Neural Computations Underpinning Anxiety in Health and Disease

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA MEDICAL SCHOOL
BY

CODY J. WALTERS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN NEUROSCIENCE

ADVISOR: A. DAVID REDISH, Ph.D.

FEBRUARY 2021

Acknowledgements

First and foremost, I would like to thank Dave for encouraging me to pursue a question I found interesting and helping to make me a better thinker along the way. I would also like to thank past and present members of the Redish lab, especially Brendan for his inexhaustible patience as well as Ayaka, Chris, and Kelsey for their outstanding technical support. Outside of the lab, I would like to acknowledge the University of Minnesota Graduate Program in Neuroscience class of 2015, especially Austin (without whom I would have never left my desk). I am beyond thankful for my family and their unflagging support over the years. Finally, I want to thank my partner, Joy, for her ability to always make me laugh.

Table of Contents

List of Figures

Appendices

— Introduction —

Chapter 1

Anxiety through the ages

The nature of fear and anxiety has been debated for centuries. Despite not having a concrete definition, anxiety has long been suspected to involve some form of prospection, or mental simulation. Seneca, the Roman philosopher and statesman born in the year 4 BCE, observed that “memory brings back the agony of fear while foresight brings it on prematurely” (Seneca, 65 CE). Centuries later, Kierkegaard famously stated in his seminal 1844 treatise *The Concept of Anxiety* that “anxiety is the dizziness of freedom” (Kierkegaard, 1844). The idea that anxiety results from the conflict inherent in choice was elaborated upon at the turn of the century, with existentialist philosophers and psychiatrists considering anxiety as the defining feature of human experience. Indeed, Freud asserted that “there is no question that the problem of anxiety is a nodal point at which the most various and important questions converge, a riddle whose solution would be bound to throw a flood of light on our whole mental existence” (Freud, 1917). Freud came to believe that we repress our socially unacceptable libidinal impulses in order to avoid the anxiety that they cause us to feel. Psychoanalysis ultimately fell out of favor in the scientific community, but William James and Carl Lange, contemporaries of Freud, succeeded in developing a more lasting psychological framework of emotional phenomena. The James-Lange theory posits that emotions are the result of the brain’s interpretation of internal physiological states. The idea that the cognitive interpretation of bodily states generates emotions and affects ongoing reasoning, as opposed to emotions generating bodily states, was a profound insight that has been reformulated by various theorists in recent years.

In the early 1900s, radical forms of behaviorism dominated the brain sciences and constrained discussions of fear and anxiety to strictly observable behavior. Ultimately, cognitivist frameworks overturned decades of behaviorism by seeking to explain behavior with models of cognition, and researchers began to classify anxiety in terms of risk-assessment. Emphasizing the use of ethologically valid foraging paradigms, Blanchard et al. modeled responses to threat as dependent on the perceived proximity to the threat, with anxiety-like behaviors (hesitation and avoidance) being elicited when threats were distant and fear-like behaviors (fight or flight) being elicited when threats were near (Blanchard et al., 1990; Blanchard et al., 1993).

Fanselow and Lester further developed this notion with their ‘Threat Imminence Continuum’ model which characterizes the progression through four states of apprehension that depend on the visibility and proximity of the threat: **(1)** The preferred phase, during which there is no threat; **(2)** the pre-encounter phase, during which the prey is vulnerable to threat (e.g., foraging) but no threat is detected; **(3)** the post-encounter phase, during which a threatening agent is detected but does not pursue the prey; and **(4)** the circa-strike phase, during which the threat actively pursues the prey (Fanselow and Lester, 1988). Mobbs et al. integrated and elaborated on Fanselow’s model with their ‘Survival Optimization System’ in which there are five strategy systems that align with the four stages of Fanselow's Threat Imminence Continuum: **(1)** Prediction strategies, **(2)** threat orienting strategies, **(3)** threat assessment strategies, **(4)** defense strategies, **(5)** prevention strategies (Mobbs et al., 2015). Prediction and prevention strategies correspond to the preferred phase, threat orienting and threat assessment strategies correspond to the post-encounter phase, and defense strategies correspond to circa-strike.

In 1982, Gray argued that conflicts in goal-based decision-making (e.g., approach reward/avoid threat) lead to heightened sensitivity to aversive stimuli and risk-avoidance behavior (Gray, 1982). Furthermore, Gray hypothesized that the hippocampus and the septum play critical roles in fear and anxiety.

In line with these efforts to frame fear and anxiety as distinct but related phenomena, Walker and Davis argued that the amygdala and the bed nucleus of the stria terminalis (BNST) regulate phasic and sustained fear, respectively. Specifically, they assert that basolateral amygdala (BLA) projections to one division of the amygdala (the central nucleus) drives 'phasic fear' via short-term hypothalamic and brainstem activation. Similarly, inputs to the lateral BNST from another division of the amygdala (the lateral nucleus) as well as the BLA corresponds to 'sustained fear' (i.e., anxiety) through prolonged activation of hypothalamic and brainstem nuclei (Walker, 2008).

LeDoux has similarly argued for two dissociable threat-related systems: a fast, amygdala-mediated nonconscious threat-response system for generating defensive behaviors and a slow, hippocampo-cortical-mediated cognitive system for generating the subjective experience of fear and anxiety (LeDoux, 2015). LeDoux argues that the fast system has been incorrectly taken as a proxy for the conscious feeling of fear or anxiety, and that there is presently little evidence that what we see in non-human animals is consciously experienced emotion as opposed to simply the release of non-conscious defensive behaviors (LeDoux, 2015).

The neuroanatomy of fear conditioning (i.e., the fast threat-response circuit) is well characterized (LeDoux, 2000). In the case of Pavlovian fear conditioning, information regarding the conditioned stimulus (CS) (i.e., a light or an auditory cue) projects onto the lateral amygdala (LA), which also receives incoming information about the unconditioned stimulus (US) (i.e., an electric shock). The LA in turn projects to the central nucleus of the amygdala (CeA), which activates structures involved in behavioral inhibition (e.g., via the periaqueductal gray), autonomic regulation (e.g., via the lateral hypothalamus), and hormone secretion (e.g., via the paraventricular nucleus of the hypothalamus). The concerted effort of these downstream targets generates a conditioned response (CR) (e.g., freezing, fleeing, or fighting). Importantly, the coincident discharge of these two sensory inputs onto the LA induces synaptic remodeling (via an intracellular signaling pathway; LeDoux, 2007) such that the experience of the CS in isolation is sufficient to elicit the CR (LeDoux, 2000).

Unlike the fast threat-response system, the mechanisms underpinning the cognitive system remain unclear. Traditional models have assumed that emotions are discrete entities that are identifiable by a set of fundamental attributes. This framework holds that there are core emotions with defined physiological and behavioral signatures, and that the circuitry that mediates these emotional states is modular and conserved. This view, a type of essentialism that sees emotions as natural kinds, has been challenged by

Barrett who has argued for a constructionist view of emotions. Barrett argues that the brain relies on predictions (learned from past experience) to categorize internal and external sensory inputs, and that this sensory categorization actively constructs emotion (Barrett, 2017). Contrary to the natural kinds view of emotion, this sensory categorization process is distributed, dynamic, and varies depending upon an agent's goals and history.

In a variation on the embodiment view espoused by James and Lange, Damasio has argued that somatic states become associated through experience with event outcomes, thus allowing these 'somatic markers', a type of emotional signal, to gain predictive value. Later, somatic markers are elicited by primary inducers (an innate or learned stimulus in the immediate environment) or secondary inducers (recall of a primary inducer from memory). Damasio argues that somatic markers are advantageous because they simplify complex decisions by reducing time spent deliberating between available options; instead, these body-state signals bias decision-making and facilitate rapid and decisive action (Damasio, 2005).

Much of the modern work exploring the neural basis of threat-processing has been focused on elucidating which neural circuits map onto and generate specific threat-response behaviors. In particular, optogenetic techniques have proven to be a popular method for isolating specific cell populations and exciting or inhibiting select terminals to uncover their role in threat-responsivity. This line of research has been highly informative. However, these tools, while effective at dissecting the circuit underpinnings of the fast 'somatic' threat-response system (e.g., freeze, fight, flight), are not by themselves sufficient to give us access to the ‘cognitive’ dimension of fear and anxiety systems.

Large-scale population activity in freely behaving animals provides a window into the representational dynamics of fear and anxiety in a way that circuit stimulation cannot. Specifically, computational methods allow us to decode neural activity and relate it to mental states. This approach will allow us to not only account for the neural correlates of defensive behaviors produced by the threat-response system but also to explore the cognitive dimension of fear and anxiety. In this dissertation, I will argue that in order to expose fear and anxiety networks to neuroscientific inquiry we must track information as it flows through behaviorally-relevant neural circuits; this approach requires that we turn our focus to neural algorithms, a computational scale found at the intersection of circuit dynamics and neural representations.

A computational approach: defining neural algorithms

Recent efforts to integrate a computational perspective into psychology, neuroscience, and psychiatry have unlocked novel approaches to modeling and treating the nervous system. Here, a computational perspective simply refers to characterizing how neural structures represent and process information. In adopting a computational approach to understanding the nervous system and behavior we are primarily concerned with three levels of analysis: **(1)** characterizing how information is represented (e.g., which parameters of a given task can be decoded from the neural activity in a given structure?), **(2)** characterizing how information is processed over time (e.g., what are the algorithms employed to address the task demands?), and **(3)** merging the preceding two levels of analysis to explain behavior.

The steps taken by an information processing system to solve a problem are collectively called an algorithm. Thus, the representation (both within and distributed across nodes) and transmission (between nodes) of information in the nervous system in response to a given array of inputs, both internal and external, constitutes a neural algorithm. In the following sections, we will review three fear- and anxiety-related neural algorithms. That is, we will trace how information is represented within and across structures as well as how that information flows through neural circuits during threat-processing.

Moving toward a computational understanding of mental illness

Anxiety disorders are heterogeneous, and their heterogeneity highlights that different anxiety tasks involve unique cues, contexts, and cognitive demands that elicit distinct patterns of network activity. There is no single cure for cancer because cancer is not a unitary disease — it is a complex group of diseases, each with its own set of underlying causes. As such, treatments are tailored on a patient-by-patient basis. Similarly, there likely can be no single cure for anxiety. Classifying these different anxious states and the physiology that governs them is critical to making strides toward understanding and treating anxiety disorders and the specific biological substrates that each disorder uniquely exploits. By understanding the anatomical and neurophysiological properties of these various anxious states, we can adjust our therapies to more selectively and effectively target them in instances of disease.

A corollary of having multiple threat-processing systems is that there are multiple ways for those systems to fail. Having a map of the potential failure modes (i.e., the vulnerable components and processes) of a system is not only a valuable diagnostic tool but also a starting point for identifying effective remedies (Walters and Redish, 2018). By seeing anxiety disorders as resulting from failure modes in threat-processing systems that result in cognitive distortions, we can shed light on the latent variables driving cognitive pathology. It is to this approach that we will be turning next.

— Theoretical perspective —

Chapter 2

Computational modeling in psychiatry

The concept of computational psychiatry derives from the more general field of computational neuroscience which explores how the nervous system represents and processes information to guide adaptive behavior. Breakthroughs in neuroscience over the last several decades have elucidated how these computations work both in terms of the processes themselves and of the neural circuits involved in those processes (Dayan et al., 2001; Redish, 2013). Computational psychiatry entails applying reliability engineering techniques to those brain information processing systems – if we understand how the system works, we can identify its "vulnerabilities" and tailor treatment to address those vulnerabilities (Knill and Pouget, 2004; Redish et al., 2008; Montague et al., 2012; Huys et al., 2016; MacDonald et al., 2016; Moutoussis et al., 2018).

