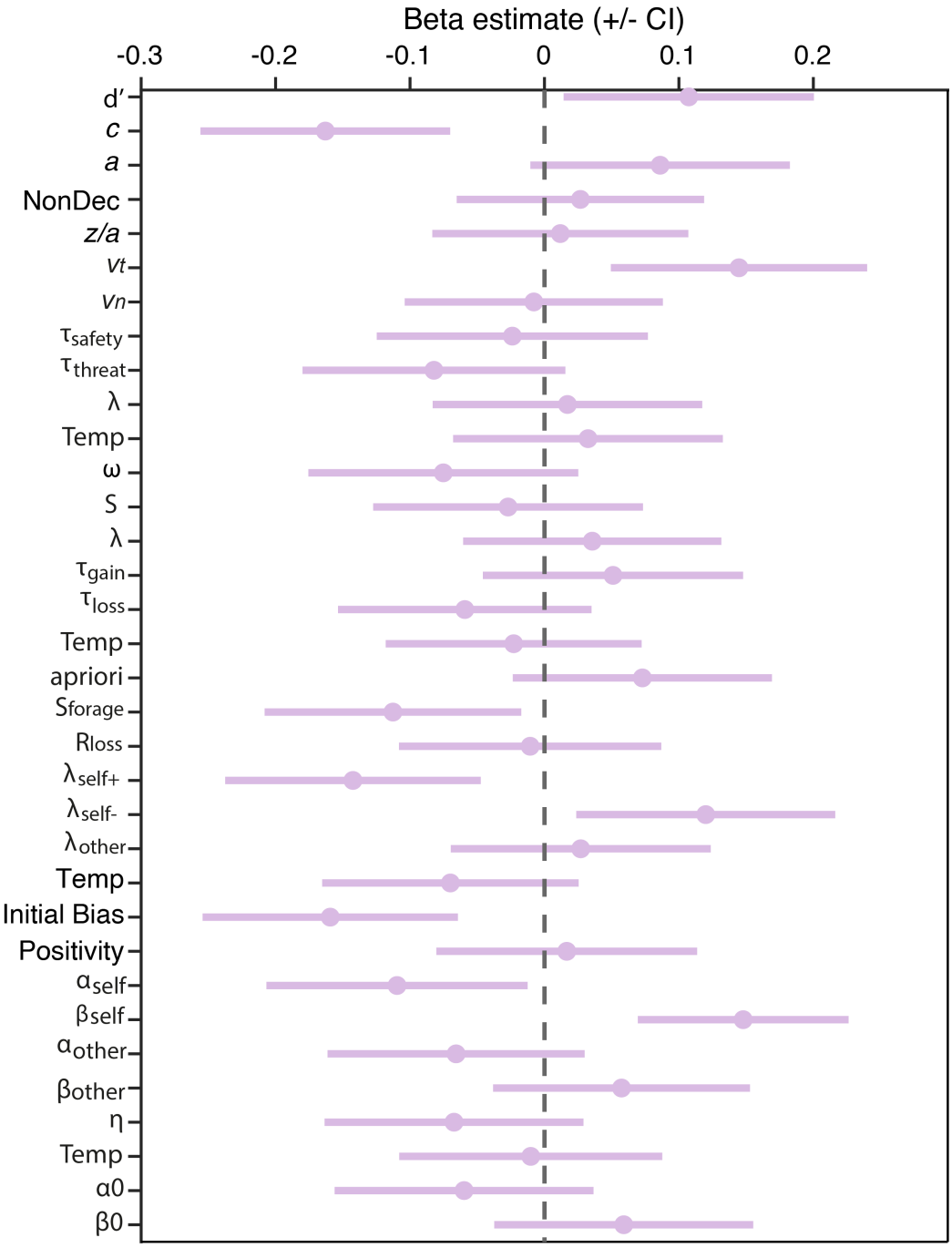

Figure S5.2: Results of the Bayesian regression models for all model parameters using our trait anxiety measures as predictor variable. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

Demographic Analyses

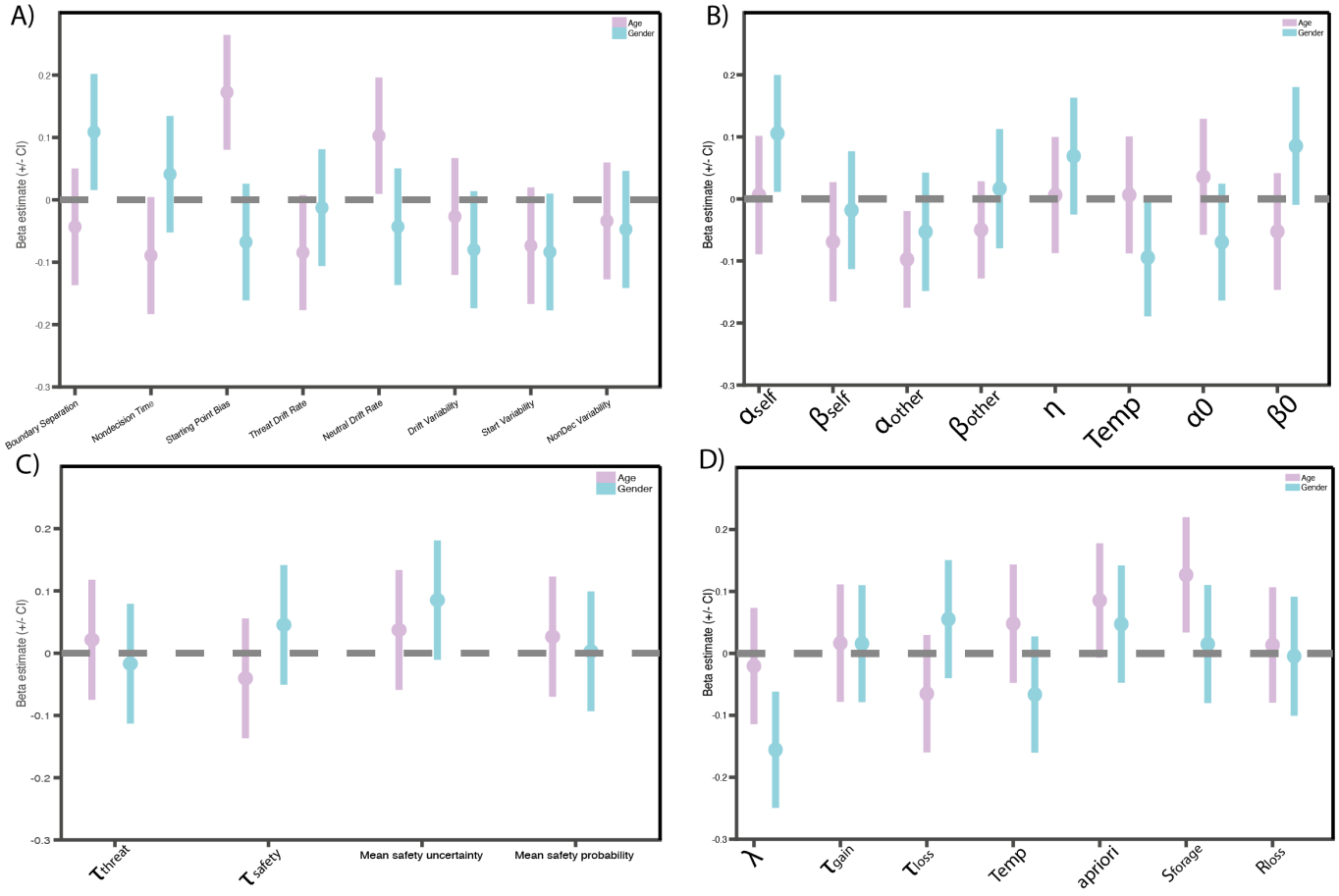

Figure S5.3: Results of the Bayesian regression of gender and age as predictors for each of the model parameters for A) the threat bias task, B) the social evaluation task C) the aversive learning task D) the avoidance task. Markers indicate the mean of the posterior distribution with error bars representing the 95% equitailed credible interval.

We performed an exploratory analysis of age and gender demographics on model parameters. For the threat bias task, there was an effect of age on the non-decision time, starting point bias (z/a) and neutral drift rate parameters (Vn), suggesting older individuals made decisions faster, had a greater starting bias towards neutral words, as well as faster evidence

towards neutral words, suggestive of a reduced threat bias overall. This is in line with a positivity bias observed in older adults generally (Carstensen and DeLiema, 2018). Women had a greater boundary separation, meaning they tended to respond more slowly, but also more accurately, similar to sex differences in speed-accuracy trade-offs in other domains (von Kluge, 1992). In the social evaluation task, older individuals had a reduced other-positivity parameter (α_{other}), suggesting their trait other beliefs are more negatively skewed. Women had a larger trait positivity parameter, (α_{self}) suggesting they have more positive self-views, which is surprising given women generally tend to have lower self-esteem and less positive self evaluations of themselves (Allen et al., 2001). Finally, for the avoidance task, older individuals had a greater stickiness parameter (S_{forage}), indicating they tended not to repeat foraging decisions, but also had a greater a priori tendency towards foraging decisions. Women had a lower decay parameter (λ), meaning they tended towards forgetting older information and weighted recent information more heavily.

5.7.3 Full modelling descriptions

Signal Detection Theory d' is a measure of the ability to discriminate between threat and neutral words. Larger d' indicates a greater ability to discriminate threat from neutral. d' is calculated as:

$$d' = z(hitrate) - z(false - alarmrate) \tag{51}$$

Criterion, c, reflects the decision criterion and is a measure of bias. Weaker c indicates a lower threshold for classifying words as threatening. Criterion is calculated as:

$$c = -0.5 * (z(hitrate) + z(false - alarmrate)) \tag{52}$$

Drift Diffusion Model There were 8 parameters in the DDM, consisting of 5 key parameters that describe different components of the evidence accumulation process and 3 that capture variability in some of these components. a = boundary separation is a measure of response caution and the speed/accuracy trade-off, with a larger boundary separation leading to slower but more accurate responses. Ter = non-decision time is a measure of the processes that are considered external to the decision process itself such as motor execution. A larger non-decision time parameter suggests slower motor processes. z/a = starting point bias is a measure of the start point relative to the decision boundaries for threat and neutral words. If the start point is biased towards either decision boundary, this suggests an initial prepotent tendency towards threat or neutral classifications. Here, greater values for the starting bias parameter reflect a greater tendency towards neutral classification. The two drift rate parameters, v_t = drift rate for threatening words and v_n = drift rate for neutral words capture evidence accumulation for threat and neutral words respectively. Larger drift rate parameters suggest a bias towards threat or neutral words. The variability parameters, s_t = variability in non-decision time, s_z = variability in starting point, eta = variability in drift rate capture the trial-by-trial variance in the primary components.

Avoidance learning modelling

Associative learning models. In addition to the leaky beta models, we fit AL models, which do not directly provide a quantity of uncertainty, but capture the learning process in a model-

	Model	NP	α	q_0	Temp	S	Rloss	LL	AIC	BIC	Pseudo r^2
1	Simple	4	1	1	1		1	-274440.60	552449.21	561529.91	0.26
2	Valence	5	2	1	1		1	-272350.84	549161.67	560512.54	0.27
3	Valence + sticky	6	2	1	1	1	1	-253799.48	512950.96	**526572.00**	**0.32**
4	Valence + 2sticky	7	2	1	1	2	1	**-253328.86**	**512901.72**	528792.94	**0.32**

Table S5.1: Associative Learning models parameter descriptions and fit statistics. Best model fit statistics highlighted in bold.

	Model	NP	λ	τ	Temp	apriori	S	Rloss	LL	AIC	BIC	Pseudo r^2
1	Simple	4	1	1	1	1			-331330.65	666229.31	675310.01	0.11
2	Valence	5	1	2	1	1			-280714.13	565888.26	577239.13	0.24
3	Valence + sticky	6	1	2	1	1	1		-264852.29	535056.59	548677.63	0.29
4	Valence + 2sticky	7	1	2	1	1	2		-264863.98	535971.97	551863.19	0.29
5	Simple	5	1	1	1	1		1	-284107.08	572674.16	584025.03	0.23
6	Valence	6	1	2	1	1		1	-261023.86	527399.72	541020.77	0.30
7	Valence + sticky	7	1	2	1	1	1	1	-247845.36	**501934.72**	**517825.94**	**0.33**
8	Valence + 2sticky	8	1	2	1	1	2	1	**-247568.30**	502272.61	520434.00	**0.33**

Table S5.2: Leaky Beta models parameter descriptions and fit statistics. Best model fit statistics highlighted in bold. The final model selected was model number 7.

free manner. In this framework, participants learn the action value of the two choice options, forage vs avoid, based on the outcomes they observed. These action values are updated trial-by-trial according to the PE, which is calculated as the difference between the expected and observed outcomes, multiplied by a learning rate, λ, which provides a weight onto value update, such that larger learning rates give larger weight to recent evidence. We allowed the α values to be separate for gains vs losses (α_{gain}, α_{loss}) to account for asymmetries in learning from positive vs negative outcomes. Similar to AL models in previous chapters, models could also allow the first trial to vary as a free initial bias parameter ($-1 < q_0 < 1$), which measured the pre-potent response to forage or avoid, with larger values reflecting bias towards forage. We further incorporated a choice stickiness parameter, S, which was applied to the forage and/or avoid option (S_{forage}, S_{avoid} respectively) and increased the propensity to repeat the last choice. The Q values were then fed into a standard softmax function, described in the main text. Details of AL models are given in Table S5.1.

Posterior Parameter Correlations Figure S5.4 shows the posterior correlations of the best-fitting models and their parameters. High correlations suggest the parameters trade-off against each other, suggesting redundancy which impacts interpretation of parameter contributions. The threat bias task had a number of moderate to high correlations between parameters, suggesting improvements could be made to the model formulation and/or estimation procedure. The model was fit using the fast-DM package (Voss and Voss, 2007) which does not allow bounds to be applied to parameters, which may aid in parameter estimation (White et al., 2016). Hierarchical model fitting may also improve parameter estimates (Wiecki et al., 2013). The social evaluation BU model was largely uncorrelated except for the initial bias parameters (α_0, β_0), which were almost perfectly correlated, suggesting these parameters could be represented by a single initial bias parameter. For the avoidance learning task, the two update (τ) parameters were anti-correlated, which was previously reported in Wise and Dolan (2020). For the avoidance task, the update parameter for gains was anti-correlated with the (inverse) temperature parameter suggesting people who forage more show increased decision variability. This fits the intuition that people who forage more might be less sure of their decisions and thus might sample more information throughout the task, increasing variability.

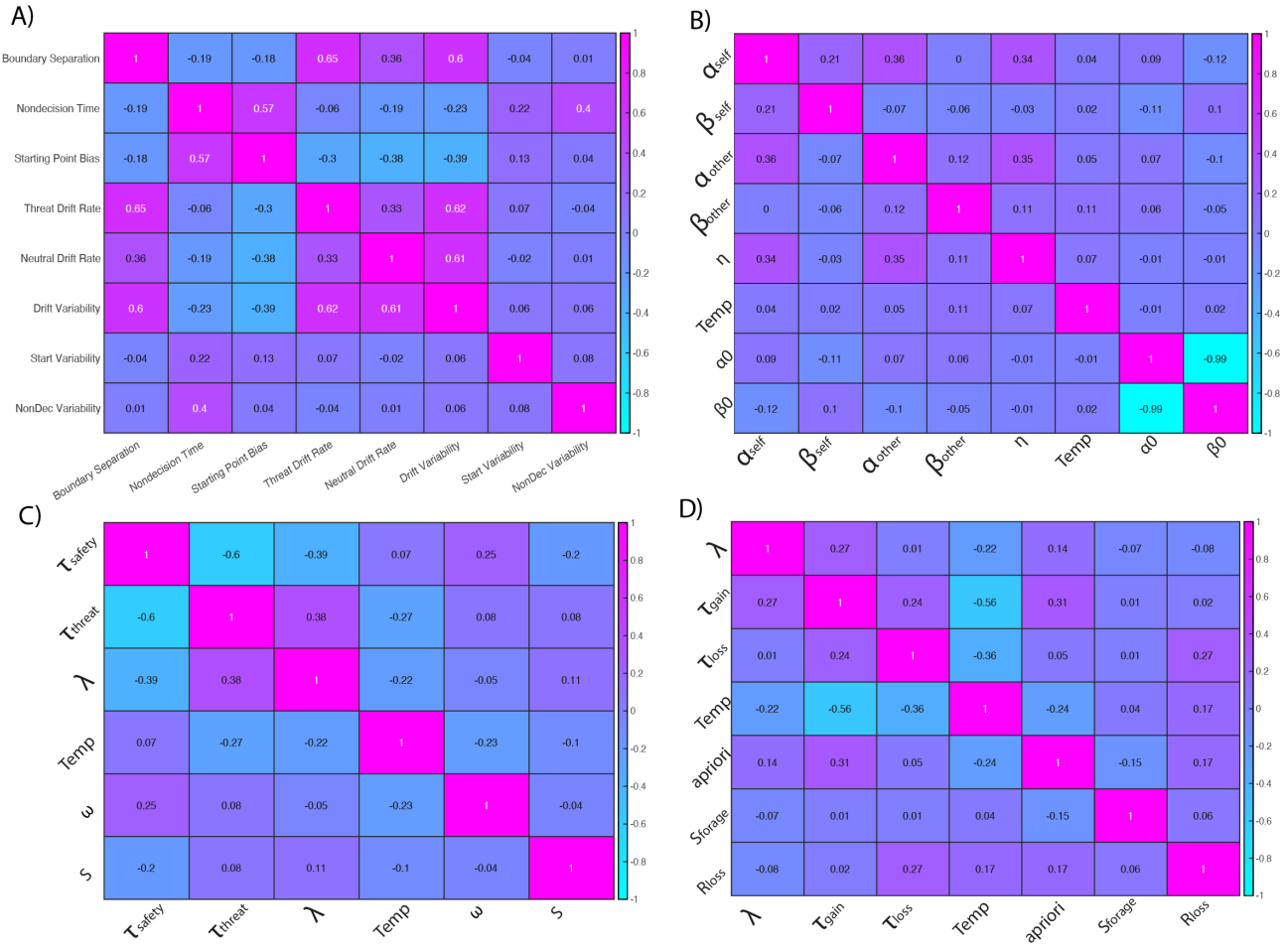

Figure S5.4: Posterior parameter correlations for A) DDM model for the threat bias task. B) The best-fitting belief-update model fit using MCMC. C) The leaky beta model fit to the aversive learning task. D) The leaky beta model for the avoidance task, fit using MLE.

6 Chapter 6: General Discussion

We demand rigidly defined areas of doubt and uncertainty! Douglas Adams, The Hitchhiker's Guide to the Galaxy

We cannot escape uncertainty in our everyday lives and for people with anxiety, the estimation (or mis-estimation) of uncertainty has been theorised to be especially important for cognition, affect and behaviour (Pulcu and Browning, 2019). The work presented in this thesis aimed to elucidate the role that uncertainty plays in anxiety, by developing and applying novel computational models that explicitly quantify uncertainty to elucidate key mechanisms important for anxiety symptomatology. The research in this thesis gives strong evidence that uncertainty models outperform traditional AL models which do not make uncertainty explicit and are useful for understanding differences between anxious and non-anxious individuals in processes such as social evaluation, aversive learning and threat perception.

Studies have proposed that the amygdala may be especially important for uncertainty monitoring and value computation. Reviewing the literature on the amygdala in **Chapter 2** revealed a prominent yet complicated role in value and prediction error signalling. The key suggestions that arose from this review are important for the progression of research within this field. There is also evidence that the amygdala plays a key role in uncertainty monitoring, which is in line with the view of the amygdala as a key area for vigilance monitoring. More research is needed to understand which levels of uncertainty the amygdala plays a key role in tracking. The review also highlights the need for more studies that decompose the structural specificity of the CMA and BLA subregions in tracking different computational

processes such as value learning and research is vitally needed for understanding the role of the amygdala in model-based learning.

The study presented in **Chapter 3** aimed to investigate the neuro-computational mechanisms underlying the relationship between sensory uncertainty and threat perception. Following on from the review in **Chapter 2**, the investigation was targeted at the amygdala, alongside other uncertainty monitoring areas such as the insula, as key areas for sensory uncertainty monitoring. The novel Bayesian Hierarchical model of sensory uncertainty enabled the quantification of trial-by-trial uncertainty computations and the modelling revealed that individuals higher in intolerance of uncertainty had prior expectations about greater uncertainty in the environment. The model-based fMRI results revealed significant insula, but not amygdala activations in correspondence with uncertain stimuli in individuals highly intolerant of uncertainty. The results of this work aid in the understanding of the neuro-computational mechanisms involved in sensory uncertainty processing in the context of threat and suggest that individual differences in intolerance of uncertainty are important to aid understanding of the responses to uncertainty that are not outcome related.

The belief-update uncertainty model that was developed in **Chapter 4** for the investigation of social evaluation learning was essential towards understanding the process underpinning self-negative bias in subclinical social anxiety. The key finding of reduced trait self-positive beliefs suggests that individuals who are fearful of negative evaluation activate less positive beliefs about themselves in social evaluative contexts. These results have clinical implications as it suggests that therapy needs to consider the strength of belief as well as the direction of belief. Moreover, the focus on positive beliefs about the self as therapeutical

target may benefit the understanding of the core mechanisms of treatment programmes that aim to change negative self-belief cycles, such as CBT (Beck, 1971).

The work presented in **Chapter 5** is the first attempt within the Computational Psychiatry literature to understand how different key computational parameters may relate to each other and to anxiety/mood symptoms. This new method is a powerful tool for investigating transdiagnostic computational mechanisms that cut across anxiety and mood disorder categories and make the first steps towards being able to capture anxiety subtypes with a computational profile (Patzelt et al., 2018). The results show that different forms of anxiety show different computational patterns, which could aid in understanding why some people with anxiety respond to treatment, but not others. The results also show that despite these divisions, high anxiety symptoms, alongside low mood have a computational profile that provides useful assays for assessing intervention outcomes (Huys, 2018).