With this paradigm shift, psychopathology can now be understood as a failure of various brain information processing systems to generate an adaptive response to dynamic environmental contingencies. It is important to recognize that this failure lies in the interaction between the environment and the individual – an individual susceptible to cocaine addiction who never tries cocaine never becomes a cocaine addict. Moreover, psychiatric symptoms depend on complex feedback loops between neural information processing and the environment – for example, excessive anxiety can produce insomnia, which produces fatigue, which produces an inability to provide the self-control to reduce anxiety.

Chapter modified from:
Walters, C. J., Vinogradov, S., Redish, A. D. (2020), Computational Modeling in Psychiatry, in 'The Cambridge Handbook of Computational Cognitive Sciences', (R. Sun, ed). Cambridge University Press, Chapter 24. (In press)

1 Approaches to psychiatry

The field of computational psychiatry is often described as including both theory-driven and data-driven approaches (Huys et al., 2016). Theory-driven approaches are like those described above: one can derive “failure modes” or “vulnerabilities” from a theory-driven understanding of the underlying information processing (Redish et al., 2008; MacDonald et al., 2016) and design treatment options that target or bypass those vulnerabilities. With sufficient recognition of those potential vulnerabilities, one could even engineer tests that can provide warning signs and allow treatments that prevent them from becoming active in the first place. In contrast, data-driven approaches use unsupervised learning techniques to identify clusters of behaviors that co-occur (Huys et al., 2016). Historically, the DSM-III was built on this model, in which the authors attempted to find symptom clusters from surveys and interviews with practicing psychiatrists (Lieberman, 2015). While big-data approaches are still being touted and tried (Borsboom et al., 2019), I argue that the major breakthroughs that have occurred within the field of computational psychiatry so far have been from the theory-driven side (see Appendix A for an overview of computational modeling), and thus I will focus on them in this chapter.

Behavior arises from a complex interaction of genetics, biochemistry, and the environment, which includes (because humans are social animals) our social interactions; however, all of those underlying causes are translated through the brain and its interaction with the environment (Fig. 1). This means that we can conceptualize the brain’s information operations as the key step in translating underlying causes (genetics, biochemistry, the physical and social environment, etc.) to adaptive or maladaptive behavior, including psychopathology. These computational processes are implemented through complex neural (and hormonal and glial) networks, and understanding the interaction between these processes and the environment leads to a recognition of where that interaction can fail and the vulnerabilities within these complex networks. In the sections below, we will review seven cases that highlight how a theoretical approach to neural information processing can be applied to psychiatric phenomena: addiction, psychosis, depression, obsessive-compulsive disorder, autism spectrum disorder, attention-deficit hyperactivity disorder, and anxiety disorders.

2 Addiction

Addiction is broadly defined as an inability to stop engaging in a behavior despite negative consequences. This takes many forms: gambling, alcoholism, smoking, shopping, drugs, video games – almost any rewarding behavior can become inelastic to social and financial costs as well as physical and psychological harm. In addition to an inability to stop the behavior, people with an addiction often experience cravings or withdrawal as well as an escalation of their addictive behavior over time often with evidence of sensitization (e.g., taking larger and larger doses of the drug or going on longer or more expensive gambling sprees).

2.1 Temporal-delay reinforcement learning models

Current computational models of addiction are generally based on reinforcement learning (RL) models in which a decision-making agent performs actions and receives environmental feedback. In RL models, the world communicates to the agent by providing observations and rewards (which can be positive rewards or negative punishments/costs), and the agent communicates actions back to the world, which have the effect of changing the state of the world. The first RL model of addiction with computational simulations is that of Redish (2004), in which drugs of abuse are assumed to modify parameters of what is now referred to as a "model-free temporal difference reinforcement learning (TDRL)" model. In this classic TDRL model, value is defined as the amount of expected future reward given a decision policy (Sutton et al., 1998), taking an action in any given state of the world is associated with an expected value, and that value is learned through "temporal difference reinforcement learning". If the agent takes an action and finds more (or less) reward than expected, then the agent increases (or decreases) the stored value of taking that action in that environment. This difference is known as the "reward prediction error" or "value prediction error", and there is evidence that some aspects of dopamine signaling carry this value prediction error signal (Schultz et al., 1997). When the agent leaves one state (S_1) and enters another (S_2), we define the value prediction error (δ) to be:

$$\delta(t) \ = \ \gamma^{d} \left[R(S_2) \ + \ V(S_2)\right] \ - \ V(S_1)$$

where d is the delay spent in state S_1, $R(S_2)$ is the reward received in state S_2, and V is the value of a given state. $R(S_2) + V(S_2)$ is discounted by γ^{d}(so that the larger the temporal distance between S_1 and S_2, the smaller $R(S_2) + V(S_2)$ becomes). The value of state S_1 is then adjusted by δ such that if the observed value in state S_2 is better ($\delta > 0$) or worse ($\delta < 0$) than expected, the agent will increase or decrease the stored value of S_1, respectively. Eventually, $V(S_1)$ will approach $\gamma^{d}[R(S_2) + V(S_2)]$ and δ will approach 0. Redish (2004) noted that given the evidence that many drugs of abuse produce dopamine neuropharmacologically, one can model the effect of these drugs as a non-compensable δ signal:

$$\delta(t) = max\{\gamma^{d}[R(S_2) + V(S_2)] - (S_1) + D(S_2), D(S_2)\}$$

where $D(S_2)$ is the neuropharmacological effect of receiving the drug upon entering state S_2. Through simulations, Redish (2004) found that the agent would develop preferences for drug-taking, preferences for drug-seeking, and would become increasingly inelastic with continued drug use.

The Redish (2004) model can serve as an introduction to the concept of failure modes. According to these RL computational models of decision-making, the brain evolved to use dopamine as a learning signal driving the recognition of future value. A chemical that bypasses the normal function of dopamine as a value prediction error signal (δ) provides a signal that is interpreted by the rest of the brain as always being "better than expected" and driving an increased willingness to take the action that led to drug use,no matter how pleasant or rewarding it actually was. This is a vulnerability in the brain's reinforcement learning processes.

An important issue in action-selection models is that there is now very strong evidence that decision-making arises from multiple algorithms (Kahneman, 2011; Redish, 2013), each of which has different vulnerabilities. For example,the incentive-sensitization theory of addiction (Robinson and Berridge, 2001) distinguishes between pleasure (liking or craving, encoded in endogenous opiate signals and vulnerable to exogenous drugs of abuse like morphine, heroin, oxycodone) and value (wanting or incentive salience, encoded in endogenous dopamine signals and vulnerable to exogenous drugs of abuse like cocaine and amphetamine). Robinson and Berridge (2001) suggest that these two aspects are dissociable and can change independently of one another (e.g., an increase in wanting with a decrease in liking, a common phenomenon in addiction).

While RL models of addiction like those described above are all based on positive (reinforcing) outcomes, addiction likely has a darker side as well in which drug-seeking becomes a means of escaping negative affective states (anxiety, depression, anhedonia, social isolation) which can result from withdrawal and other effects of drug taking (Koob and Volkow, 2010). These components are included in other models of addiction, such as pharmacological homeostatic models (Tsibulskyand Norman, 1999) and opponent process models (Koob and Volkow, 2010). For a more complete review of computational models of addiction, see Walters and Redish (2018).

This multi-vulnerability model has important consequences for both our understanding of psychiatric phenomena and treatment. It suggests that symptom clusters (such as addiction and drug-seeking) reflect processes that are equifinal (multiple causes are capable of generating a given outcome) and multifinal (similar initial conditions are capable of generating a variety of outcomes). It also suggests that treatment should address the underlying impairments rather than the symptom clusters (Redish et al., 2008; Friston et al., 2014; Redish and Gordon, 2016). I will return to this discussion at the end of the chapter (see section 9).

3 Psychosis

Schizophrenia is a heterogeneous psychiatric disorder characterized by three kinds of symptom clusters: positive symptoms (hallucinations and delusions), negative symptoms (blunted affect, reduced speech, and social withdrawal), and cognitive symptoms (impairments in processing speed, working memory, executive function, and social cognition).

An individual’s first psychotic episode is often preceded by a prodromal phase which can last anywhere from weeks to years during which they progressively exhibit symptoms such as depression, suspiciousness, magical thinking, and social isolation. This period then culminates in a psychotic episode, known as the acute phase, during which some combination of the above symptoms are exhibited. The acute phase is generally followed by treatment and a degree of recovery, with variable periods of time separating episodes of acute psychosis.

3.1 Basin of attraction models

Neurophysiological theories suggest that cortical systems carry information about the world – where information is defined mathematically as the degree to which knowing something about the state of one system (e.g., a neuron's firing rate) reduces your uncertainty about the state of another system (e.g., a visual stimulus) (Shannon, 1948) – by categorizing stimuli into “basins of attraction”, a concept from dynamical systems theory (Hertz et al., 1991). In these models, both perception and memory are encoded as specific firing patterns across a population of neurons. Computational models of these networks have shown that appropriate connection structures will recover remembered patterns from noisy or partial patterns of activity (Hebb, 1957; Hopfield, 1982; Hertz et al., 1991).

This phenomenon – called an attractor state – is a mathematical description of pattern completion wherein a remembered pattern is retrieved from noisy or partial cues. The set of points in this n-dimensional space that flow into a stored state is called a "basin of attraction". One can imagine this process of pattern completion as a ball falling down into a valley, with higher network energy corresponding to greater potential energy of the ball on the energy landscape; thus, the system evolves toward a state that minimizes the network energy. In perception, this process produces categorization whereby similar patterns (e.g., the many shades of blue) can flow into a single pattern and become recognized as part of that category (e.g., blue). In memory, this process implements content-addressable memory whereby retrieving part of a memory results in the memory being recalled in full. Attractor dynamics depend on the depth of the basin, where deeper basins occur with stronger synapses (which produce a stronger vector field), while shallower basins are more sensitive to noise and thus more susceptible to small changes in input (Seamans and Yang, 2004).

Attractor models can provide valuable insights into the biological dynamics underlying psychosis. In a study using a recurrent integrate-and-fire biophysical network model, Loh et al. (2007) found that a decrease in NMDA conductance not only reduced firing rates of neurons in a stable network state but also resulted in a failure to maintain a persistent network pattern. They argue that this shift in dynamics could relate to negative symptoms (e.g., blunted affect which is thought to result from reduced activity in orbitofrontal and anterior cingulate cortices) and cognitive symptoms (e.g., working memory deficits which are thought to result from temporal instability in prefrontal circuits), both of which often appear together and precede the exhibition of positive symptoms. Furthermore, they found that decreasing both NMDA and GABA conductance resulted in a failure to maintain both an immediate and a persistent network pattern, thus giving rise to spontaneous jumps between attractors, a finding consistent with experimental evidence showing that disrupting NMDA-receptor activity disrupts spike timing and decouple prefrontal circuits in non-human primate models of schizophrenia (Zick et al., 2018). This effectively makes the network less resilient to stochastic neural activity and as a result liable to meander from basin to basin.

Going beyond attractor models, abnormalities in the neurotransmission systems that regulate synaptic gain (e.g., NMDA-R function, dopamine, and acetylcholine) are a common focus in other models of psychosis, such as Bayesian models (Adams et al., 2013). Bayesian models allow for the incorporation of new observations (the likelihood) with established knowledge (the prior) in order to continuously infer the probable cause of new observations:

$$p(cause \mid observation) \propto p(observation \mid cause)\; p(cause)$$

which is more concisely denoted as:

$$posterior \propto likelihood \cdot prior$$

where the posterior distribution is simply the updated expectation after making an observation. Thus, the posterior at one time step becomes the prior at the next time step, with the aim being to continuously update expectations (i.e., beliefs) so that they predict new observations with increasing accuracy. Given that the posterior, the likelihood, and the prior are all probability distributions, the width of the prior and the posterior reflect belief uncertainty and the width of the likelihood reflects the observation (or stimulus) noise. Additionally, the difference between the prior and the likelihood corresponds to the prediction error (i.e., the surprise), and the difference between the prior and the posterior can be thought of as the information gained, or, more precisely, how much the belief changes to fit the new observation.

A helpful mathematical reframing of Bayes theorem is that it is describing how to best update beliefs about the world when new observations deviate from expectations (i.e., the prior). This deviation from expectation is the prediction error mentioned above, but in Bayesian models these prediction errors are weighted in accordance with the number of observations that have been made (Adams et al., 2013; Mathys et al., 2016). For example, the prior is more precise when it is based on more observations, thus the weight placed on the prediction error is inversely proportional to the precision of the prior. This means that if the prior is highly precise as the result of many observations having been made, then a new observation that drastically deviates from that prior expectation will not result in a large belief change. Bayes theorem is therefore mathematically equivalent to a precision-weighted prediction error (Mathys et al., 2016):

$$new\ belief \propto old\ belief \quad + \quad weight(prediction\ error)$$

Bayesian accounts of psychosis hold that schizophrenic symptoms result from faulty Bayesian inference. According to these models, psychosis is driven by inaccuracies in beliefs (i.e., priors) and the confidence in those beliefs (i.e., the precision, or the inverse variance, of the prior) (Adams et al., 2013). Confidence in this context is a direct function of synaptic gain in neurons signaling surprise, where discrepancies between predictions (priors) and sensory data (likelihood) drive Bayesian belief updating. Psychotic symptoms can be understood in terms of an imbalance in synaptic gain, much in the same way as the basins of attraction model discussed above.

4 Depression

Major Depressive Disorder (MDD) is a mood disorder characterized by persistent feelings of dysphoria, fatigue, helplessness, hopelessness, and loss of interest and pleasure. Individuals suffering from MDD commonly have somatic symptoms that include changes in sleep patterns (often with difficulty sleeping), changes in appetite, and lethargy or agitation. Additionally, people with MDD may experience suicidal ideation and behavior.

4.1 Decision-theoretic models

Many theories assert that the brain represents a model of its environment, and that this model can be thought of as a set of beliefs (i.e., predictions) about the structure of the world and the likely causes of sensory observations (Huang and Rao, 2011). The manner by which these beliefs get updated in light of new sensory evidence can be described as form of Bayesian inference (see section 3 for more on Bayesian inference):

$$\Delta belief \propto precision \cdot prediction\ error$$

where Δbelief is the degree to which the agent updates its belief; precision is the certainty, or inverse variance, of the prior belief; and prediction error, or surprise, is the difference between the prior belief and the new sensory observation.

Chekroud (2015) proposed a framework in which depression is viewed through the lens of the free energy principle, a cognitive framework which, in the context of perception, asserts that the brain represents a model of the environment in order to infer the causes of sense data and minimize surprise (mathematically, free energy), where surprise simply means unexpected states, via sensory prediction errors (i.e., the disagreement between the model's predictions and the inputs it receives) (Friston, 2010).

Importantly, there are two ways an agent can minimize prediction error: they can change their model to fit the environment or they can change the environment to fit their model. Chekroud argues that depression results from a set of depressive beliefs (owing to aberrant neural information processing) that are immune to countervailing evidence; therefore, an individual with a depressive model of the world behaves in a way that reinforces their depressive model (e.g., by not engaging in rewarding behaviors) as opposed to altering the model itself, thus resulting in a self-reinforcing feedback loop. It is worth noting that this cyclic notion of an individual's actions reinforcing their psychopathology is likely true of other psychiatric conditions (e.g., anxiety and obsessive-compulsive disorders).

Others have used decision-theoretic approaches to explore the nature of these depressive models of the world. It has been suggested that many depressive symptoms (e.g., anergia) can be explained as the result of pessimistic evaluations of the future where predicted utility is consistently low (Huys et al., 2015). This dovetails with another symptom of clinical depression, learned helplessness, in which patients feel that their actions have no impact on the outcomes they experience in the world, thus they resign to a state of inaction and exhibit signs of indifference and lethargy in the face of adversity (Seligman, 1972).

Within this context, rumination (the consideration of alternative past and potential future events), which is commonly seen in depression, entails search processes through a potentially very large transition function T:

$$T: s_{a,t} \rightarrow\ p(\widehat{s}_{t+1})\ \forall\ s \in S,\ a \in A$$

where T is a matrix of all transition probabilities between an initial states at timestep t and any other state $\widehat{s}$ at timestep t+1 in the set of all possible states S after taking an action a from the set of available actions A. Rumination can be interpreted as exploration of the possible paths in a POMDP state space. Models of depression have suggested that the over-rumination seen in depression may be a pathological extension of a normal re-evaluation and reconsideration process evolved to determine useful paths within a large and potentially unknown state space.

Indeed, a modeling study found that the extent to which one prunes the mental search tree of possible future states correlates with sub-clinical symptoms of depression (Huys et al., 2012). This suggests that non-depressed individuals underexplore aversive prospects while individuals with depression will overexplore negative prospects. Huys et al. (2012) interpret these findings in the context of a theoretical model of serotonin, supported by some experimental evidence suggesting that behavioral inhibition in the context of threat prediction may be mediated by serotonergic activity (Dayan and Huys, 2009), which posits that serotonin curtails the contemplation of aversive outcomes. Given that some forms of depression are characterized by reduced serotonergic activity and that patients with depression benefit from medications that increase serotonergic neurotransimission, this framework suggests that the result of such an imbalance could be an inability to prune the mental search tree, thus leading to an increased consideration of negative outcomes.

Anhedonia, another hallmark symptom of depression, is characterized by a reduction in motivation and the enjoyment of formerly rewarding stimuli. Two possible causes have been suggested: disrupted reward learning or decreased sensitivity to reward itself (Huys et al., 2013). Some data suggest that aberrant prediction error signaling may underlie anhedonia (Gradin et al., 2011) while reward sensitivity to positive and negative outcomes might be modulated by serotonin (Seymour et al., 2012). Attempts to sharpen the distinction between these two hypotheses, most commonly in the language of opponent-processes attempting to make sense of the functional interplay between serotonin and dopamine, have not been conclusive (Daw et al., 2002); however, MDD is a heterogeneous condition and abnormalities in reward learning and action-selection are only two of the many symptomatic factors which might manifest in a patient.

5 Obsessive-compulsive disorder, tics, and Tourette's syndrome

Obsessive-compulsive disorder (OCD) is a psychiatric condition characterized by obsessive thoughts (e.g., a preoccupation with a perceived threat such as germs) that cause negative affect and repetitive, ritualized behaviors (e.g., excessive hand washing) which are thought to provide (temporary) relief from the distressing obsessions (Dougherty et al., 2018).

Tourette's syndrome is a related but distinct neurological condition in which individuals exhibit tics – spontaneous and repetitive movements or vocalizations (e.g., facial twitches, eye-blinking, humming, throat clearing, etc.) which can escalate in complexity over time (Swain et al., 2007).

5.1 Models of habit and sequence learning

That action-selection is not mediated by a unitary system has been a long-held view in psychology and neuroscience (O'Keefe and Nadel, 1978; Kahneman, 2011; Redish, 2013), with evidence pointing to there being non-overlapping neural systems underpinning at least two differentiable mode of action-selection (Scoville and Milner, 1957). Procedural processes encompass the largely automated habit system while declarative processes refer to the more episodic goal-directed system. Operationally, habitual behavior can be said to be insensitive to changes in contingency, such as outcome devaluation, while goal-directed behavior is defined by its flexibility in response to novel circumstances and environmental rules. In non-human animals, the habit system has been labeled the stimulus-response (S-R) system while the declarative system has been labeled the action-outcome (A-O) system (Adams and Dickinson, 1981).

This distinction is further supported at the level of anatomy, with individuals suffering from medial temporal lobe damage exhibiting impairments in the declarative system while maintaining a functioning procedural system (Scoville and Milner, 1957) and damage to the basal ganglia disrupting procedural function and leaving declarative abilities intact (Saint-Cyr et al., 1995). Similarly, in non-human animals, lesioning the basal ganglia impairs habit-like S-R learning (O'Keefe and Nadel, 1978; Saint-Cyr et al., 1995; Redish, 1999, 2013) while behaviors involving goal-directed A-O planning require the hippocampus (O'Keefe and Nadel, 1978; Redish, 1999, 2013, 2016).

There is now considerable evidence implicating dysfunction in the cortico-basal ganglia-thalamo-cortical (CBGTC) loop, a critical circuit in the habit system, in OCD. Key hubs in this network include the orbitofrontal, anterior cingulate, and medial prefrontal cortices as well as the caudate nucleus (Graybiel and Rauch, 2000). Individuals with lesions to the striatum (or its downstream target the pallidum), for example, show signs of obsessions, compulsions, and stereotyped behaviors reminiscent of OCD (Laplane et al.,1989).

While obsessions and compulsions are often co-expressed, there is some evidence suggesting that they might be developmentally dissociable (Freeman et al., 2012). Furthermore, individuals with OCD display signs of impaired goal-directed planning and an over-reliance on habitual heuristics in a variety of tasks with no indication of the presence of obsessions (Gillan et al., 2011). Though it has been commonly thought that obsessions instigate compulsions, these and other data have led to the supposition that this causal relationship might in fact run in the other direction, with compulsions being the primary feature of OCD which precede obsessions (Gillan et al., 2011). In this ‘COD’ model, compulsions are viewed as being egodystonic, meaning they generate behaviors that are in conflict with one’s self-image. This results in cognitive dissonance, and obsessions are posited as confabulatory reactions attempting to rationalize that mismatch (e.g., I feel the urge to wash my hands therefore I must be worried about germs, as opposed to I am worried about germs therefore I feel the need to wash my hands) (Gillan and Robbins, 2014). In support of this COD model, confabulation has been shown to be a key factor in dealing with dysfunction (Gazzaniga et al., 1965; Ramachandran et al.,1998).

Neural network models consisting of coupled excitatory and inhibitory units have been shown to recapitulate many of the defining features of OCD when the E-I balance is disrupted (specifically when the inhibition parameter is reduced) (Verduzco-Flores et al., 2012). Maia and McClelland (2012) underscore how this parameter change is likely equivalent to the levels of network excitation increasing, which is consistent with prior modeling work (Rolls et al., 2008) showing that glutamatergic hyperactivity generates deeper basins of attraction which could be the cause of the tenacious habitual responses characteristic of OCD (see section 3 for more on attractors). However, unlike point attractors which stabilize around a set pattern of activity, the Verduzco-Flores model captures attractor dynamics that cycle through stereotyped sequences of activity, a property which more closely resembles the motor and thought sequences experienced by those with OCD. Sequence learning has been a long-standing problem in psychology and cognitive science (Lashley, 1951). While previous theoretical and experimental efforts have underscored the role of the basal ganglia in sequence production (Berns and Sejnowski, 1998; Graybiel, 1995), they have not explored how sequences could become pathologically expressed in conditions like OCD.

While OCD and Tourette's syndrome are both behaviorally and neurologically similar, as well as highly comorbid, the two conditions are dissociable (George et al., 1993). Anatomically, evidence implicates the degeneration of parvalbumin-containing neurons in the striatum and pallidum in Tourette's syndrome (Kalanithi et al., 2005), two structures often compromised in OCD. Functional magnetic resonance imaging (fMRI) data has shown that volitional suppression of tics correlates with an increased fMRI BOLD signal in the caudate nucleus and prefrontal cortex and a decreased signal in the putamen and pallidum relative to BOLD activity observed during the free expression of vocal or motor tics (Peterson et al., 1998).