Across the experimental studies, there was no evidence that people with high anxiety have differences in their uncertainty *estimation*. This was the case for different forms of uncertainty, such as uncertainty about the self, sensory uncertainty and safety uncertainty. This was also true for uncertainty that arose from action selection, in the form of RT variability. Moreover, the computational models tested in **Chapter 4** that used an uncertainty weighting in the link functions performed less well than models that ignored uncertainty in action selection. This is in contrast to a study by (Moutoussis et al., 2016) that found a relationship between choice variability and uncertainty, however this study utilised a much more complex task that required social inference as well as discounting. The task utilised in **Chapter 4** required only the reduction of estimation uncertainty and learning curves show that this

was achieved fairly quickly, suggesting this task design might not be optimal for detecting differences in uncertainty. Although sensory uncertainty was unrelated to trait anxiety in **Chapter 3**, there were differences found in individuals who were intolerant of uncertainty. Although these measures are highly correlated, this discrepancy could suggest that the subjective reporting of uncertainty being aversive in self-report more strongly relates to tasks that also assay subjective reporting. The subjective reporting of uncertainty also appeared to be more related to neural differences in the insula, as choice behaviour was unrelated to our measures of anxiety and IUS. The subjective reporting of being averse to uncertainty and the prominence of intolerance of uncertainty in anxiety renders it surprising that none of the uncertainty measures in **Chapter 5** were related to any dimensions of anxiety. One explanation could be that stronger uncertainty manipulations are needed in order to provoke behavioural differences that can be quantified computationally. Studies employing electrical shocks, such as threat-of-shock paradigms (Robinson et al., 2015), for example, induce state uncertainty over aversive outcomes. Using electrical shocks as uncertain outcomes, de Berker et al. (2016) observed that uncertainty was more stress-inducing than certain pain. It is a limitation of online studies that aversive stimuli such as electrical stimulation cannot be used to induce uncertainty, therefore future studies may focus on other ways to induce aversive uncertainty, such as using emotionally shocking images (Simmons et al., 2008) or loud sounds presented unpredictably (Herry et al., 2007; Shankman et al., 2014). Future studies should also explore the possibility that self-reported uncertainty aversion is captured more by subjective in-task reporting, such as probability tracking or the reporting of trial-by-trial fluctuations in emotions. Tasks and computational models that incorporate subjective beliefs and the interplay between beliefs and choice will better able to understand

this role of uncertainty these two levels of cognition (Will et al., 2017, 2020; Gu et al., 2019).

Despite there being no evidence of an explicit mis-estimation of uncertainty relationship to anxiety linked to individual model-derived quantities, a consistent theme throughout this thesis was that computational models that explicitly capture uncertainty outperform (although sometimes only marginally) traditional learning models that do not explicitly quantify uncertainty, such as RW models. Thus, explicit uncertainty modelling is a necessary step in understanding learning and decision-making processes for these individuals. In **Chapter 3**, the novel Bayesian model that tracks trial-by-trial estimates of sensory uncertainty and uses this uncertainty to inform shock probability estimates outperforms RW models that solely maintain probability of shock estimates. Similarly, in **Chapter 4**, the novel belief-update models which update and maintain entire belief distributions over the probability of social approval outperform RW models that maintain only point estimates. Finally, in **Chapter 5**, the belief-update model again outperforms the RW model for social evaluative learning and the novel process of avoidance learning is best captured by a leaky-beta model rather than an associative learning model. These studies add to the growing research that demonstrates the advantageous performance of uncertainty models (Wise et al., 2019; Wise and Dolan, 2020; de Berker et al., 2016) and presents useful model formulations for exploring self-beliefs, even for simple task processes without much uncertainty, sensory uncertainty and avoidance learning. To further develop these models to assess contributions of uncertainty, they could be further adapted to incorporate choice imprecision (Wyart and Koechlin, 2016).

6.1 Limitations

A limitation of this thesis is the generalisability of the results to individuals with clinical anxiety disorders. The sample for **Chapter 3** was taken from SONA and consisted largely of a student population and contained only a small minority of symptomatic individuals. The sample selected for **Chapter 5** was not selected based on any clinical criteria, but used targeted recruitment to enrich the sample according to high scores on trait anxiety measures, in line with taking a dimensional approach to understanding mental health problems (Huys, 2018). This nevertheless requires caution with regards to clinical applications, and further work is required within clinical samples to determine whether the same results hold, or interact with other processes at the more extreme ends of the clinical spectrum.

As stated in **Chapter 4**, a limitation with models with large number of parameters is that they are unreliable at the level of the individual. This can be seen in the relatively weak parameter recoverability for some (albeit non-essential) parameters. Thus, although the models as formulated provide important insights into the psychological process of interest, they cannot be used to computationally characterise individuals in a stable manner, one of the goals of precision psychiatry (Fernandes et al., 2017; Salazar de Pablo et al., 2021).

6.2 Future Directions

6.2.1 Patient samples

As mentioned in the limitations, it would be of great benefit to investigate whether the findings presented in this thesis can also be found in patients. As one of the primary goals of computational psychiatry (and by extension, the work presented here) is to benefit patient

populations, it would be pertinent to determine whether clinical and sub-clinical anxiety groups are similar in terms of their computational biases. It has been proposed that patients lie at the extreme end of a mental health spectrum and studies could aim to recruit samples that include non symptomatic controls, subclinical samples with mid-level symptomatology and patient samples who are highly symptomatic. This would enable researchers to understand whether specific biases are linearly related to symptoms, as has been suggested in information gathering along a 'compulsivity spectrum' (Hauser et al., 2017).

6.2.2 Subjective computational modelling

An important future development to this thesis, especially for **Chapter 3** will be the development of computational models to capture subjective beliefs about perception of threat. It would be of use to determine how subjective perception relates to choice behaviour and whether they have a shared representation, or if one is more informative than the other for understanding behaviour. Thus far, computational modelling has predominantly focused on the modelling of choice behaviour. However, subjective states are important, especially for psychopathology, which is often characterised by negative subjective affect (Gu et al., 2019). Computational modelling studies that make use of subjective models are beginning to be seen (Will et al., 2017, 2020; Low et al., 2021) and therefore paradigms that probe subjective feelings such as in **Chapter 3** may be useful for furthering these investigations.

6.2.3 Carving nature at its joints

Bridging the gap between different studies, experimental paradigms and computational models is an important step towards understanding how these processes interact within

individuals. Future studies could build on and finesse the approach used in the computational phenotyping study of **Chapter 5**. The aim of this study was to determine the relationships between key computational parameters relevant to anxiety, however this could be extended to other psychiatric disorders, using experimental paradigms specifically tailored to each specific disorder. Discovering the patterns and relationships between important learning and decision-making processes within the same individuals is crucial to understanding how the vast array of literature fits together. Studies can also extend this framework more broadly to look at specific concepts, for example 'avoidance' or 'outcome uncertainty', using a combination of measures that assay different features but are united under the same conceptual framework. Similar work using an array of decision-making paradigms for example, has suggested the existence of a 'decision-acuity' factor, akin to the general intelligence factor, which facilitates optimal decision-making (Moutoussis et al., 2021).

6.2.4 Precision psychiatry

In line with dimensional frameworks (Insel et al., 2010; Kotov et al., 2018) and towards precision psychiatry (Fernandes et al., 2017; Salazar de Pablo et al., 2021), studies could aim to accurately and reliably characterise computational profiles at the level of the individual. This is difficult for a number of reasons, one of which being parameter recoverability, an issue that we explored in **Chapter 4**. Building from work presented **Chapter 5**, future studies could seek to use combined measurements of multiple computational processes, reducing noise, and determine whether these measures are reliable over time. If they are, this could be a first step towards creating computational profiles of people at the individual level, an important part of being able to tailor treatments and interventions to the problems specific to

the individual (Fernandes et al., 2017).

6.3 Concluding remarks

This thesis began with a quote by Ursula K. Le Guin, one which perhaps highlights the catch-22 nature of uncertainty for those who find it aversive, such as in high anxiety. On the one hand, uncertainty is completely necessary - if we knew exactly what our lives involved in advance, how would we deal with the negative events we know will inevitably happen? Life would be unrecognisable as we know it, however, we are forced to square this notion with the fact that people, for the most part, prefer certainty. This thesis has explored only a tiny fraction of the possible questions arising from this topic of research and only scratched the surface of how the estimation of different forms of uncertainty is important (or not) for anxiety. However, through the development of computational models that explicitly quantify uncertainty such as those developed here, I hope that this work will be useful for future researchers who wish to build upon these ideas.

7 References

A. T. Beck, Steer, R., and Brown, G. (1996). *Beck Depression Inventory Manual.* The Psychological Corporation., San Antonio, TX.

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, 1191(1):42–61.

Adrián-Ventura, J., Costumero, V., Parcet, M. A., and Ávila, C. (2019). Reward network connectivity "at rest" is associated with reward sensitivity in healthy adults: A resting-state fMRI study. *Cogn Affect Behav Neurosci*, 19(3):726–736.

Ahadi, B. and Basharpoor, S. (2010). Relationship Between Sensory Processing Sensitivity, Personality Dimensions and Mental Health. *Journal of Applied Sciences*, 10.

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer, New York, NY.

Allen, M., Preiss, R. W., Gayle, B. M., and Burrell, N. (2001). *Interpersonal Communication Research: Advances Through Meta-analysis*. Routledge. Google-Books-ID: lLaRAgAAQBAJ.

Alvarez, R. P., Kirlic, N., Misaki, M., Bodurka, J., Rhudy, J. L., Paulus, M. P., and Drevets, W. C. (2015). Increased anterior insula activity in anxious individuals is linked to diminished perceived control. *Transl Psychiatry*, 5(6):e591–e591. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Neuroscience;Scientific community Subject_term_id: diagnostic-markers;neuroscience;scientific-community.

Anderson, A. K. (2007). Feeling emotional: the amygdala links emotional perception and experience. *Social Cognitive and Affective Neuroscience*, 2(2):71–72.

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage*, 13(5):903–919.

Andreatta, M., Michelmann, S., Pauli, P., and Hewig, J. (2017). Learning processes underlying avoidance of negative outcomes. *Psychophysiology*, 54(4):578–590.

Andreatta, M. and Pauli, P. (2015). Appetitive vs. Aversive conditioning in humans. *Front Behav Neurosci*, 9.

Aquino, T. G., Minxha, J., Dunne, S., Ross, I. B., Mamelak, A. N., Rutishauser, U., and O'Doherty, J. P. (2020). Value-Related Neuronal Responses in the Human Amygdala during Observational Learning. *J. Neurosci.*, 40(24):4761–4772. Publisher: Society for Neuroscience Section: Research Articles.

Arditte, K. A., Çek, D., Shaw, A. M., and Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychol Assess*, 28(6):684–691.

Armony, J. L., Corbo, V., Clément, M.-H., and Brunet, A. (2005). Amygdala Response in Patients With Acute PTSD to Masked and Unmasked Emotional Facial Expressions. *American Journal of Psychiatry*, 162(10):1961–1963.

Arnoldini, M., Mostowy, R., Bonhoeffer, S., and Ackermann, M. (2012). Evolution of Stress Response in the Face of Unreliable Environmental Signals. *PLOS Computational Biology*, 8(8):e1002627. Publisher: Public Library of Science.

Aronsson, M., Husberg, M., Kalkan, A., Eckard, N., and Alwin, J. (2014). Differences Between Hypothetical and Experience-Based Value Sets for Eq-5d: Implications for Decision Makers. *Value in Health*, 17(7):A331–A332. Publisher: Elsevier.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C.: American Psychiatric Publishing., Arlington, VA, 5th edition edition.

Atlas, L. Y. (2019). How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. *Current Opinion in Behavioral Sciences*, 26:121–129.

Aylward, J., Hales, C., Robinson, E., and Robinson, O. J. (2019). Translating a rodent measure of negative bias into humans: the impact of induced anxiety and unmedicated mood and anxiety disorders. *Psychol Med*, pages 1–10.

Bach, D. R. (2015). Anxiety-Like Behavioural Inhibition Is Normative under Environmental Threat-Reward Correlations. *PLOS Computational Biology*, 11(12):e1004646.

Bach, D. R. and Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.*, 13(8):572–586.

Bach, D. R., Hulme, O., Penny, W. D., and Dolan, R. J. (2011). The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J Neurosci*, 31(13):4811–4820.

Balderston, N. L., Hsiung, A., Ernst, M., and Grillon, C. (2017). Effect of Threat on Right dlPFC Activity during Behavioral Pattern Separation. *J. Neurosci.*, 37(38):9160–9171. Publisher: Society for Neuroscience Section: Research Articles.

Balleine, B. W. and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4):407–419.

Bardeen, J. R., Fergus, T. A., and Orcutt, H. K. (2017). Examining the specific dimensions of distress tolerance that prospectively predict perceived stress. *Cognitive Behaviour Therapy*, 46(3):211–223. Publisher: Routledge _eprint: https://doi.org/10.1080/16506073.2016.1233454.

Bartra, O., McGuire, J. T., and Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, 76:412–427.

Baxter, A. J., Scott, K. M., Vos, T., and Whiteford, H. A. (2013). Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychological Medicine*, 43(5):897–910.

Baxter, M. G. and Murray, E. A. (2002). The amygdala and reward. *Nat Rev Neurosci*, 3(7):563–573.

Bechara, A., Damasio, H., Damasio, A. R., and Lee, G. P. (1999). Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making. *Journal of Neuroscience*, 19(13):5473–5481.

Beck, A. T. (1971). Cognition, Affect, and Psychopathology. *Arch Gen Psychiatry*, 24(6):495–500.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *Am J Psychiatry*, 165(8):969–977.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry*, 4:561–571.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221. Number: 9 Publisher: Nature Publishing Group.

Beltzer, M. L., Adams, S., Beling, P. A., and Teachman, B. A. (2019). Social Anxiety and Dynamic Social Reinforcement Learning in a Volatile Environment. *Clinical Psychological Science*, 7(6):1372–1388. Publisher: SAGE Publications Inc.

Berker, A. O. d., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., and Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, 7:ncomms10996.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., and Flor, H. (2005). Deficient Fear Conditioning in Psychopathy: A Functional Magnetic Resonance Imaging Study. *Archives of General Psychiatry*, 62(7):799–805.

Bishop, S. J. and Gagne, C. (2018). Anxiety, Depression, and Decision Making: A Computational Perspective. *Annual Review of Neuroscience*, 41(1):371–388.

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., and Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience*, 37(5):758–767. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.12094.

Bonnet, L., Comte, A., Tatu, L., Millot, J.-l., Moulin, T., and Medeiros de Bustos, E. (2015). The role of the amygdala in the perception of positive emotions: an "intensity detector". *Frontiers in Behavioral Neuroscience*, 9.

Boswell, J. F., Thompson-Hollands, J., Farchione, T. J., and Barlow, D. H. (2013). Intolerance of Uncertainty: A Common Factor in the Treatment of Emotional Disorders. *J Clin Psychol*, 69(6).

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Brown, V. M., Chen, J., Gillan, C. M., and Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(6):601–609.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., and Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, 18(4):590–596. Number: 4 Publisher: Nature Publishing Group.

Bröker, F., Marshall, L., Bestmann, S., and Dayan, P. (2018). Forget-me-some: General versus special purpose models in a hierarchical probabilistic task. *PLOS ONE*, 13(10):e0205974. Publisher: Public Library of Science.

Buchkremer, E. M. and Reinhold, K. (2010). The emergence of variance-sensitivity with successful decision rules. *Behavioral Ecology*, 21(3):576–583.

Buhr, K. and Dugas, M. (2002). The Intolerance of Uncertainty Scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, 40(8):931–946. Place: Netherlands Publisher: Elsevier Science.

Burke, C. J. and Tobler, P. N. (2011). Reward skewness coding in the insula independent of probability and loss. *J Neurophysiol*, 106(5):2415–2422.

Butler, G. and Mathews, A. (1983). Cognitive processes in anxiety. *Advances in Behaviour Research and Therapy*, 5(1):51–62.

Button, K. S., Browning, M., Munafò, M. R., and Lewis, G. (2012). Social inference and social anxiety: evidence of a fear-congruent self-referential learning bias. *J Behav Ther Exp Psychiatry*, 43(4):1082–1087.

Button, K. S., Kounali, D., Stapinski, L., Rapee, R. M., Lewis, G., and Munafò, M. R. (2015). Fear of Negative Evaluation Biases Social Evaluation Inference: Evidence from a Probabilistic Learning Task. *PLOS ONE*, 10(4):e0119456.

Calder, A. J. (1996). Facial Emotion Recognition after Bilateral Amygdala Damage: Differentially Severe Impairment of Fear. *Cognitive Neuropsychology*, 13(5):699–745. Publisher: Routledge _eprint: https://doi.org/10.1080/026432996381890.

Calvete, E., Orue, I., and 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.

Camerer, C. and Ho, T. H. (1999). Experience-weighted Attraction Learning in Normal Form Games. *Econometrica*, 67(4):827–874. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00054.

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., Tettamanti, M., and Cappa, S. F. (2013). The Functional and Structural Neural Basis of Individual Differences in Loss Aversion. *J. Neurosci.*, 33(36):14307–14317. Publisher: Society for Neuroscience Section: Articles.

Carleton, R. N., Collimore, K. C., and Asmundson, G. J. G. (2010). "It's not just the judgements—It's that I don't know": Intolerance of uncertainty as a predictor of social anxiety. *Journal of Anxiety Disorders*, 24(2):189–195.

Carleton, R. N., Collimore, K. C., McCabe, R. E., and Antony, M. M. (2011). Addressing revisions to the Brief Fear of Negative Evaluation scale: Measuring fear of negative evaluation across anxiety and mood disorders. *Journal of Anxiety Disorders*, 25(6):822–828.

Carleton, R. N., Duranceau, S., Freeston, M. H., Boelen, P. A., McCabe, R. E., and Antony, M. M. (2014). "But it might be a heart attack": intolerance of uncertainty and panic disorder symptoms. *J Anxiety Disord*, 28(5):463–470.

Carleton, R. N., Mulvogue, M. K., Thibodeau, M. A., McCabe, R. E., Antony, M. M., and Asmundson, G. J. G. (2012). Increasingly certain about uncertainty: Intolerance of uncertainty across anxiety and depression. *Journal of Anxiety Disorders*, 26(3):468–479.

Carleton, R. N., Norton, M. A. P. J., and Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders*, 21(1):105–117.

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

Carstensen, L. L. and DeLiema, M. (2018). The positivity effect: a negativity bias in youth fades with age. *Curr Opin Behav Sci*, 19:7–12.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., and Moffitt, T. E. (2014). The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clin Psychol Sci*, 2(2):119–137.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276.

Cavanagh, S. E., Lam, N. H., Murray, J. D., Hunt, L. T., and Kennerley, S. W. (2020). A circuit mechanism for decision-making biases and NMDA receptor hypofunction. *eLife*, 9:e53664. Publisher: eLife Sciences Publications, Ltd.

Chapman, J. P., Chapman, L. J., and Kwapil, T. R. (1995). Scales for the measurement of schizotypy. In *Schizotypal personality*, pages 79–106. Cambridge University Press, New York, NY, US.

Charpentier, C. J., Aylward, J., Roiser, J. P., and Robinson, O. J. (2017). Enhanced Risk Aversion, But Not Loss Aversion, in Unmedicated Pathological Anxiety. *Biological Psychiatry*, 81(12):1014–1022.

Charpentier, C. J., Bromberg-Martin, E. S., and Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *PNAS*, 115(31):E7255–E7264.

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., and Sharot, T. (2016a). Models of Affective Decision Making: How Do Feelings Predict Choice? *Psychol Sci*, 27(6):763–775.

Charpentier, C. J., Martino, B. D., Sim, A. L., Sharot, T., and Roiser, J. P. (2016b). Emotion-induced loss aversion and striatal-amygdala coupling in low-anxious individuals. *Social Cognitive and Affective Neuroscience*, 11(4):569–579.

Chase, H. W., Kumar, P., Eickhoff, S. B., and Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2):435–459.

Chen, J. T.-H. and Lovibond, P. F. (2016). Intolerance of Uncertainty Is Associated With Increased Threat Appraisal and Negative Affect Under Ambiguity but Not Uncertainty. *Behav Ther*, 47(1):42–53.

Claridge, G. and Beech, T. (1995). Fully and quasi-dimensional constructions of schizotypy. *Schizotypal personality*, pages 192–216.

Clark, D. M. and Wells, A. (1995). A cognitive model of social phobia. In *Social phobia: Diagnosis, assessment, and treatment*, pages 69–93. The Guilford Press, New York, NY, US.

Clarkson, T., Kang, E., Capriola-Hall, N., Lerner, M. D., Jarcho, J., and Prinstein, M. J. (2020). Meta-Analysis of the RDoC Social Processing Domain across Units of Analysis in Children and Adolescents. *Journal of Clinical Child & Adolescent Psychology*, 49(3):297–321.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, 24(4):385–396.

Conversano, C., Rotondo, A., Lensi, E., Della Vista, O., Arpone, F., and Reda, M. A. (2010). Optimism and Its Impact on Mental and Physical Well-Being. *Clin Pract Epidemiol Ment Health*, 6:25–29.

Cooley, C. (1902). *Human nature and the social order*. Human nature and the social order.

Costafreda, S. G., Brammer, M. J., David, A. S., and Fu, C. H. Y. (2008). Predictors of amygdala activation during the processing of emotional stimuli: A meta-analysis of 385 PET and fMRI studies. *Brain Research Reviews*, 58(1):57–70.

Craig, A. D. and Bushnell, M. C. (1994). The thermal grill illusion: Unmasking the burn of cold pain. *Science*, 265(5169):252–255. Place: US Publisher: American Assn for the Advancement of Science.

Craske, M. G., Hermans, D., and Vansteenwegen, D., editors (2006). *Fear and Learning: From basic processes to clinical implications*. American Psychological Association, Washington, DC.

Cuthbert, B. N. and Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, 11(1):126.

Davis, M., Walker, D. L., Miles, L., and Grillon, C. (2010). Phasic vs Sustained Fear in Rats and Humans: Role of the Extended Amygdala in Fear vs Anxiety. *Neuropsychopharmacology*, 35(1):105–135.

Daw, N., Gershman, S., Seymour, B., Dayan, P., and Dolan, R. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215.

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

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.

Dayan, P. and Abbot, L. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press, Cambridge, MA.

Dayan, P. and Huys, Q. J. M. (2008). Serotonin, Inhibition, and Negative Mood. *PLOS Computational Biology*, 4(2):e4. Publisher: Public Library of Science.

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., and Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, 7(1):10996. Number: 1 Publisher: Nature Publishing Group.

Delgado, M. R., Jou, R. L., and Phelps, E. A. (2011). Neural Systems Underlying Aversive Conditioning in Humans with Primary and Secondary Reinforcers. *Frontiers in Neuroscience*, 5.

Delgado, M. R., Li, J., Schiller, D., and Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511):3787–3800.

Deschênes, S. S., Dugas, M. J., and Gouin, J.-P. (2016). Intolerance of uncertainty, worry catastrophizing, and heart rate variability during worry-inducing tasks. *Personality and Individual Differences*, 90:199–204.

Dillon, D. G., Rosso, I. M., Pechtel, P., Killgore, W. D. S., Rauch, S. L., and Pizzagalli, D. A. (2014). Peril and Pleasure: An RDoC-inspired examination of threat responses and reward processing in anxiety and depression. *Depress Anxiety*, 31(3):233–249.

Dorfman, H. M., Bhui, R., Hughes, B. L., and Gershman, S. J. (2019). Causal Inference About Good and Bad Outcomes. *Psychol Sci*, 30(4):516–525. Publisher: SAGE Publications Inc.

Dugas, M. J., Freeston, M. H., and Ladouceur, R. (1997). Intolerance of uncertainty and problem orientation in worry. *Cognitive Therapy and Research*, 21(6):593–606. Place: Germany Publisher: Springer.