Tic disorders and Tourette's syndrome may result from aberrantly reinforced motor behaviors (Maia and Conceicao, 2017). As in OCD, individuals withTourette's syndrome often report an escalating sense of discomfort leading up to tic expression known as a premonitory urge, and this discomfort is often dissipated by expression of the tic. A recent model of premonitory urges argues that sensory signals originating in structures like the somatosensory cortex get projected to cortical regions such as the insula, and that the resulting aversive sensations are successfully terminated by tic execution (Conceicao et al., 2017). This generates a positive prediction error (conveyed via phasic dopamine) which then reinforces the tic via the CBGTC loop (Conceicao et al., 2017). Other models suggest that elevated levels of tonic striatal dopamine (or changes in striatal dopamine receptor density or sensitivity) result in hyperactivity in the direct GO pathway in the CBGTC loop, thus amplifying the expression of motor and vocal tics (Maia and Frank, 2011). This is consistent with the efficacy of D1 receptor antagonists in suppressing tics in individuals with Tourette’s (Gilbert et al., 2014) and the ability of D1 receptor agonists to cause spontaneous tic-like motor behaviors (Bergstrom et al., 1987).

6 Autism spectrum disorders

The autism spectrum refers to a continuum of neurodevelopmental disorders associated with impaired social communication, a preference for sameness, and sensory hypersensitivity. Individuals with autism often exhibit a narrow range of interests (e.g., an intense preoccupation with a specific topic) and repetitive behaviors (e.g., rocking or repeating certain words or phrases).

6.1 Bayesian observer models

As mentioned above (see section 4), Bayesian models assert that the brain weighs bottom-up sensory information (the likelihood) using an internal predictive model of the environment in the form of top-down expectations (the priors). This operation serves the purpose of inferring the probable cause of a given sensory state using prior knowledge of how the world works to form a percept (the posterior), and is thought to be implemented by hierarchical prediction error signaling wherein higher order brain areas compare their predictions against incoming sensory information from lower order brain areas (Van Boxtel and Lu, 2013).

This model, known as the Bayesian brain hypothesis, posits a fundamental trade-off between having a veridical representation of the external world (weak priors, which is equivalent to overweighting the likelihood) and the ability to extract statistical patterns from experience and skew perception in line with those expectations (strong priors). Individuals on the autism spectrum appear to have attenuated priors (i.e., abnormal internal predictive models of the environment) which results in incoming sensory information being less heavily weighted by top-down expectations (Pellicano and Burr, 2012).

Impaired priors results in perception being more accurate in the sense that the trial-by-trial variability of sensory experience is not smoothed out and biased toward the mean of those experiences (as is the case in non-autistic individuals). Instead, the hypersensitivity to fluctuations in sensory information characteristic of autism is akin to overfitting noisy data. This model furnishes an explanation for a variety of non-social symptoms observed in individuals on the autism spectrum. For example, people with autism are often overwhelmed by certain sensory stimuli (such as loud sounds or being touched) and are resistant to change in their environment – an inability to leverage past experience (via priors) in order to generalize and respond adaptively to novel stimuli would make the world confusing and unpredictable. This model predicts that the near-constant feeling of being overwhelmed by novel sensory information (hypersensitivity) leads to a preference for routine (which minimizes exposure to novel scenarios). In support of this model, experimental evidence shows a reduction in the amount of temporally correlated mutual information (a measure of representational stability over time) in the hippocampus of individuals with autism (Gómez et al., 2014), suggesting impairments in top-down processing in individuals with autism consistent with the notion of weak priors.

A cognitive framework consistent with the Bayesian brain model of autism is known as the weak central coherence theory (Frith, 2003; Happé and Frith, 2006). This theory posits that while non-autistic individuals have an innate perceptual bias towards Gestalt perceptions (privileging the coherent whole over its constituent parts), autism is characterized by an anti-Gestalt perceptual bias (a bias toward perceiving local features at the expense of global properties) (Frith, 2003). There is a considerable body of experimental evidence in favor of the weak coherence account with a variety of neurobiological mechanisms having been proposed (Happé and Frith, 2006).

The model of weak priors in autism does not, however, make much headway in explaining the social and emotional dysfunctions experienced by those with autism. These symptoms have been suggested to be a result of abnormalities in interoception, the ability to detect sensations from the body and viscera (heart rate, chemoreceptors, respiration, gastrointestinal tract, etc.) and interpret those physiological signals as feeling states (hunger, anxiety, excitement, etc.). Garfinkel et al. (2016) argue that there are several dimensions to interoception, two of which are accuracy (objective ability to detect bodily states) and sensibility (one's belief about one's accuracy), and that individuals with autism exhibit poor interoceptive accuracy and high interoceptive sensibility. This complements embodied theories of social cognition and attachment which suggest that we mentally simulate the emotional state of others in order to empathize with them (Niedenthal, 2007). These and other data suggest that impairments in interpreting one's own interoceptive states could drastically impair one's ability to infer the emotional states of others (Friston et al., 2014).

7 Attention-deficit hyperactivity disorder

Attention-deficit hyperactivity disorder (ADHD) is characterized by extreme difficulty sustaining attention during conversation or any task requiring persistent mental effort. Individuals with ADHD often exhibit signs of restlessness, poor concentration, and distractibility (e.g., fidgeting) and can be highly disorganized (e.g., regularly losing personal items) or display impulsive behavior.

7.1 Normalization models

Agents must arbitrate between stable behavior, exploiting what they currently know about the environment to maximize value, and unstable behavior, exploring potentially less fruitful alternatives in order to gain new information. The brain, then, is confronted with this explore-exploit dilemma and needs to strike a balance between these two competing strategies (Daw et al., 2006). Hauser et al. (2016) frame ADHD in relation to this trade-off, arguing that ADHD biases an agent toward more exploratory (i.e., information gathering) behavior at the cost of stability (i.e., exploitation), a policy that can be advantageous in highly uncertain environments.

Hauser et al. (2016) model attention in terms of neural gain by building on a standard softmax model of exploration versus exploitation (Sutton et al.,1998; Williams and Dayan, 2005) which uses a sigmoid function that takes an input signal and either amplifies or dampens the probability of taking an action given that signal:

$$f_G(x) = \frac{1}{1+e^{-Gx+B}}$$

where G is the gain parameter and B is a bias term that allows the equation to shift the sigmoid along the horizontal axis. Hauser et al. (2016) then relate this more general principle of neural gain, which dictates sensitivity to incoming signals, to action-selection and choice stochasticity. They do this by employing a variant of the softmax decision function wherein the value of performing a given action is weighted relative to the value of performing all other available actions (Williams and Dayan, 2005):

$$p(a_i) = \frac{e(\frac{a_i}{\tau})}{\sum_{k=1}^{N} e(\frac{a_k}{\tau})}$$

where $p(a_i)$ is the probability of taking action i, a_i denotes the value of action i, a_k is a vector of the value of all N possible actions, and τ is the decision temperature. What this softmax function does in practice is to convert the value associated with a set of actions into probabilities of taking those actions. A low τ is equivalent to the neural gain being high and choice being more exploitative while a high τ is equivalent to the neural gain being low and choice being more exploratory.

Indeed, there are now converging lines of evidence that attentional computations involve some form of normalization (Lee et al., 1999; Reynolds and Heeger, 2009; Schmitz and Duncan, 2018). The neural gain model of ADHD thus provides a comprehensive perspective which first outlines the computational problem (the explore-exploit trade-off), characterizes an algorithm that can model the phenomenon of interest (neural gain), and links the algorithm to a biological mechanism (catecholaminergic tone in the striatum). This framing is consistent with other efforts to relate ADHD symptomatology to variations in decision temperature (Williams and Dayan, 2005) as well as experimental findings from individuals with ADHD (Hauser et al., 2014).

Both modeling (Frank et al., 2007) and experimental evidence (Tripp and Wickens, 2008) support ADHD as a condition of low neural gain (i.e., increased decision temperature) owing to impaired catecholaminergic signalling (i.e., dopaminergic or noradrenergic neurotransmission). This decreases the neural signal-to-noise ratio between competing actions, making attention unstable and behavior more stochastic. This idea is consistent with a long-standing theory which posits that ADHD results from an impairment in behavioral inhibition and excessive impulsiveness (Sagvolden and Sergeant, 1998). The notion that ADHD is associated with a hypersensitivity to delayed rewards is supported by data showing excessive discounting of future outcomes in individuals with ADHD (Tripp and Wickens, 2008). This cognitive model of excessive delay discounting is in agreement with the dopaminergic account described above given modeling data which suggests that low levels of dopamine in the ventral striatum decreases motivation to pursue distal rewards (Smith et al., 2006).

8 Anxiety disorders

It is important to distinguish between fear and anxiety, as they are separate emotional states with distinct behavioral correlates (Mobbs et al., 2007; Blanchard and Blanchard, 2008; Perusini and Fanselow, 2015). Broadly speaking, fear corresponds to immediate threat while anxiety is elicited when threat is spatially or temporally distant and uncertain (Mobbs et al., 2007; Blanchard and Blanchard, 2008; Perusini and Fanselow, 2015). Both fear and anxiety are adaptive and elicit evolutionarily advantageous defensive behaviors aimed at avoiding bodily harm and predation; however, they can become pathological, being excessively or inappropriately expressed such that they significantly interfere with one's daily activities.

There are various disorders of anxiety with examples ranging from generalized anxiety disorder and specific phobias to social anxiety disorder, agoraphobia, and panic disorder (DSM-5). While symptoms for each disorder differ, somatic symptoms common to most forms of anxiety include periods of intense physiological arousal, restlessness, muscle tension, heart palpitations, fatigue, shortness of breath, and avoidance behaviors (Beck et al., 2005; NIMH, 2019a). Anxiety is also characterized by cognitive symptoms such as sustained periods of rumination and worry (Nolen-Hoeksema, 2000; NIMH, 2019a). There is often a positive feedback loop component between the somatic symptoms and the cognitive dimension of anxiety, such that one initiates and exacerbates the other (Ehlers et al., 1988).

8.1 Belief-state models

Many theories view anxiety as involving negative beliefs about the future (MacLeod and Byrne, 1996; Beck et al., 2005); however, to understand anxiety as a form of negative future thinking requires identifying the neural and cognitive processes that support prospection. Episodic future thinking (i.e., the ability to perform mental simulations) has become an increasingly studied topic in recent years, and there is now a growing body of evidence that both humans and non-human animals engage in episodic future thinking to some degree (Clayton et al., 2003; Suddendorf, 2013; Redish, 2016). One facet of mental simulation involves the representation of spatio-contextual information stored in the hippocampus (Hassabis et al., 2007; Schacter et al., 2008; Redish, 2016). The hippocampus encodes spatial and contextual maps of experienced environments which can then be explored offline to facilitate learning even when the animal is not currently occupying that environment (O'Keefe and Nadel, 1978; Redish, 1999).

Animals perform this prospective planning during periods of hippocampal theta, the 4-10 Hz oscillation prominently observed in the hippocampal local field potential (Redish, 2016). Furthermore, there is high hippocampal theta power during reward-based (Johnson and Redish, 2007) and threat-based (Kim et al., 2015) conflict in rodents, as well as in humans during avoid-approach conflict (Ito and Lee, 2016). The theta-suppression model of anxiolytic drug action suggests that anxiolytics (particularly barbiturates and benzodiazepines) function by attenuating hippocampal theta (Yeung et al., 2012), thus possibly impairing the ability to engage in hippocampal-dependent episodic future thinking (Walters et al., 2019).

While there have been a few models of fear focusing on amygdalar circuitry, biases in threat processing, and defensive behaviors, there have not been many computational models of anxiety per se (Raymond et al., 2017). Gray (1982) was the first to suggest that the septo-hippocampal circuit plays a role in anxious prospection and the resolution of conflict between competing goals (e.g., during avoid-approach conflict). More recently, Dayan and Huys (2008) used reinforcement learning to model future-oriented thoughts that terminate in either positively or negatively valued predicted future states. They further modeled a hypothesized effect of serotonin on pruning by stopping these trains of thought when they transition to the consideration of aversive outcomes. Avoid-approach conflict models of anxiety in humans suggests that behavioral inhibition, a hallmark readout of anxiety, coincides with goal-directed planning and acts as a cost-minimizing strategy in environments where threat and reward are correlated (Bach, 2015), thus supporting the case that subjects are considering future outcomes during anxiogenic decision-making.