Dymond, S. and Roche, B. (2009). A Contemporary Behavior Analysis of Anxiety and Avoidance. *Behav Anal*, 32(1):7–27.

El Khoury-Malhame, M., Reynaud, E., Soriano, A., Michael, K., Salgado-Pineda, P., Zendjidjian, X., Gellato, C., Eric, F., Lefebvre, M.-N., Rouby, F., Samuelian, J.-C., Anton, J.-L., Blin, O., and Khalfa, S. (2011). Amygdala activity correlates with attentional bias in PTSD. *Neuropsychologia*, 49(7):1969–1973.

Eldar, E., Hauser, T. U., Dayan, P., and Dolan, R. J. (2016). Striatal structure and function predict individual biases in learning to avoid pain. *PNAS*, 113(17):4812–4817. Publisher: National Academy of Sciences Section: Biological Sciences.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, 338(6111):1202–1205.

Elliott, R., Friston, K. J., and Dolan, R. J. (2000). Dissociable Neural Responses in Human Reward Systems. *J. Neurosci.*, 20(16):6159–6165. Publisher: Society for Neuroscience Section: ARTICLE.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Engel-Yeger, B. and Dunn, W. (2011). The Relationship between Sensory Processing Difficulties and Anxiety Level of Healthy Adults. *British Journal of Occupational Therapy*, 74(5):210–216. Publisher: SAGE Publications Ltd STM.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., and Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of

self-regulation measures. *Proceedings of the National Academy of Sciences*, 116(12):5472–5477.

Enoch, M.-A., White, K. V., Waheed, J., and Goldman, D. (2008). Neurophysiological and genetic distinctions between pure and comorbid anxiety disorders. *Depression and Anxiety*, 25(5):383–392. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.20378.

Eppinger, B., Walter, M., and Li, S.-C. (2017). Electrophysiological correlates reflect the integration of model-based and model-free decision information. *Cogn Affect Behav Neurosci*, 17(2):406–421.

Ercan, I., Hafizoglu, S., Ozkaya, G., Kirli, S., Yalcintas, E., and Akaya, C. (2015). Examining Cut-Off Values for the State-Trait Anxiety Inventory / Examinando Los Puntajes De Corte Para El Inventario De Ansiedad Estado-Rasgo. *Revista Argentina de Clínica Psicológica; Buenos Aires*, XXIV(II):143.

Farrell, S. and Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Curr Dir Psychol Sci*, 19(5):329–335. Publisher: SAGE Publications Inc.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., and Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive and Affective Neuroscience*, 7(7):743–751.

Fergusson, D. M. and Horwood, L. J. (1995). Predictive validity of categorically and dimensionally scored measures of disruptive childhood behaviors. *J Am Acad Child Adolesc Psychiatry*, 34(4):477–485; discussion 485–487.

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., and Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, 15(1):80.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, 299(5614):1898–1902. Publisher: American Association for the Advancement of Science Section: Report.

Foulds, A., G. and Bedford, A. (1975). Hierarchy of classes of personal illness. *Psychological Medicine*, 5(2):181–92.

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila Parcet, A., and Radua, J. (2016). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*, 21(4):500–508. Number: 4 Publisher: Nature Publishing Group.

Gagne, C., Zika, O., Dayan, P., and Bishop, S. J. (2020). Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife*, 9:e61387. Publisher: eLife Sciences Publications, Ltd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

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

Gelskov, S. V., Henningsson, S., Madsen, K. H., Siebner, H. R., and Ramsøy, T. Z. (2015). Amygdala signals subjective appetitiveness and aversiveness of mixed gambles. *Cortex*, 66:81–90.

Gentes, E. and Ruscio, A. (2011). A meta-analysis of the relation of intolerance of uncertainty to symptoms of generalized anxiety disorder, major depressive disorder, and obsessive-compulsive disorder. *Clinical psychology review*, 31:923–33.

George, S. A., Sheynin, J., Gonzalez, R., Liberzon, I., and Abelson, J. L. (2019). Diminished Value Discrimination in Obsessive-Compulsive Disorder: A Prospect Theory Model of Decision-Making Under Risk. *Frontiers in Psychiatry*, 10.

Gershman, S. J. (2019). The Generative Adversarial Brain. *Front. Artif. Intell.*, 2. Publisher: Frontiers.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., and Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife Sciences*, 5:e11305.

Giorgetta, C., Grecucci, A., Zuanon, S., Perini, L., Balestrieri, M., Bonini, N., Sanfey, A. G., and Brambilla, P. (2012). Reduced risk-taking behavior as a trait feature of anxiety. *Emotion*, 12(6):1373–1383. Place: US Publisher: American Psychological Association.

Gorka, S. M., Nelson, B. D., Phan, K. L., and Shankman, S. A. (2016). Intolerance of uncertainty and insula activation during uncertain reward. *Cogn Affect Behav Neurosci*, 16(5):929–939.

Gorsuch, R. and Nelson, J. (1981). CNG scree test: an objective procedure for determining the number of factors. *In annual meeting of the Society for Multivariate Experimental Psychology.*

Gottfried, J. A., O'Doherty, J., and Dolan, R. J. (2003). Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. *Science*, 301(5636):1104–1107.

Gregory, B. and Peters, L. (2017). Changes in the self during cognitive behavioural therapy for social anxiety disorder: A systematic review. *Clinical Psychology Review*, 52:1–18.

Grillon, C. (2002). Associative learning deficits increase symptoms of anxiety in humans. *Biological Psychiatry*, 51(11):851–858.

Grillon, C., Lissek, S., Rabin, S., McDowell, D., Dvir, S., and Pine, D. S. (2008). Increased Anxiety During Anticipation of Unpredictable But Not Predictable Aversive Stimuli as a Psychophysiologic Marker of Panic Disorder. *AJP*, 165(7):898–904.

Grupe, D. W. and Nitschke, J. B. (2011). Uncertainty Is Associated with Biased Expectancies and Heightened Responses to Aversion. *Emotion*, 11(2):413–424.

Grupe, D. W. and Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat Rev Neurosci*, 14(7):488–501.

Grupe, D. W., Oathes, D. J., and Nitschke, J. B. (2013). Dissecting the anticipation of aversion reveals dissociable neural networks. *Cereb. Cortex*, 23(8):1874–1883.

Grös, D. F., Antony, M. M., Simms, L. J., and McCabe, R. E. (2007). Psychometric properties of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA): comparison to the State-Trait Anxiety Inventory (STAI). *Psychol Assess*, 19(4):369–381.

Gu, X., FitzGerald, T. H. B., and Friston, K. J. (2019). Modeling subjective belief states in computational psychiatry: interoceptive inference as a candidate framework. *Psychopharmacology (Berl)*, 236(8):2405–2412.

Gu, Y., Gu, S., Lei, Y., and Li, H. (2020). From Uncertainty to Anxiety: How Uncertainty Fuels Anxiety in a Process Mediated by Intolerance of Uncertainty. *Neural Plasticity*, 2020:e8866386. Publisher: Hindawi.

Haines, T. (2011). Gaussian Conjugate Prior Cheat Sheet.

Hall, G. (1991). *Perceptual and associative learning.* Clarendon, Oxford.

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *PNAS*, 105(18):6741–6746. Publisher: National Academy of Sciences Section: Biological Sciences.

Hart, A. S., Rutledge, R. B., Glimcher, P. W., and Phillips, P. E. M. (2014). Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. *J Neurosci*, 34(3):698–704.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258.

Hauser, T. U., Moutoussis, M., Dayan, P., and Dolan, R. J. (2017). Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. *Transl Psychiatry*, 7(12):1–10. Number: 12 Publisher: Nature Publishing Group.

Heekeren, H. R., Wartenburger, I., Marschner, A., Mell, T., Villringer, A., and Reischies, F. M. (2007). Role of ventral striatum in reward-based decision making. *NeuroReport*, 18(10):951–955.

Hendriks, S. M., Spijker, J., Licht, C. M. M., Beekman, A. T. F., Hardeveld, F., de Graaf, R., Batelaan, N. M., and Penninx, B. W. J. H. (2014). Disability in anxiety disorders. *J Affect Disord*, 166:227–233.

Herry, C., Bach, D. R., Esposito, F., Salle, F. D., Perrig, W. J., Scheffler, K., Lüthi, A., and Seifritz, E. (2007). Processing of Temporal Unpredictability in Human and Animal Amygdala. *J. Neurosci.*, 27(22):5958–5966. Publisher: Society for Neuroscience Section: Articles.

Hilbert, K., Pine, D. S., Muehlhan, M., Lueken, U., Steudte-Schmiedgen, S., and Beesdo-Baum, K. (2015). Gray and white matter volume abnormalities in generalized anxiety disorder by categorical and dimensional characterization. *Psychiatry Res*, 234(3):314–320.

Hill, W. F. (1960). Learning theory and the acquisition of values. *Psychological Review*, 67(5):317–331.

Hirsch, C. and Mathews, A. (1997). Interpretative inferences when reading about emotional events. *Behaviour Research and Therapy*, 35(12):1123–1132.

Holt, C. A. and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92(5):1644–1655.

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., Pietrzak, R. H., Southwick, S., Krystal, J. H., Harpaz-Rotem, I., and Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature neuroscience*, 22(3):470–476.

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

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, 310(5754):1680–1683.

Huang, H., Thompson, W., and Paulus, M. P. (2017). Computational Dysfunctions in Anxiety: Failure to Differentiate Signal From Noise. *Biol. Psychiatry*, 82(6):440–446.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., and Platt, M. L. (2006). Neural Signatures of Economic Preferences for Risk and Ambiguity. *Neuron*, 49(5):765–775.

Huys, Q. J. M. (2018). Advancing Clinical Improvements for Patients Using the Theory-Driven and Data-Driven Branches of Computational Psychiatry. *JAMA Psychiatry*, 75(3):225–226.

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., and Dayan, P. (2015). Decision-Theoretic Psychiatry. *Clinical Psychological Science*, 3(3):400–421.

Huys, Q. J. M., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., and Popovich, D. L. (2015). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology*, 25(4):652–665. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jcps.2014.12.002.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *AJP*, 167(7):748–751.

Janak, P. H. and Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, 517(7534):284–292.

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford. Google-Books-ID: vh9Act9rtzQC.

Jenison, R. L., Rangel, A., Oya, H., Kawasaki, H., and Howard, M. A. (2011). Value Encoding in Single Neurons in the Human Amygdala during Decision Making. *Journal of Neuroscience*, 31(1):331–338.

Jiang, J., Beck, J., Heller, K., and Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nat Commun*, 6(1):8165. Number: 1 Publisher: Nature Publishing Group.

Johansen, J. P., Cain, C. K., Ostroff, L. E., and LeDoux, J. E. (2011). Molecular Mechanisms of Fear Learning and Memory. *Cell*, 147(3):509–524.

Kahneman, D. (1979). Prospect Theory : An Analysis of Decisions under Risk. *Econometrica*, 47:278.

Kalloniatis, M. and Luu, C. (1995). The Perception of Color. In Kolb, H., Fernandez, E., and Nelson, R., editors, *Webvision: The Organization of the Retina and Visual System*. University of Utah Health Sciences Center, Salt Lake City (UT).

Kaminska, O. K., Magnuski, M., Olszanowski, M., Gola, M., Brzezicka, A., and Winkielman, P. (2020). Ambiguous at the second sight: Mixed facial expressions trigger late electrophysiological responses linked to lower social impressions. *Cogn Affect Behav Neurosci*, 20(2):441–454.

Khdour, H. Y., Abushalbaq, O. M., Mughrabi, I. T., Imam, A. F., Gluck, M. A., Herzallah, M. M., and Moustafa, A. A. (2016). Generalized Anxiety Disorder and Social Anxiety Disorder, but Not Panic Anxiety Disorder, Are Associated with Higher Sensitivity to Learning from Negative Feedback: Behavioral and Computational Investigation. *Frontiers in Integrative Neuroscience*, 10:20.

Kim, M. J., Shin, J., Taylor, J. M., Mattek, A. M., Chavez, S. J., and Whalen, P. J. (2017). Intolerance of uncertainty predicts increased striatal volume. *Emotion*, 17(6):895–899.

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., and Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion*, 17(8):1144–1155.

Koelsch, S. and Skouras, S. (2014). Functional centrality of amygdala, striatum and hypothalamus in a "small-world" network underlying joy: An fMRI study with music. *Human Brain Mapping*, 35(7):3485–3498.

Kohn, A., Kohn, W. K., and Staddon, J. E. (1992). Preferences for constant duration delays and constant sized rewards in human subjects. *Behav Processes*, 26(2-3):125–142.

Kool, W., Gershman, S. J., and Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychol Sci*, 28(9):1321–1333. Publisher: SAGE Publications Inc.

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., and Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *J. Neurosci.*, 32(47):16832–16844.

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., and Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, 44(3):579–592.

Kotov, R., Krueger, R. F., and Watson, D. (2018). A paradigm shift in psychiatric classification: the Hierarchical Taxonomy Of Psychopathology (HiTOP). *World Psychiatry*, 17(1):24–25.

Krain, A. L., Gotimer, K., Hefton, S., Ernst, M., Castellanos, F. X., Pine, D. S., and Milham, M. P. (2008). A Functional Magnetic Resonance Imaging Investigation of Uncertainty in Adolescents with Anxiety Disorders. *Biological Psychiatry*, 63(6):563–568.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1):417–446.

Kriegeskorte, N. and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, 36(3):210–226.

Kube, T., Glombiewski, J. A., Gall, J., Touissant, L., Gärtner, T., and Rief, W. (2019). How to modify persisting negative expectations in major depression? An experimental study comparing three strategies to inhibit cognitive immunization against novel positive experiences. *Journal of Affective Disorders*, 250:231–240.

Kullback, S. (1997). *Information Theory and Statistics*. Courier Corporation. Google-Books-ID: luHcCgAAQBAJ.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., and Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron*, 20(5):937–945.

Lagorio, C. H. and Hackenberg, T. D. (2010). Risky Choice in Pigeons and Humans: A Cross-Species Comparison. *Journal of the Experimental Analysis of Behavior*, 93(1):27–44. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1901/jeab.2010.93-27.

Lake, J. I. and Labar, K. S. (2011). Unpredictability and uncertainty in anxiety: a new direction for emotional timing research. *Front Integr Neurosci*, 5:55.

LaMotte, R. H., Thalhammer, J. G., Torebjork, H. E., and Robinson, C. J. (1982). Peripheral neural mechanisms of cutaneous hyperalgesia following mild injury by heat. *J. Neurosci.*, 2(6):765–781. Publisher: Society for Neuroscience Section: Articles.

Lara, D. R., Pinto, O., Akiskal, K., and Akiskal, H. S. (2006). Toward an integrative model of the spectrum of mood, behavioral and personality disorders based on fear and anger traits: I. Clinical implications. *Journal of Affective Disorders*, 94(1):67–87.

Lavric, A., Rippon, G., and Gray, J. R. (2003). Threat-Evoked Anxiety Disrupts Spatial Working Memory Performance: An Attentional Account. *Cognitive Therapy and Research*, 27(5):489–504.

Leary, M. R. (1983). A Brief Version of the Fear of Negative Evaluation Scale. *Pers Soc Psychol Bull*, 9(3):371–375.

Lebeau, R. T., Glenn, D. E., Hanover, L. N., Beesdo[U+2010]Baum, K., Wittchen, H., and Craske, M. G. (2012). A dimensional approach to measuring anxiety for DSM[U+2010]5. *Int J Methods Psychiatr Res*, 21(4):258–272.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeDoux, J. (2003). The Emotional Brain, Fear, and the Amygdala. *Cellular and Molecular Neurobiology*, 23(4):727–738.

LeDoux, J. (2012). Rethinking the Emotional Brain. *Neuron*, 73(4):653–676.

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences*, 111(8):2871–2878.

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699.

Lewis, G., Pelosi, A. J., Araya, R., and Dunn, G. (1992). Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine*, 22(2):465–486. Publisher: Cambridge University Press.

Li, J. and Daw, N. D. (2011). Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction. *Journal of Neuroscience*, 31(14):5504–5511.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., and Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, 14(10):1250–1252.

Li, S. S. Y. and McNally, G. P. (2014). The conditions that promote fear learning: Prediction error and Pavlovian fear conditioning. *Neurobiology of Learning and Memory*, 108:14–21.

Limongi, R., Pérez, F. J., Modroño, C., and González-Mora, J. L. (2016). Temporal Uncertainty and Temporal Estimation Errors Affect Insular Activity and the Frontostriatal Indirect Pathway during Action Update: A Predictive Coding Study. *Frontiers in Human Neuroscience*, 10:276.

Lindström, B. and Tobler, P. N. (2018). Incidental ostracism emerges from simple learning mechanisms. *Nat Hum Behav*, 2(6):405–414. Number: 6 Publisher: Nature Publishing Group.

Liss, M., Timmel, L., Baxley, K., and Killingsworth, P. (2005). Sensory processing sensitivity and its relation to parental bonding, anxiety, and depression. *Personality and Individual Differences*, 39(8):1429–1439.

Locey, M. L., Pietras, C. J., and Hackenberg, T. D. (2009). Human risky choice: Delay sensitivity depends on reinforcer type. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(1):15–22. Place: US Publisher: American Psychological Association.

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., and Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *PNAS*, 113(35):9763–9768. ISBN: 9781603198110 Publisher: National Academy of Sciences Section: Social Sciences.

Lockwood, P. L., Wittmann, M. K., Apps, M. A. J., Klein-Flügge, M. C., Crockett, M. J., Humphreys, G. W., and Rushworth, M. F. S. (2018). Neural mechanisms for learning self and other ownership. *Nature Communications*, 9(1):4747. Number: 1 Publisher: Nature Publishing Group.

Low, A. A. Y., Hopper, W. J. T., Angelescu, I., Mason, L., Will, G.-J., and Moutoussis, M. (2021). Self-Esteem depends on beliefs about the rate of change of social approval. Technical report, PsyArXiv. type: article.

Ludvig, E. A. and Spetch, M. L. (2011). Of Black Swans and Tossed Coins: Is the Description-Experience Gap in Risky Choice Limited to Rare Events? *PLOS ONE*, 6(6):e20262. Publisher: Public Library of Science.

MacLeod, C., Mathews, A., and Tata, P. (1986). Attentional bias in emotional disorders. *J Abnorm Psychol*, 95(1):15–20.

Maner, J. K., Richey, J. A., Cromer, K., Mallott, M., Lejuez, C. W., Joiner, T. E., and Schmidt, N. B. (2007). Dispositional anxiety and risk-avoidant decision-making. *Personality and Individual Differences*, 42(4):665–675.

Mansell, W., Ehlers, A., Clark, D., and Chen, Y.-P. (2002). Attention to Positive and Negative Social-Evaluative Words: Investigating the Effects of Social Anxiety, Trait Anxiety and

Social Threat. *Anxiety, Stress, & Coping*, 15(1):19–29. Publisher: Routledge _eprint: https://doi.org/10.1080/10615800290007263.

Marks, I. f. and Nesse, R. M. (1994). Fear and fitness: An evolutionary analysis of anxiety disorders. *Ethology and Sociobiology*, 15(5):247–261.

Martino, B. D., Camerer, C. F., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian Foundation for Individual Learning Under Uncertainty. *Front. Hum. Neurosci.*, 5. Publisher: Frontiers.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Front. Hum. Neurosci.*, 8. Publisher: Frontiers.

McClure, S. M., Berns, G. S., and Montague, P. R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron*, 38(2):339–346.

McDannald, M. A., Takahashi, Y. K., Lopatina, N., Pietras, B. W., Jones, J. L., and Schoenbaum, G. (2012). Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. *Eur J Neurosci*, 35(7):991–996.

McEvoy, P. M. and Mahoney, A. E. J. (2011). Achieving certainty about the structure of intolerance of uncertainty in a treatment-seeking sample with anxiety and depression. *Journal of Anxiety Disorders*, 25(1):112–122.

McEvoy, P. M. and Mahoney, A. E. J. (2012). To Be Sure, To Be Sure: Intolerance of Uncertainty Mediates Symptoms of Various Anxiety Disorders and Depression. *Behavior Therapy*, 43(3):533–545.

McGuire, J. T., Nassar, M. R., Gold, J. I., and Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, 84(4):870–881.

McKerchar, T. L. and Mazur, J. E. (2016). Human choices between variable and fixed rewards in hypothetical variable-delay and double-reward discounting procedures. *Journal of the Experimental Analysis of Behavior*, 106(1):1–21. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jeab.214.

Meacham, F. and Bergstrom, C. (2016). Adaptive behavior can produce maladaptive anxiety due to individual differences in experience. *Evol Med Public Health*, 2016(1):270–285.

Michely, J., Rigoli, F., Rutledge, R. B., Hauser, T. U., and Dolan, R. J. (2020). Distinct Processing of Aversive Experience in Amygdala Subregions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(3):291–300.

Mitte, K. (2007). Anxiety and risky decision-making: The role of subjective probability and subjective costs of negative events. *Personality and Individual Differences*, 43(2):243–253.

Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., and Robinson, O. J. (2017). Modeling Avoidance in Mood and Anxiety Disorders Using Reinforcement Learning. *Biological Psychiatry*, 82(7):532–539.

Mobbs, D., Petrovic, P., Marchant, J. L., Hassabis, D., Weiskopf, N., Seymour, B., Dolan, R. J., and Frith, C. D. (2007). When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science*, 317(5841):1079–1083.

Monosov, I. E. (2017). Anterior cingulate is a source of valence-specific information about value and uncertainty. *Nat Commun*, 8(1):134.

Moore, M. T. and Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6):496–509.

Morriss, J., Christakou, A., and van Reekum, C. M. (2015). Intolerance of uncertainty predicts fear extinction in amygdala-ventromedial prefrontal cortical circuitry. *Biology of Mood & Anxiety Disorders*, 5(1):4.

Morriss, J., Christakou, A., and van Reekum, C. M. (2016). Nothing is safe: Intolerance of uncertainty is associated with compromised fear extinction learning. *Biological Psychology*, 121:187–193.

Moutoussis, M., Dolan, R. J., and Dayan, P. (2016). How People Use Social Information to Find out What to Want in the Paradigmatic Case of Inter-temporal Preferences. *PLOS Computational Biology*, 12(7):e1004965.

Moutoussis, M., Garzón, B., Neufeld, S., Bach, D., Rigoli, F., Goodyer, I. M., Bullmore, E. T., Guitart-Masip, M., Dolan, R. J., Fonagy, P., Jones, P., Hauser, T., Romero-Garcia, R., St Clair, M., Vértes, P., Whitaker, K. L., Inkster, B., Prabhu, G., Ooi, C., Toseeb, U., Widmer, B., Bhatti, J., Villis, L., Alrumaithi, A., Birt, S., Bowler, A., Cleridou, K., Dadabhoy, H., Davies, E., Firkins, A., Granville, S., Harding, E., Hopkins, A., Isaacs, D., King, J., Kokorikou, D., Maurice, C., McIntosh, C., Memarzia, J., Mills, H., O'Donnell, C., Pantaleone, S., Scott, J., Fearon, P., Suckling, J., van Harmelen, A.-L., and Kievit, R. (2021). Decision-making ability, psychopathology, and brain connectivity. *Neuron*. Publisher: Elsevier BV.