Some have used a discrete state model known as a partially observable Markov decision process (POMDP) to model belief states and their relation to mood and action-selection. In such models, the environment is treated as noisy and uncertain, and thus agents represent probabilistic beliefs over the states to inform action-selection. In these models, the agent's beliefs are updated on the basis of observations obtained from performing actions (Paulus and Yu, 2012). POMDP models also allow agents to perform mental simulations in addition to physical actions. The resulting fictive observations can inform state estimations (and thus decision-making), with these mental simulations having specific representational elements (e.g., space, value, and state inference) supported by distinct neurobiological substrates (Walters et al., 2019).

Data supports the theory that impairments in episodic foresight may in fact be central to certain anxiety disorders (Miloyan et al., 2016). Avoidance behaviors which reduce the probability of experiencing a future aversive outcome are fundamental to most anxiety disorders and have been shown to be anxiolytic (Lovibond et al., 2008). The expectancy-based model of anxiety claims that expectations about aversive future events generates anxiety which avoidance behaviors serve to alleviate (Declercq et al., 2008). Exposure therapy is aimed at subverting avoidance behaviors and forcing the individual to learn from experience that their expectations are largely inaccurate. Such expectations about the future appear to become pathological in individuals with generalized anxiety disorder, who, for example, have difficulties constructing positively valenced episodic simulations and perceive negatively valued simulated events as being more likely to happen than their non-anxious counterparts (Wu et al., 2015).

9 Where to from here: moving (slowly) toward precision psychiatry

The cases described above reveal a field in flux. Some disorders, such as schizophrenia and addictions, have received more focus, while others, such as anxiety and depression, have not been as heavily modeled. While early computational models of psychiatric disorders show a great deal of promise and a clear potential for future breakthroughs, there are as yet no current examples where these new perspectives have actually changed clinical practice (Redish and Gordon, 2016; Stephan et al., 2016). However, mounting evidence suggests that a biologically-informed, computationally-grounded approach to psychiatry will lead to a richer etiological understanding of these disorders and allow not only better disease progression prediction but better treatment options in a personalized, patient-specific manner (see Appendix B). Indeed, taking a computational approach to psychiatry has already positively impacted our understanding of the nature of mental illness at various levels, and these insights do appear to have diagnostic and therapeutic value (Redish and Gordon, 2016; Bzdok and Meyer-Lindenberg, 2018). Many groups are working to bring these insights into the clinic, with these efforts representing a collaboration between fundamental neuroscientists studying the underlying neuroscience of psychiatric phenomena, clinicians and clinical scientists who treat and study patients, and computational neuroscientists working to bridge the gap between the two.

If one looks at the process of scientific discovery, one tends to find a 30 year (or longer) path from initial breakthrough to implementation (Contopoulos-Ioannidis et al., 2008; Redish et al., 2018). This occurs due to the fact that this path requires three stages. First, in the **fundamental science stage**, one must find the space of a discovery – *Where does it apply? What are the parameters of the discovery? What are the regularities? What are the correct constructs, the correct language, with which to talk about these parameters? How does one measure them?* Second, in the **engineering stage**, one must find the space of control – *How does knowing about that discovery allow us to take action? What are the subtleties of specific instantiations of control?* Third, in the **implementation stage**, one must find a way to make that control ubiquitous – *How can we make that control reliable that it works under every appropriate condition? And not to be applied in inappropriate conditions? How can we make it simple enough for everyone to use?* Of course, these three stages do not occur in a completely linear manner, and there are multiple recursive interactions as engineering observations lead to new fundamental discoveries or implementation considerations require re-engineering. Nonetheless, this basic sequence is a good description of many breakthroughs.

Computational psychiatry as a field is presently at the boundary between the fundamental science and engineering stages. We know that the new language of psychiatry will be grounded in an understanding of information-processing and a thoughtful approach to delineating the continually evolving interactions between decision-making systems, their underlying network dynamics, and the environment. We know that measuring these phenomena will require behavioral assays and neural measurements obtained from EEG, fMRI, and other technologies. We know that there are important unresolved questions about the underlying neural processing occurring within the brain's decision systems and their malleability. We also know that nosology is going to depend on complex interactions between underlying neurocomputational dysfunction and observable clinical phenotypes as outlined in the examples discussed in this chapter. Lastly, we know that successful treatment will depend on neural manipulations (e.g., TMS, tDCS, focal ECT, ketamine and other pharmacological infusions, invasive neurostimulation, etc.), behavioral manipulations (e.g., cognitive and social-affective training), and meta-cognitive therapies that induce both restorative and compensatory processes.

The promise of computational psychiatry is a new view on psychiatry itself and on how we approach mental disorders. Characterizing a complex phenomenon mathematically accelerates our understanding of it, and the ability to use those mathematical models and test their predictions against experimental data allows us to do this in a quantitative way. Successfully integrating the most recent insights and methods from computational neuroscience into psychiatry will have large and meaningful consequences for the future of mental healthcare (Huys et al., 2016; Redish and Gordon, 2016; Vinogradov, 2017; Lynn and Bassett, 2019).

10 Conclusion

In this chapter, I demonstrated how specific vulnerabilities in neural (and more generally biological) information processing systems produce particular psychiatric phenomena. Importantly, I highlighted how cognitive and behavioral processes are computationally multifaceted — addiction, for example, can be split into separate liking and wanting components, each with its own unique biological underpinnings. This computational approach can be applied to better understand a diverse range of disorders at all scales, from molecular and cellular to network and behavioral.

In the next chapter, I will take a similar approach to anxiety: as with addiction, we will **(1)** see how anxiety can similarly be parsed out into dissociable neural algorithms, and **(2)** explore the role that the amygdala, hippocampus, and prefrontal cortex play in these distinct anxiety systems.

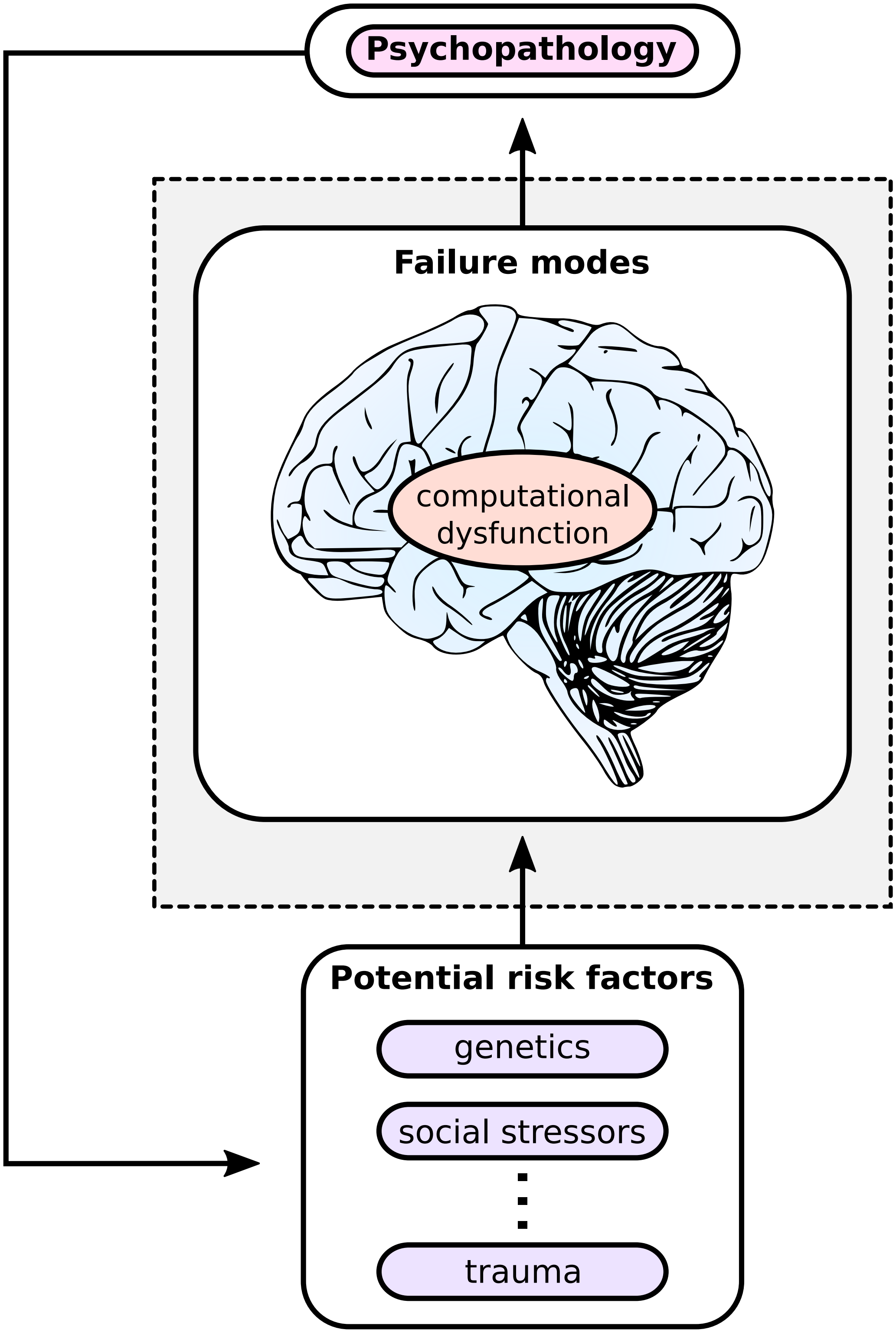

Psychopathology
Failure modes
computational
dysfunction
Potential risk factors
genetics
social stressors
⋮
trauma

Figure 2.1 Underlying causes of psychiatric disease in the form of potential risk factors lead to computational dysfunctions in the nervous system. These computational dysfunctions then lead to psychopathology, which in turn influences the array of potential risk factors.

— Theoretical perspective —

Chapter 3

Parsing the neural algorithms underpinning anxious states

1 Dissecting anxiety

Despite being studied for well over a century, both fear and anxiety lack a widely agreed-upon scientific definition (Mobbs et al., 2019). Though anxiety is often regarded as a unitary phenomenon conserved across anxiogenic scenarios, the available data suggest that anxiety actually describes a heterogeneous family of threat-processing algorithms. Below, I will argue that we can tease apart distinct neural algorithms underpinning various states of anxiety by taking a computational approach to the question of anxiety. I will argue that by looking at the problem through this lens, we will identify not one unified anxiety, but rather several anxiety subsystems that must be accounted for. I will make the case that by dissecting these algorithms we can generate a more clear and natural definition for anxiety that captures its full neurophysiological as well as phenomenological complexity, and that, importantly, this approach has implications for improving clinical treatment. A primary emphasis will be placed upon three such neural algorithms in the context of anxiety: associative learning, generalization, and prospection. Associative learning refers to the pairing of two stimuli (or a stimulus and a behavior) via Hebbian plasticity (see Chapter 1), generalization refers to the ability to transfer learning that occurs in one context to novel scenarios, and prospection refers to the ability to mentally imagine potential future scenarios (see Chapter 2). While fear- and anxiety-related associative learning is well-understood at the neural level (LeDoux, 2000; LeDoux, 2007), how generalization and prospection are instantiated in the brain are two looming questions in the field of neuroscience. As such, the aim of this chapter is twofold: **(1)** I will outline how anxiety can be parsed into at least three distinct neural algorithms (i.e., associative learning, generalization, and prospection), and **(2)** I will emphasize the privileged role that the amygdala, hippocampus, and prefrontal cortex play in these threat-processing subsystems.