Moutoussis, M., Hopkins, A. K., and Dolan, R. J. (2018). Hypotheses About the Relationship of Cognition With Psychopathology Should be Tested by Embedding Them Into Empirical Priors. *Front. Psychol.*, 9.

Moutoussis, M., Shahar, N., Hauser, T. U., and Dolan, R. J. (2017). Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Computational Psychiatry*, 2:50–73.

Mueller, E. M., Nguyen, J., Ray, W. J., and Borkovec, T. D. (2010). Future-oriented decision-making in Generalized Anxiety Disorder is evident across different versions of the Iowa Gambling Task. *Journal of Behavior Therapy and Experimental Psychiatry*, 41(2):165–171.

Nastase, S., Iacovella, V., and Hasson, U. (2014). Uncertainty in visual and auditory series is coded by modality-general and modality-specific neural systems. *Hum Brain Mapp*, 35(4):1111–1128.

Nastase, S. A., Davis, B., and Hasson, U. (2018). Cross-modal and non-monotonic representations of statistical regularity are encoded in local neural response patterns. *bioRxiv*, page 243550. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Nelson, B. D. and Shankman, S. A. (2011). Does intolerance of uncertainty predict anticipatory startle responses to uncertain threat? *Int J Psychophysiol*, 81(2):107–115.

Norbury, A., Robbins, T. W., and Seymour, B. (2018). Value generalization in human avoidance learning. *eLife*, 7:e34779. Publisher: eLife Sciences Publications, Ltd.

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H. B., Coello, C., Wall, M. B., Dolan, R. J., and Howes, O. D. (2018). Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. *PNAS*, 115(43):E10167–E10176. Publisher: National Academy of Sciences Section: PNAS Plus.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, 38(2):329–337.

Oglesby, M. E., Boffa, J. W., Short, N. A., Raines, A. M., and Schmidt, N. B. (2016). Intolerance of uncertainty as a predictor of post-traumatic stress symptoms following a traumatic event. *J Anxiety Disord*, 41:82–87.

Pagnoni, G., Zink, C. F., Montague, P. R., and Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, 5(2):97–98.

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, 21(6):425–433.

Patel, N., Vytal, K., Pavletic, N., Stoodley, C., Pine, D. S., Grillon, C., and Ernst, M. (2016). Interaction of threat and verbal working memory in adolescents. *Psychophysiology*, 53(4):518–526. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.12582.

Patzelt, E. H., Hartley, C. A., and Gershman, S. J. (2018). Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personality Neuroscience*, 1. Publisher: Cambridge University Press.

Patzelt, E. H., Kool, W., Millner, A. J., and Gershman, S. J. (2019). The transdiagnostic structure of mental effort avoidance. *Sci Rep*, 9(1):1689. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Decision;Human behaviour;Motivation Subject_term_id: decision;human-behaviour;motivation.

Payzan-LeNestour, E. and Bossaerts, P. (2011). Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLOS Computational Biology*, 7(1):e1001048. Publisher: Public Library of Science.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J. P. (2013). The Neural Representation of Unexpected Uncertainty During Value-Based Decision Making. *Neuron*, 79(1):191–201.

Pearce, J. M. and Hall, G. B. C. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6):532–552.

Perry, C. J. and Barron, A. B. (2013). Honey bees selectively avoid difficult choices. *PNAS*, 110(47):19155–19159. Publisher: National Academy of Sciences Section: Biological Sciences.

Pierce, M., Hope, H., Ford, T., Hatch, S., Hotopf, M., John, A., Kontopantelis, E., Webb, R., Wessely, S., McManus, S., and Abel, K. M. (2020). Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population. *The Lancet Psychiatry*, 7(10):883–892.

Pinto-Gouveia, J., Castilho, P., Galhardo, A., and Cunha, M. (2006). Early Maladaptive Schemas and Social Phobia. *Cogn Ther Res*, 30(5):571–584.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.

Preuschoff, K., Quartz, S. R., and Bossaerts, P. (2008). Human Insula Activation Reflects Risk Prediction Errors As Well As Risk. *J. Neurosci.*, 28(11):2745–2752. Publisher: Society for Neuroscience Section: Articles.

Prévost, C., McCabe, J. A., Jessup, R. K., Bossaerts, P., and O'Doherty, J. P. (2011). Differentiable contributions of human amygdalar subregions in the computations underlying reward and avoidance learning. *European Journal of Neuroscience*, 34(1):134–145.

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., and O'Doherty, J. P. (2013). Evidence for Model-based Computations in the Human Amygdala during Pavlovian Conditioning. *PLOS Computational Biology*, 9(2):e1002918.

Pulcu, E. and Browning, M. (2019). The Misestimation of Uncertainty in Affective Disorders. *Trends in Cognitive Sciences*, 23(10):865–875. Publisher: Elsevier.

Qin, S., Young, C. B., Duan, X., Chen, T., Supekar, K., and Menon, V. (2014). Amygdala Subregional Structure and Intrinsic Functional Connectivity Predicts Individual Differences in Anxiety During Early Childhood. *Biological Psychiatry*, 75(11):892–900.

Ramnani, N., Elliott, R., Athwal, B. S., and Passingham, R. E. (2004). Prediction error for free monetary reward in the human prefrontal cortex. *NeuroImage*, 23(3):777–786.

Rapee, R. M. and Heimberg, R. G. (1997). A cognitive-behavioral model of anxiety in social phobia. *Behav Res Ther*, 35(8):741–756.

Raymond, J. G., Steele, J. D., and Seriès, P. (2017). Modeling Trait Anxiety: From Computational Processes to Personality. *Front Psychiatry*, 8.

Rescorla, R. and Wagner, A. R. (1972). 3 A Theory of Pavlovian Conditioning : Variations in the Effectiveness of Reinforcement and Nonreinforcement.

Ritchie, H. and Roser, M. (2018). Mental Health. *Our World in Data*.

Robinson, O. J., Bond, R. L., and Roiser, J. P. (2015). The impact of threat of shock on the framing effect and temporal discounting: executive functions unperturbed by acute stress? *Front. Psychol.*, 6. Publisher: Frontiers.

Robinson, O. J., Charney, D. R., Overstreet, C., Vytal, K., and Grillon, C. (2012). The adaptive threat bias in anxiety: Amygdala–dorsomedial prefrontal cortex coupling and aversive amplification. *NeuroImage*, 60(1):523–529.

Robinson, O. J., Letkiewicz, A. M., Overstreet, C., Ernst, M., and Grillon, C. (2011). The effect of induced anxiety on cognition: threat of shock enhances aversive processing in healthy individuals. *Cogn Affect Behav Neurosci*, 11(2):217.

Robinson, O. J., Pike, A. C., Cornwell, B., and Grillon, C. (2019). The translational neural circuitry of anxiety. *J Neurol Neurosurg Psychiatry*, 90(12):1353–1360. Publisher: BMJ Publishing Group Ltd Section: Neuropsychiatry.

Robinson, O. J., Vytal, K., Cornwell, B. R., and Grillon, C. (2013). The impact of anxiety upon cognition: perspectives from human threat of shock studies. *Front Hum Neurosci*, 7.

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., and Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *J Abnorm Psychol*, 125(6):840–851.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., and Schoenbaum, G. (2012a). Surprise! Neural Correlates of Pearce-Hall and Rescorla-Wagner Coexist within the Brain. *Eur J Neurosci*, 35(7):1190–1200.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., and Schoenbaum, G. (2012b). Surprise! Neural Correlates of Pearce-Hall and Rescorla-Wagner Coexist within the Brain. *The European Journal of Neuroscience*, 35(7):1190–1200.

Rosenberg, M. (1965). Rosenberg Self-Esteem Scale (RSES). *APA PsycTests.*

Rouault, M., Seow, T., Gillan, C. M., and Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, 84(6):443–451.

Roy, A. K., Shehzad, Z., Margulies, D. S., Kelly, A. M. C., Uddin, L. Q., Gotimer, K., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2009). Functional connectivity of the human amygdala using resting state fMRI. *NeuroImage*, 45(2):614–626.

Russo, S. J. and Nestler, E. J. (2013). The brain reward circuitry in mood disorders. *Nat Rev Neurosci*, 14(9):609–625.

Rutledge, R. B., Dean, M., Caplin, A., and Glimcher, P. W. (2010). Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. *J. Neurosci.*, 30(40):13525–13536. Publisher: Society for Neuroscience Section: Articles.

Rutledge, R. B., Skandali, N., Dayan, P., and Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *PNAS*, 111(33):12252–12257.

Sabatinelli, D., Fortune, E. E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W. T., Beck, S., and Jeffries, J. (2011). Emotional perception: Meta-analyses of face and natural scene processing. *NeuroImage*, 54(3):2524–2533.

Sah, P., Faber, E. S. L., Lopez De Armentia, M., and Power, J. (2003). The Amygdaloid Complex: Anatomy and Physiology. *Physiological Reviews*, 83(3):803–834.

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., and Fusar-Poli, P. (2021). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophr Bull*, 47(2):284–297.

Salzman, C. D. and Fusi, S. (2010). Emotion, Cognition, and Mental State Representation in Amygdala and Prefrontal Cortex. *Annual Review of Neuroscience*, 33(1):173–202. _eprint: https://doi.org/10.1146/annurev.neuro.051508.135256.

Sanborn, A. N. and Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, 20(12):883–893.

Sarinopoulos, I., Grupe, D. W., Mackiewicz, K. L., Herrington, J. D., Lor, M., Steege, E. E., and Nitschke, J. B. (2010). Uncertainty during Anticipation Modulates Neural Responses to Aversion in Human Insula and Amygdala. *Cerebral Cortex*, 20(4):929–940.

Satterthwaite, T. D., Wolf, D. H., Pinkham, A. E., Ruparel, K., Elliott, M. A., Valdez, J. N., Overton, E., Seubert, J., Gur, R. E., Gur, R. C., and Loughead, J. (2011). Opposing amygdala and ventral striatum connectivity during emotion identification. *Brain and Cognition*, 76(3):353–363.

Savage, L. J. (1972). *The Foundations of Statistics*. Courier Corporation. Google-Books-ID: zSv6dBWneMEC.

Schienle, A., Köchel, A., Ebner, F., Reishofer, G., and Schäfer, A. (2010). Neural correlates of intolerance of uncertainty. *Neuroscience Letters*, 479(3):272–276.

Schiller, D., Levy, I., Niv, Y., LeDoux, J. E., and Phelps, E. A. (2008). From Fear to Safety and Back: Reversal of Fear in the Human Brain. *The Journal of Neuroscience*, 28(45):11517–11525.

Schultz, D. H., Balderston, N. L., Baskin-Sommers, A. R., Larson, C. L., and Helmstetter, F. J. (2016). Psychopaths Show Enhanced Amygdala Activation during Fear Conditioning. *Front. Psychol.*, 7. Publisher: Frontiers.

Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C. D., Tobler, P. N., and Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511):3801–3811. Publisher: Royal Society.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.*, 6(2):461–464.

Sege, C. T., Bradley, M. M., and Lang, P. J. (2018). Avoidance and escape: Defensive reactivity and trait anxiety. *Behaviour Research and Therapy*, 104:62–68.

Seymour, B., Daw, N., Dayan, P., Singer, T., and Dolan, R. (2007). Differential Encoding of Losses and Gains in the Human Striatum. *Journal of Neuroscience*, 27(18):4826–4831.

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, 8(9):1234–1240.