2 Associative learning algorithms: cues and contexts

The temporal pairing of a stimulus (a cue or a context) with an outcome (e.g., delivery of electric shock) is at the core of associative learning. As discussed in Chapter 1, stimulus-outcome associations, once acquired, can be leveraged to release conditioned responses upon presentation of the CS: A CS associated with an appetitive outcome (e.g., sucrose delivery) elicits approach-type behaviors whereas a CS associated with an aversive outcome (e.g., electric shock) elicits avoidance-type behaviors (e.g., freezing or fleeing).

Importantly, conditioned stimuli need not always be discrete — exposure to an environment with multiple elemental cues (e.g., a room with a particular smell and arrangement of specific objects) can become holistically encoded as a context. Conditioned responses to cues and contexts can be innate (genetically hardwired) or learned over the course of a lifetime. An agent must compute three core components in order to perform associate learning: **(1)** value (was the experienced stimulus appetitive or aversive?), **(2)** context (what was the structure of the environment where I experienced the stimulus?), and **(3)** behavioral control (which behaviors should be expressed in response to specific cues and contexts?). Firstly, I will briefly review the role of various structures (with an emphasis on the amygdala, hippocampus, and prefrontal cortex) in representing these three features of associative learning; secondly, I will discuss the role of neural oscillations in orchestrating activity within and between structures; and thirdly, I will review the effect of anxiolytics on these structures and neural processes.

2.1 Representations of value

The amygdala plays a well established role in the generation of defensive behaviors in response to threat (LeDoux, 2000; LeDoux, 2007). The basolateral amygdala (BLA) is composed of several neural subsets that project to distinct targets. For example, there are BLA neurons that project to the ventral hippocampus (vHPC) that respond to positively and negatively valenced conditioned stimuli (tone+sucrose and tone+quinine, respectively); there are BLA neurons that project to the central nucleus of the amygdala (CeA) that respond to negatively valenced conditioned stimuli; and there are BLA neurons that project to the nucleus accumbens (NAc) that respond to positively valenced conditioned stimuli (Beyeler et al., 2016). Altogether, these data support a view of the amygdala as categorizing inputs as aversive or appetitive and then routing that information to the ventral striatum (vStr), vHPC, and even back to itself to be integrated with other inputs.

2.2 Representations of context

A wealth of data confirms that lesioning the dorsal hippocampus (dHPC) impairs both the acquisition (Phillips and LeDoux, 1992; Phillips and LeDoux, 1994; Maren and Fanselow, 1997) and expression (Maren et al., 1997; Frankland et al., 1998; Anagnostaras et al., 1999; Trivedi and Coover, 2004) of contextual fear. Similarly, lesioning the vHPC prevents the acquisition (Maren, 1999; Richmond et al., 1999) and expression (Maren, 1999; Richmond et al., 1999; Maren and Holt, 2004) of contextual fear. This apparent functional overlap of the dHPC and vHPC disappears, however, when we dissect the contribution of these two structures in different tasks (McHugh et al., 2004; Xu et al., 2016) and at varying time points over the course of conditioning (Beeman et al., 2013), thus supporting the view that the dHPC and vHPC play functionally dissociable roles during aversive memory acquisition, consolidation, and retrieval.

Altogether, the available data suggests that each stage of aversive memory acquisition, consolidation, and expression are mediated by distinct structures in a temporal- (e.g., short term versus long term) and stimulus-specific (i.e., cues versus contexts) fashion. Namely, **(1)** the dHPC appears to play a role in generating a cognitive map of the task space (e.g., the pre-exposure facilitation effect, Rudy et al., 2002), and **(2)** the vHPC, in tandem with other structures, appears to play a role in encoding the valence of a context (McHugh et al., 2004; Sierra-Mercado et al., 2011; Jin and Maren, 2015; Xu et al., 2016; Kim and Cho, 2017).

2.3 Representations of behavioral control

This takes us to the third component of associative learning: how to control behavior in response to specific cues and contexts. Agents learn the value of cues and contexts through experience. Once these associations are formed, presentation of a given cue or context must drive the expression of the appropriate behavior while inhibiting the expression of inappropriate behaviors. Additionally, this process needs to be reversible, so that behavior can flexibly adapt to a changing environment. In this section, we will explore the structures and interactions that are involved in this process.

It is believed that fear conditioning and subsequent fear extinction create two separate memory traces (Quirk, 2002). The prefrontal cortex contains multiple subregions, among them being the prelimbic cortex (PL) and infralimbic cortex (IL). In an effort to dissociate the contribution of PL and IL during cued (i.e., auditory) fear conditioning, Quirk et al. reversibly inactivated each structure independently, finding that **(1)** fear expression (i.e., freezing) was impaired with PL but not IL inactivation, and **(2)** that extinction (i.e., unlearning CS-US pairings) was impaired with IL but not PL inactivation. Similar results have been reported in contextual fear conditioning (Orsini et al., 2011). These data fit into a popular model of prefrontal function (Giustino and Maren, 2015) that suggests that PL is involved in promoting the behavior associated with the conditioned memory (i.e., freezing), whereas IL promotes the behavior associated with the extinction memory (i.e., extinguished freezing).

2.4 The role of neural oscillations during fear and anxiety

The medial prefrontal cortex (mPFC) and vHPC are known to increase theta synchrony during innately anxiogenic tasks such as the open field and elevated plus maze (EPM) (Adhikari et al., 2010), and there is increased theta synchrony between the lateral nucleus of the amygdala (LA) and layer CA1 of the dHPC during cued and contextual fear conditioning (Seidenbecher et al., 2003; Narayanan et al., 2007). This raises the more general question of the role that neural oscillations might play in facilitating multi-structure communication during periods of fear and anxiety.

Tasks such as the EPM provide the opportunity to study the neural basis of innate context-induced anxiety owing to rodents' hardwired aversion to elevated, open spaces. For instance, when rodents pause at the intersection between the open and closed arms of the EPM before deciding to venture out onto an open arm or retreat back into a closed arm, the vHPC (but not the dHPC) exhibits theta oscillations (Jacinto et al., 2016). However, increased theta power is seen in the dHPC during periods of avoid-approach conflict, and lesioning the amygdala eliminates this dHPC theta modulation (Kim et al., 2015). In agreement with these data, lesioning the vHPC increases open arm exploration on the EPM, whereas lesioning the dHPC has no effect on EPM anxiety-like behavior (Degroot and Treit, 2004).

Altogether, the neural data agrees with the lesion data: theta power increases in dHPC during the presentation of conditioned cues or contexts, and high dHPC theta power corresponds to the temporal window during which dHPC is involved in generating threat-related conditioned responses. This is consistent with the dHPC's role in recognizing episodes, both spatial and non-spatial (Aronov et al., 2017), and consolidating those episodic events to cortex and vHPC (Komorowski et al., 2013). Furthermore, these data suggest that coordination between neural assemblies in the dHPC and LA are required for both the consolidation and reconsolidation of aversive engrams. Altogether, these data once again support that the dHPC and vHPC each play dissociable roles in responding to threat and encoding aversive events: the dHPC encodes episodes of cued and contextual conditioning early on and exhibits theta modulation during learned conflict (but not innate conflict) whereas the vHPC is required for innate contextual freezing during which it is engaged in theta synchrony with mPFC.

2.5 Anxiolytics

The hippocampal theta oscillation has long been suspected to play a critical role in anxiety (Gray, 1982). Building on this suspicion, much work has outlined how lesions to the septo-hippocampal system appear to reduce anxiety-like behaviors (Degroot and Treit, 2003; Gray and McNaughton, 2003). Furthermore, it has been observed that most anxiolytic drugs impair the hippocampal theta rhythm (McNaughton, 2007). This has led to the hypothesis (known as 'the theta suppression model of anxiolytic drug action') that any drug that impairs or reduces hippocampal theta is a potentially viable anxiolytic compound.

Yeung et al. tested this hypothesis by using phenytoin, a compound that inhibits persistent sodium currents (thus likely to suppress hippocampal theta) but with no documented anxiolytic effects. They found that, like diazepam, intraperitoneal phenytoin alleviates anxiety-like behavior on the EPM (Yeung et al., 2012). These data provide support for the theta suppression model of anxiolytic drug action. It is interesting to note here that this effect of theta suppression on EPM is likely exerted on the vHPC given that (as discussed in section 2.4) there is no significant change in dHPC theta during EPM deliberation (Jacinto et al., 2016).

2.6 Summary

Altogether, associative learning systems are differentially activated contingent on the nature of the environment and task. These differences in circuit activation highlight the existence of multiple threat processing algorithms *within the associative learning system alone*, each being largely mediated by the amygdalo-hippocampo-prefrontal network. Developing a more nuanced appreciation of these algorithms will be essential if we hope to gain a deeper understanding of what fear and anxiety are and how they are processed and represented in the brain.

3 Generalization algorithms: the problem of induction

Cues and contexts encountered in the environment do not always take the same form nor do they appear reliably in precisely the same setting. Thus, the ability to abstract information gleaned from limited experience and determine its applicability in new environments is highly advantageous, while the inability to extract shared features between non-identical experiences impairs predictive learning. Striking a balance such that a given memory is malleable enough to generalize to novel scenarios but sufficiently rigid to ensure predictive value is crucial; however, if the scales are tipped and abstraction or specificity becomes too extreme, cognitive dysfunction results.

In this section we will discuss how generalization is represented in the brain. We will explore the notion of representation learning (Niv et al., 2019), and how algorithms such as high-dimensional representational clustering (Bernadi et al., 2020) might offer a coding scheme for generalization. This will be grounded in the underlying neurobiology, and we will review current data suggesting that sensory factorization (Whittington et al., 2020) and ordinal representation (Sun et al., 2020), operations that could confer the ability to perform flexible inference and abstraction, might be implemented in structures like the hippocampus and prefrontal cortex.

3.1 Pattern completion

Attractor networks have illuminated how an engram might generalize to similar but non-identical situations (Hopfield, 1982). Attractor networks can be biologically realized by applying the principle of Hebbian plasticity to a recurrent neural network. Evidence exists that the hippocampus employs attractor networks to promote memory recall and shift between cognitive maps (Wills et al., 2005). Briefly, this is achieved by a given stimulus generating a certain pattern of neural activity and potentiating the weights (synaptic strength) of the network accordingly, thus forming a ‘basin of attraction’. Future inputs that elicit similar patterns of activity in the network (whether because they are similar to the initial stimulus or because they are a degraded version of that same initial stimulus) will result in the subset of activated neurons in the network recurrently activating the remaining neurons that constitute that ‘basin’, thus ‘dragging’ an incomplete or noisy pattern of activity toward a stable state (e.g., a category). This process is a form of pattern completion.

Examples of pattern completion in the recall and generalization of memory traces abound in the neuroscience literature and are essential to understanding the neural computations that underpin fear and anxiety. For example, Südhof et al. showed that inactivation of mPFC projections to nucleus reuniens (NR) results in overgeneralization of contextual fear memories (i.e., elevated freezing in a similar but non-identical chamber to where tone+shock pairings occurred) (Xu and Südhof, 2013). Furthermore, Sudhof et al. demonstrated that NR bidirectionally controls fear memory generalization; inhibition of NR synaptic transmission results in overgeneralization of contextual fear memories, while suppression of synaptic inhibition onto the NR results in reduced generalization of contextual fear memories (i.e., decreasing activity in the NR drives engrams to be hypergeneralized while increasing activity in the NR drives engrams toward the hypogeneralized side of the continuum). Modulating mPFC and NR activity had no effect on cued or contextual conditioning, however, suggesting that the mPFC-NR-HPC circuit is specifically involved in tuning engram generalizability. This might be accomplished by mPFC accessing real-time episodic traces in dHPC and using that data to refine prefrontal representations of the current environment and its possible contingencies given its similarity to previous environments with which the animal has experience.

3.2 Engram abstraction

Data suggest that PL is critical for the modification of reactivated contextual memory traces (Vanvossen et al., 2017). For example, temporary inactivation of PL (but not IL) after aversive memory retrieval impairs reconsolidation at multiple time points (1-, 7-, and 21 days) (Stern et al., 2013). Indeed, Homberg et al.

suggest that the generalizability of emotional engrams is controlled by the interaction between the hippocampus and mPFC. The authors argue that vHPC projections to GABAergic interneurons in PL are capable of gating the context in which conditioned responses are released, and that, as a result, dysfunction in this pathway can result in the type of overgeneralization seen in post-traumatic stress disorder (Lopresto et al., 2016). One interpretation is that while prelimbic appears to augment retrieved engrams (the generalizability axis), infralimbic seems to monitor the predictive value of CS-US associations (the predictive potency axis). In this view, IL-PL form a complement, with PL abstracting individual aversive episodes and computing whether past experiences in dissimilar contexts generalize to novel scenarios (perhaps, as discussed in section 3.1, through a dialogue with the hippocampus via the NR) while IL is engaged in 'fact checking' the cache of S-R associations and adjusting their predictive value (which is functionally similar to mediating extinction learning). To further complicate matters, in addition to mediating extinction learning, IL has been shown to play a key role in regulating habitual behavior (Killcross and Coutureau, 2003; Smith et al., 2013). Chandler et al. offer a framework that attempts to reconcile these two reported roles of IL, claiming that both habit and extinction learning fall under the category of action-outcome inhibition (Barker et al., 2014). An interesting area for future research will be to further delineate how this model of IL-PL function during threat memory generalization relates to the model of IL-PL function during fear conditioning (see section 2.3).

Moving beyond PL and IL, Nader et al. explored the degree of involvement that the anterior cingulate cortex (ACC) has in encoding remote contextual fear memories and how the ACC interacts with the dHPC during retrieval and reconsolidation. They found that **(1)** inactivation of ACC 24 hours (but not 6 hours) after contextual fear memory reactivation resulted in impaired fear expression, **(2)** only the simultaneous inactivation of both the ACC and the dHPC impaired fear expression 6 hours after contextual fear memory reactivation, and **(3)** inactivation of the ACC 6 hours after contextual fear memory reactivation impaired the ability to generalize the contextual fear memory to a novel setting while leaving fear expression in the training context intact (Einarsson et al., 2015). This study highlights the critical role of the mPFC in developing a schematic representation of the memory over time and using that abstracted representation to generalize from a single episode to novel contexts.

3.3 Summary

In conclusion, the available data suggest that **(1)** dHPC represents recently acquired memories and over time gradually disengages with the memory trace via consolidation with the mPFC, **(2)** the mPFC develops an abstracted, schematic version of the hippocampal engram, and **(3)** remote reactivation of the memory trace engages the mPFC. One possibility is that, broadly speaking, the dHPC encodes episodes (i.e., things, places, and the sequence in which they occurred) and the vHPC encodes contexts (i.e., the environments in which episodes take place) and the salience of those contexts. While this is likely overly simplistic, there is some support for this framework.

There are drastic differences in functional connectivity between the dHPC and vHPC: while the dHPC sends projections to structures implicated in spatial navigation, visuospatial information, and memory, the vHPC projects to structures involved in motivated behaviors and affect, specifically aversion

(Fanselow and Dong, 2010). In light of this difference in functional connectivity, some have proposed that vHPC conveys contextual information to mPFC while BLA routes valence information to various downstream structures (e.g., the vStr, mPFC, and vHPC). These inputs might then get integrated in mPFC networks which form schemas — data structures that extract shared features from dissimilar events in the form of rules and contingencies that confer the ability to generalize from limited experience to novel scenarios (Preston and Eichenbaum, 2014; Bernadi et al., 2020). In addition to the mPFC, there is emerging evidence that the hippocampal-entorhinal system is involved in abstraction, specifically by learning low-dimensional representations of complex high-dimensional tasks (Bernadi et al., 2020; Whittington et al., 2020). Finally, during recall, the generalizability and applicability of these schemas appear to be tuned and assessed via mPFC-NR-dHPC projections (Benoit et al., 2011; Wu and Südhof, 2013; Preston and Eichenbaum, 2014).

4 Prospection algorithms: negative future thinking and anxiety

So far, we have discussed how stimuli in the form of cues and contexts gain predictive value by developing associations with outcomes. We have also discussed how there is a need to generalize these associations obtained from limited experience to novel settings that are similar but non-identical to the precise setting where the conditioning occurred. Associative learning is a well-characterized phenomenon and generalization is a highly active area of ongoing research, yet the neural and behavioral foundations of prospection remain poorly understood. It is to this topic that we will be turning next. Specifically, I will focus on internal drives (e.g., motivation and goal-directed decision-making); how they interface with complex, real-world environments that require cost-benefit analysis; and how this interaction can lead to uncertainty, conflict, and negative future thinking. Below, I will argue that **(1)** motivational conflict initiates mental simulations of the future, **(2)** these mental simulations are utilized to perform hypothetical cost-benefit analysis, **(3)** that this simulate-and-evaluate algorithm is computationally synonymous with certain types of anxiety, and **(4)** that dysfunction in this algorithm produces certain types of pathological anxiety.

4.1 A brief history of mental time travel

How decisions are made in the face of uncertainty is a central question in the fields of neuroscience, psychology, cognitive science, philosophy, and economics. If an agent has complete information of the task space, then the means by which it calculates the utility of a given action simply becomes a matter of understanding the degree to which it weighs and attends to the various factors influencing the decision-making process. Agents, however, rarely have complete information of the task space, and so it is imperative that we understand how decisions are made when decision-relevant features of the environment are partially or entirely unknown.

Below, I will briefly outline what is presently known about how decisions are made during uncertainty. As discussed in the previous chapter, I will further underscore future thinking (or ‘mental simulation’) as a critical component in this process. Future thinking, broadly speaking, is the ability to mentally generate hypothetical scenarios. It is generally thought that future thinking confers the ability to imagine potential outcomes, and thus, the ability to plan despite having incomplete information (analogous to how a chess-playing AI uses a tree search algorithm to ‘simulate’ possible moves).

Much of the literature that speaks to the importance of future thinking in decision-making has been conducted using reward-based paradigms, and the extent to which this phenomenon of future thinking occurs during threat-based decision-making remains unclear. Importantly, natural environments are complex and require agents to incorporate competing costs (i.e., threats) and benefits (i.e., rewards) into their decisional calculus. The aim of this section (and the following two chapters) is to address this imbalance in the decision-making literature and shed light on how decisions are made during uncertainty in complex, naturalistic environments.

4.2 What is known about future thinking?

In the 1930s, rodent spatial navigation studies suggested that hesitation behaviors reflected internal deliberation (Muenzinger and Gentry, 1931; Tolman, 1939). Over 40 years later, it was found that there were location-specific neurons in the hippocampus and entorhinal cortex that support spatial cognition (O’Keefe and Nadel, 1978). These spatially-tuned neurons are thought to construct a ‘cognitive map’ of the task space that aids the animal in spatial navigation.

In the 1980s, Tulving coined the term mental time travel (Tulving, 1985) to refer to the ability for humans to imagine potential future scenarios. Around the same time, the Bischof-Kohler hypothesis was formulated, stating that, unlike humans, non-human animals are ‘stuck in time’ and unable to anticipate their future desires and needs (Bischof-Kohler, 1985). Yet clearly this hypothesis is not true, given such counterexamples as nest building, food caching, and a wide range of non-human primate behaviors (Suddendorf et al., 2009).

Nearly 30 years after the discovery of place cells, neurophysiological experiments in rats showed that **(1)** place cell activity can be both local (representing the present location of the animal) as well as non-local (representing locations that the animal is not currently occupying), and **(2)** these non-local hippocampal representations occur during periods of hesitation resulting from decisional conflict over two competing, mutually exclusive reward offers (Johnson and Redish, 2007). This finding provided neural data confirming that non-human animals are capable of some form of prospective cognition (Zentall, 2006) in which the hippocampus plays an important role (Clayton et al., 2003). A growing body of literature supports that the hippocampus plays a central role in generating fictive representations of the environment, and that these representations inform ongoing decision-making behavior (Addis and Schacter, 2007; Buckner and Carroll, 2007; Gilbert and Wilson, 2007; Suddendorf and Corbaillis, 2007; Schacter and Addis, 2011; Redish, 2016; Kay, 2020).

4.3 Future thinking during threat-based uncertainty

Importantly, all of the above research was conducted using value-neutral or reward-based paradigms that produce approach-approach conflict. There are generally thought to be three types of motivational conflict: approach-approach, avoid-approach, and avoid-avoid (Lewin, 1931; Miller, 1944). While approach-approach conflict paradigms have primarily been used to study the neural basis of decision-making, avoid-approach conflict paradigms are typically used to model anxiety because they capture the complex, bivalent nature of most naturalistic environments. While there are some avoid-avoid assays (Wu et al., 2017), they are relatively less studied.

Miller et al. were the first to use rats to study avoid-approach conflict behavior (Miller et al., 1943). Hungry rats were placed at the start of a linear track and trained to traverse the full length of the track to receive a food reward. Once the rats had learned this simple task, the experimenters delivered a mild electric shock to the rats as they were eating at the location on the track where they received the food reward. On future trials, after having experienced the electric shock, the rats exhibited signs of conflict, slowly approaching the food source and pausing intermittently along the way. Occasionally, rats would change their mind mid-approach and retreat back to the start point. Rats are known to exhibit similar anxiety-like hesitation behaviors in semi-naturalistic anxiogenic scenarios such as exposure to a cat (Blanchard and Blanchard, 1989). These avoid-approach conflict behaviors were suspected at the time to be signs of anxiety (Miller, 1944), but the neural correlates of these anxiety-like hesitation behaviors would not be investigated for nearly 70 years.

More recently, avoid-approach foraging paradigms have been used to study the neural basis of anxiety in a semi-naturalistic setting in both rodents and humans. Choi and Kim developed a novel robotic predator-inhabited foraging arena task that requires hungry rats to forage for food in an unsafe environment, thus mimicking natural foraging conditions (Choi and Kim, 2010; Amir et al., 2015; Kim et al., 2015). Similar tasks have been adopted for humans, whether it be a looming tarantula (Mobbs et al., 2010), a gamified avoid-approach conflict foraging task with the risk of predatory attack (Mobbs et al., 2009; Qi et al., 2018; Bach et al., 2019; Fung et al., 2019; Korn and Bach, 2019), or a virtual reality environment in which the subjects are exposed to extreme heights or are attacked by dogs, spiders, or sharks to simulate naturalistic threat scenarios (Balban et al., 2020).

4.4 From threat-based uncertainty to conflict

Conflict occurs when two or more goals are in competition with one another. For example, the need to eat requires that an animal forage for food, but foraging for food increases an animal’s risk of falling victim to predatory attack (i.e., avoid-approach conflict). Uncertainty results from an environmental variable (e.g., a cue or context) being stochastic. Decisions almost always have to be made when the state of the environment, and by extension the outcomes of one's actions, are uncertain. Below, we will examine the potential ways in which uncertainty is represented at the neural level, how these representations correspond to beliefs, and how these beliefs are used to mentally simulate fictive future scenarios.

Beliefs can be viewed as representations of uncertainty and are often modeled as probability distributions over possible states. Typically, these models are Bayesian, with the organism using its internal generative model of the world to try and infer the latent variables (i.e., the posterior) of the environment that are the most likely cause of sensory observations (i.e., the likelihood). There are two prevailing theories for how probabilities are represented at the neural level (Fiser et al., 2010; Echeveste et al., 2020): **(1)** the parametric description which asserts that populations of neurons encode parameters of probability distributions corresponding to the uncertain variables (e.g., the mean and standard deviation) and **(2)** the sampling description which asserts that neurons encode individual variables from a high-dimensional distribution (i.e., the posterior) such that discrete patterns of network activity translate to samples taken from that distribution.