Shackman, A. J., Sarinopoulos, I., Maxwell, J. S., Pizzagalli, D. A., Lavric, A., and Davidson, R. J. (2006). Anxiety selectively disrupts visuospatial working memory. *Emotion*, 6(1):40–61. Place: US Publisher: American Psychological Association.

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., and Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Comput Biol*, 15(2).

Shankman, S. A., Gorka, S. M., Nelson, B. D., Fitzgerald, D. A., Phan, K. L., and O'Daly, O. (2014). Anterior Insula Responds to Temporally Unpredictable Aversiveness: an fMRI Study. *Neuroreport*, 25(8):596–600.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11):1475–1479.

Sharp, P. B., Dolan, R. J., and Eldar, E. (2020). Disrupted state transition learning as a computational marker of compulsivity. Technical report, PsyArXiv. type: article.

Siegle, G. J., Thompson, W., Carter, C. S., Steinhauer, S. R., and Thase, M. E. (2007). Increased Amygdala and Decreased Dorsolateral Prefrontal BOLD Responses in Unipolar Depression: Related and Independent Features. *Biological Psychiatry*, 61(2):198–209.

Simmons, A., Matthews, S. C., Paulus, M. P., and Stein, M. B. (2008). Intolerance of uncertainty correlates with insula activation during affective ambiguity. *Neuroscience Letters*, 430(2):92–97.

Smith, R., Killgore, W. D. S., and Lane, R. D. (2018). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, 18(5):670–692. Place: US Publisher: American Psychological Association.

Sokol-Hessner, P., Camerer, C. F., and Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Social Cognitive and Affective Neuroscience*, 8(3):341–350.

Somerville, L. H., Wagner, D. D., Wig, G. S., Moran, J. M., Whalen, P. J., and Kelley, W. M. (2013). Interactions Between Transient and Sustained Neural Signals Support the Generation and Regulation of Anxious Emotion. *Cerebral Cortex*, 23(1):49–60.

Sowislo, J. F. and Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychol Bull*, 139(1):213–240.

Spence, S. H. and Rapee, R. M. (2016). The etiology of social anxiety disorder: An evidence-based model. *Behav Res Ther*, 86:50–67.

Spielberger, C. D. (2010). State-Trait Anxiety Inventory. In *The Corsini Encyclopedia of Psychology*, pages 1–1. American Cancer Society.

Spielberger, C. D., Gorsuch, R., Lushene, R., Vagg, P., and Jacobs, G. (1983). Manual for the State-Trait Anxiety Inventory. *Palo Alto, CA: Consulting Psychologists Press*.

Stankevicius, A., Huys, Q. J. M., Kalra, A., and Seriès, P. (2014). Optimism as a Prior Belief about the Probability of Future Reward. *PLOS Computational Biology*, 10(5):e1003605.

Steele, C. C., Gwinner, M., Smith, T., Young, M. E., and Kirkpatrick, K. (2019). Experience Matters: The Effects of Hypothetical versus Experiential Delays and Magnitudes on Impulsive Choice in Delay Discounting Tasks. *Brain Sci*, 9(12):379.

Stein, M. B., Simmons, A. N., Feinstein, J. S., and Paulus, M. P. (2007). Increased Amygdala and Insula Activation During Emotion Processing in Anxiety-Prone Subjects. *American Journal of Psychiatry*, 164(2):318–327.

Stolyarova, A., Rakhshan, M., Hart, E. E., O'Dell, T. J., Peters, M. a. K., Lau, H., Soltani, A., and Izquierdo, A. (2019). Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nature Communications*, 10(1):4704. Number: 1 Publisher: Nature Publishing Group.

Stopa, L. and Clark, D. M. (2001). Social phobia: Comments on the viability and validity of an analogue research strategy and British norms for the fear of negative evaluation questionnaire. *Behavioural and Cognitive Psychotherapy*, 29(4):423–430.

Story, G. W., Vlaev, I., Seymour, B., Winston, J. S., Darzi, A., and Dolan, R. J. (2013). Dread and the Disvalue of Future Pain. *PLOS Computational Biology*, 9(11):e1003335.

Sutton, R. S., Barto, A. G., and Williams, R. J. (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazine*, 12(2):19–22.

Tanovic, E., Gee, D. G., and Joormann, J. (2018). Intolerance of uncertainty: Neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clinical Psychology Review*, 60:87–99.

Taylor, C. T. and Alden, L. E. (2010). Safety behaviors and judgmental biases in social anxiety disorder. *Behaviour Research and Therapy*, 48(3):226–237.

Taylor, T. L. and Montgomery, P. (2007). Can cognitive-behavioral therapy increase self-esteem among depressed adolescents? A systematic review. *Children and Youth Services Review*, 29(7):823–839.

Thomas, C. L. and Cassady, J. C. (2021). Validation of the State Version of the State-Trait Anxiety Inventory in a University Sample. *SAGE Open*, 11(3):21582440211031900. Publisher: SAGE Publications.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J. Neurophysiol.*, 97(2):1621–1632.

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., and Foa, E. B. (2003). Intolerance of uncertainty in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, 17(2):233–242.

Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, 315(5811):515–518.

Tsetsos, K., Chater, N., and Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *PNAS*, 109(24):9659–9664.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., and Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *PNAS*, 113(11):3102–3107. Publisher: National Academy of Sciences Section: Biological Sciences.

Tzovara, A., Meyer, S. S., Bonaiuto, J. J., Abivardi, A., Dolan, R. J., Barnes, G. R., and Bach, D. R. (2019). High-precision magnetoencephalography for reconstructing amygdalar and hippocampal oscillations during prediction of safety and threat. *Human Brain Mapping*, 40(14):4114–4129. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.24689.

Van Dam, N. T., Gros, D. F., Earleywine, M., and Antony, M. M. (2013). Establishing a trait anxiety threshold that signals likelihood of anxiety disorders. *Anxiety Stress Coping*, 26(1):70–86.

van Marle, H. J. F., Hermans, E. J., Qin, S., and Fernández, G. (2009). From Specificity to Sensitivity: How Acute Stress Affects Amygdala Processing of Biologically Salient Stimuli. *Biological Psychiatry*, 66(7):649–655.

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

Vilares, I., Howard, J., Fernandes, H., Gottfried, J., and Kording, K. (2012). Differential Representations of Prior and Likelihood Uncertainty in the Human Brain. *Current Biology*, 22(18):1641–1648.

Vilares, I. and Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*,

1224(1):22–39. _eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2011.05965.x.

Vives, M.-L. and FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature Communications*, 9(1):2156. Number: 1 Publisher: Nature Publishing Group.

von Kluge, S. (1992). Trading Accuracy for Speed: Gender Differences on a Stroop Task under Mild Performance Anxiety. *Percept Mot Skills*, 75(2):651–657. Publisher: SAGE Publications Inc.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Voss, A. and Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, 39(4):767–775.

Vytal, K. E., Cornwell, B. R., Arkin, N. E., Letkiewicz, A. M., and Grillon, C. (2013). The complex interaction between anxiety and cognition: insight from spatial and verbal working memory. *Front. Hum. Neurosci.*, 7. Publisher: Frontiers.

Watanabe, N., Sakagami, M., and Haruno, M. (2013). Reward Prediction Error Signal Enhanced by Striatum–Amygdala Interaction Explains the Acceleration of Probabilistic Reward Learning by Emotion. *Journal of Neuroscience*, 33(10):4487–4493.

Weiner, H. (1966). Preference and Switching under Ratio Contingencies with Humans. *Psychol Rep*, 18(1):239–246. Publisher: SAGE Publications Inc.

Weiskopf, N., Hutton, C., Josephs, O., and Deichmann, R. (2006). Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3 T and 1.5 T. *NeuroImage*, 33(2):493–504.

Weymar, M., Keil, A., and Hamm, A. O. (2014). Timing the fearful brain: unspecific hypervigilance and spatial attention in early visual perception. *Soc Cogn Affect Neurosci*, 9(5):723–729.

Weymar, M. and Schwabe, L. (2016). Amygdala and Emotion: The Bright Side of It. *Frontiers in Neuroscience*, 10.

White, C. N., Ratcliff, R., Vasey, M. W., and McKoon, G. (2010a). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion*, 10(5):662–677.

White, C. N., Ratcliff, R., Vasey, M. W., and McKoon, G. (2010b). Using diffusion models to understand clinical disorders. *J Math Psychol*, 54(1):39–52.

White, C. N., Skokin, K., Carlos, B., and Weaver, A. (2016). Using decision models to decompose anxiety-related bias in threat classification. *Emotion*, 16(2):196–207. Place: US Publisher: American Psychological Association.

Wiecki, T., Sofer, I., and Frank, M. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, 7:14.

Wieser, M. J. and Keil, A. (2020). Attentional threat biases and their role in anxiety: A neurophysiological perspective. *International Journal of Psychophysiology*, 153:148–158.

Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Rutledge, R. B., and Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, 10(1):1–14. Number: 1 Publisher: Nature Publishing Group.

Will, G.-J., Rutledge, R. B., Moutoussis, M., and Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife*, 6:e28098.

Willis, M. L., Dodd, H. F., and Palermo, R. (2013). The Relationship between Anxiety and the Social Judgements of Approachability And Trustworthiness. *PLOS ONE*, 8(10):e76825. Publisher: Public Library of Science.

Winton, E. C., Clark, D. M., and Edelmann, R. J. (1995). Social anxiety, fear of negative evaluation and the detection of negative emotion in others. *Behaviour Research and Therapy*, 33(2):193–196.

Wise, T. and Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, 11(1):1–13.

Wise, T., Liu, Y., Chowdhury, F., and Dolan, R. J. (2020). Model-based aversive learning in humans is supported by preferential task state reactivation. *bioRxiv*, page 2020.11.30.404491. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Wise, T., Michely, J., Dayan, P., and Dolan, R. J. (2019). A computational account of threat-related attentional bias. *PLOS Computational Biology*, 15(10):e1007341. Publisher: Public Library of Science.

Wyart, V. and Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, 11:109–115.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., and Büchel, C. (2006). Dissociable Systems for Gain- and Loss-Related Value Predictions and Errors of Prediction in the Human Brain. *Journal of Neuroscience*, 26(37):9530–9537.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yoon, K. L. and Zinbarg, R. E. (2007). Threat is in the eye of the beholder: social anxiety and the interpretation of ambiguous facial expressions. *Behav Res Ther*, 45(4):839–847.

Yoon, K. L. and Zinbarg, R. E. (2008). Interpreting neutral faces as threatening is a default mode for socially anxious individuals. *Journal of Abnormal Psychology*, 117(3):680–685. Place: US Publisher: American Psychological Association.

Zajkowski, W., Krzemiński, D., Barone, J., Evans, L., and Zhang, J. (2019). Reward certainty and preference bias selectively shape voluntary decisions. *bioRxiv*, page 832311. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Zhang, L. and Gläscher, J. (2020). A brain network supporting social influences in human decision-making. *Science Advances*, 6(34):eabb4159. Publisher: American Association for the Advancement of Science Section: Research Article.

Zhang, S., Mano, H., Ganesh, G., Robbins, T., and Seymour, B. (2016). Dissociable Learning Processes Underlie Human Pain Conditioning. *Current Biology*, 26(1):52–58.

Zhou, X., Xu, Q., Inglés, C. J., Hidalgo, M. D., and La Greca, A. M. (2008). Reliability and Validity of the Chinese Version of the Social Anxiety Scale for Adolescents. *Child Psychiatry Hum Dev*, 39(2):185–200.

Zhou, Y., Acerbi, L., and Ma, W. J. (2020). The role of sensory uncertainty in simple contour integration. *PLOS Computational Biology*, 16(11):e1006308. Publisher: Public Library of Science.