Experimental results suggest that non-local hippocampal representations in the form of forward and reverse replay events (Carr et al., 2011; Buzsáki, 2015; Redish, 2016) are a strong candidate for this sort of belief state sampling. Indeed, work has been done that examines how an organism might optimally replay these hippocampal sequences in order to sample from this high-dimensional belief distribution in order to maximize reward (Mattar and Daw, 2018). This raises the following question: when do animals perform belief state sampling?

It is known that when rats reach a fork-in-the-road that splits off to multiple reward offers, they will often pause and look back-and-forth between the potential paths they could take. This pause-and-look behavior, known as vicarious trial-and-error, was described in the 1930s by Muenzinger, Gentry, and Tolman (Muenzinger and Gentry, 1931; Tolman, 1939). More recent work discovered that place cells in the dHPC sequentially represent prospective trajectories ahead of the animal during hippocampal theta when rats are performing vicarious trial-and-error. Furthermore, when rats are deliberating between reward offers, non-local evaluation of the expected reward is seen in the ventral striatum (Redish, 2016). This illustration of non-local representation and evaluation of imagined future events suggests that this is one system by which rodents (Redish, 2016) and humans (Buckner and Carroll, 2007; Gilbert and Wilson, 2007; Gilbert and Wilson, 2009) simulate the future during periods of conflict. These data have not been studied as extensively in threat-based conflict tasks, however, though some human literature suggests that the same system is utilized during periods of negative future thinking (Benoit et al., 2016).

Non-human animals exhibit deliberation-like behaviors during threat-based conflict (Van der Poel, 1979; Pinel et al., 1989; Molewijk et al., 1995; Grewal et al., 1997) that mirror the vicarious trial-and-error behavior seen during reward-based conflict; however, whether a similar simulate-and-evaluate computation is being performed during these behaviors has not been explored. We hypothesize that the same prospection-and-evaluation system engaged during reward-based conflict is involved in threat-based conflict, and that the anticipation of negative outcomes is a core driver of the more cognitive forms of anxiety seen during periods of threat-related anticipation, uncertainty, and conflict.

4.5 Neural dynamics at choice-points

Intriguingly, risk assessment behavior in different threat-related tasks appears to involve distinct network dynamics. For example, risk-assessment behaviors (i.e., orienting and stretch-approach posture) seen in innately aversive environments (e.g., the EPM, Jacinto et al., 2016) or following contact with a nociceptive stimulus (i.e., shock, Seidenbecher et al., 2003) are not associated with dHPC theta power increase; however, when seen during avoid-approach conflict scenarios, those same risk-assessment behaviors are associated with increases in dHPC theta power (Kim et al., 2015). As discussed in section 2.4, theta synchrony becomes more coordinated between dHPC-mPFC during spatial decision-making at choice-points (i.e., during deliberation), with synchrony being greater during correct as opposed to incorrect decisions (Jones and Wilson, 2005). A large literature points to theta synchrony as being a medium by which the hippocampus, both dorsal and ventral, functionally interacts with other structures in order to temporally coordinate activity between distant neural ensembles in relation to behavioral demands.

In support of the view that dHPC episodes are accessed by mPFC, Ito et al. demonstrated that mPFC→NR→dHPC projections are required for goal-directed changes in dHPC place cell firing rates at choice-points (Ito et al., 2015). Furthermore, inactivating mPFC reduced rule coding but not place coding in the dHPC (Guise and Shapiro, 2017), even though mPFC only receives direct projections from vHPC and not dHPC (Jay et al., 1989). These data once again support a model of mPFC-NR-dHPC interaction where mPFC abstracts common features shared across many episodes (i.e., HPC→PFC consolidation), but mPFC opens dialogue with dHPC to access specific episodic events to tune the generalizability of these abstracted contingencies to new environments (for more on generalization, see section 3).

In addition to tuning the generalizability of retrieved engrams and promoting the expression of context-appropriate conditioned responses, PL appears to also guide motor output during motivational conflict. An interesting study by Graybiel et al. discovered that PL→dorsal striatum (dStr) projections are recruited to guide decision-making specifically during anxiogenic avoid-approach conflict (Friedman et al., 2015). Furthermore, inhibition of ACC→dStr projection neurons during avoid-approach conflict increased approach to high benefit options, suggesting either an increased sensitivity to benefit or decreased sensitivity to cost. This implies that PL is not only evaluating the generalizability of a memory during anxiogenic goal-conflict (see section 3.2), but it is also using contingency information to inform behavior during conflict.

The nature of the relationship between BLA and mPFC during decisional conflict is unclear. One study found a subset of BLA neurons that route threat-related (but not reward-related) information to PL, and that these BLA→PL neurons are necessary for the expression of cued fear behavior (i.e., freezing). Interestingly, decoding the activity of BLA→PL neurons more accurately predicted decision-making behavior during motivational conflict (i.e., simultaneous presentation of cues that predict both reward and threat) than the activity of non-PL projecting BLA cells (Burgos-Robles et al., 2017). This suggests that the BLA→PL subset is conveying data that informs decision-making behavior in conflict scenarios.

Altogether, there appears to be a BLA→PL→dStr pathway at play during conflict scenarios: **(1)** the parsing of valence (threat versus reward) occurs in BLA, **(2)** then that value information gets routed to PL to be integrated with the given context (e.g., via vHPC) in which it is occurring (BLA→PL), and **(3)** PL dictates the generation of motor output on the basis of that information (BLA→PL→dStr). Importantly, this valence→integrator→action selection algorithm appears to be especially critical during periods of motivational conflict such as is seen during anxiogenic avoid-approach scenarios. The resolution of this conflict likely occurs in mPFC (ACC, PL, and IL) where representations of task contingencies, outcome history, and prior experience are stored.

4.6 Summary

Approach-approach conflict appears to initiate a distinct neural algorithm: a hippocampally-mediated mental simulation of the future that is potentially paired with evaluations of anticipated outcomes (Addis and Schacter, 2007; Buckner and Carroll, 2007; Gilbert and Wilson, 2007; Suddendorf and Corbaillis, 2007; Schacter and Addis, 2011; Redish, 2016; Kay, 2020). However, it is unknown whether future thinking also occurs during avoid-approach conflict. I hypothesize that prospection occurs during periods of anxious conflict and that it is biologically implemented through the pairing of non-local spatial representations (via hippocampal place cell sequences) with non-local representations of value (via value-assessing structures like the vStr and BLA).

In the following two chapters, I present original research I have conducted to interrogate this hypothesis. First (see Chapter 4), I used a pharmacological approach in tandem with a novel avoid-approach predator-inhabited foraging arena task to show that anxiety-like hesitation behaviors similar to those described by Miller are attenuated by anxiolytic drugs, suggesting that these hesitation behaviors reflect an internal deliberation process. I then proceed to model this hesitation process as a belief-state updating loop that involves fictive representations of potential future outcomes using a partially observable Markov decision process. Second (see Chapter 5), I explored the neural representations underpinning these anxiety-like behaviors, aiming to determine whether non-local representations occur during periods of anxious conflict. To do this, I recorded from ensembles of neurons in the dorsal hippocampus of rats as they freely behaved in the predator-inhabited foraging arena task.

— Original research —

Chapter 4

Avoid-approach conflict behaviors differentially affected by anxiolytics: implications for a computational model of risky decision-making

Abstract

Whether fear or anxiety is expressed is thought to depend on an animal's proximity to threat. In general, fear is elicited when threat is proximal, while anxiety is a response to threat that is distal and uncertain. This threat gradient model suggests that fear and anxiety involve non-overlapping neural circuitry, yet few behavioral paradigms exist that elicit both states. We studied avoid-approach conflict in rats that were behaving in a predator-inhabited foraging arena task that involved tangible threat and reward incentives. In the task, rats exhibited a variety of both fearful and anxious behaviors corresponding to proximal and distal threat, respectively. We then administered ethanol or diazepam to the rats in order to study how anxiolytics affected these fear and anxiety behaviors. We discovered that both ethanol and diazepam attenuated proximal-threat fear-like behaviors. Furthermore, we found that

diazepam, but not ethanol, increased distal-threat anxiety-like behavior but also made rats less risk-averse. Finally, we describe how decisional conflict can be modeled as a partially observable Markov decision process and characterize a potential relationship between anxious behavior, diazepam’s ability to suppress hippocampal theta oscillations, and hippocampal representations of the future.

Chapter taken from:
Walters, C. J., Jubran, J., Sheehan, A., Erickson, M. T. and Redish, A. D. (2019), ‘Avoid-approach conflict behaviors differentially affected by anxiolytics: Implications for a computational model of risky decision-making’, Psychopharmacology, pp. 1–13. DOI: 10.1007/s00213-019-05197-0.

Introduction

Fear and anxiety are distinct states (Dias et al., 2013; Perusini and Fanselow, 2015), with fear being a set of defensive responses to visible and immediate danger (e.g., fighting, fleeing, or freezing) while anxiety is a form of risk-assessment involving the anticipation of potential future threat. Evidence supports the idea that behavioral responses to threat vary depending on the perceived proximity to the threat source, with anxiety-like behaviors (e.g., hesitation and avoidance) being elicited when threat is distal and fear-like behaviors being elicited when threat is proximal (Mobbs and Kim, 2015).

Threat-processing frameworks have developed around this notion of a threat gradient. For example, Fanselow and Lester's "Threat Imminence Continuum" characterizes the progression through four states of threat-processing that depend on the visibility and proximity of threat: (i) the preferred phase during which there is no threat, (ii) the pre-encounter phase during which the prey is vulnerable to threat (e.g., foraging) but no threat is detected, (iii) the post-encounter phase during which a threatening agent is detected but does not pursue the prey, and (iv) the circa-strike phase during which the threat source actively pursues the prey (Fanselow and Lester, 1988). Fanselow and Lester's model was then elaborated on in Mobbs' "Survival Optimization System" wherein there are five strategic systems that align with the four stages of the Threat Imminence Continuum: (i) prediction strategies, (ii) prevention strategies, (iii) threat orienting strategies, (iv) threat assessment strategies, and (v) defensive strategies (Mobbs et al., 2015). Prediction and prevention occur during both the preferred and pre-encounter states, threat orienting and threat assessment occur during post-encounter states, and defensive strategies occur during circa-strike states.

These frameworks harken back to early cognitive theories of anxiety which postulated that anxiety arises from negative evaluations of episodic future construction generated by hippocampal-cortical, hippocampal-accumbens, and hippocampal-amygdala interactions (Gray and McNaughton, 1982; Beck et al., 1984/2005). These anxiety frameworks differ from models of fear wherein most instantiations of fear are thought to result from either Pavlovian associations or species-specific, genetically-inherited circuitry (Bolles, 1970; LeDoux, 2012). Together, these frameworks and anatomical demarcations suggest that fear and anxiety should be both behaviorally and pharmacologically dissociable. However, few tasks clearly elicit and behaviorally dissociate fear and anxiety, making it difficult to study both states simultaneously. Naturalistic choice conflict paradigms in which reward incentives (e.g., hunger) are pursued at the risk of incurring punishment (e.g., exposure to a threatening predator) have been used since as early as the 1940s, and these avoid-approach conflict tasks are often structured so as to elicit both fear and anxiety (Miller, 1944). In recent years, these ethological approaches have largely been neglected in favor of more controlled Pavlovian fear conditioning paradigms (see reviews: Mobbs and Kim, 2015; Pare and Quirk, 2017).

One recently developed avoid-approach conflict paradigm is the predator-inhabited foraging arena (Choi and Kim, 2010; Amir et al., 2015; Kim et al., 2015, 2018). In this task, food-deprived rats are trained to forage for food pellets on a linear track with an enclosed nest-space at one end. A robotic predator is then introduced at the opposite end of the linear track from the enclosed nest-space, and the predator probabilistically surges forward and attacks when the rat approaches the feeder site near the robot. Following predatory attack, rats typically flee back to the enclosed nest-space (Choi and Kim, 2010) and proceed to exhibit various fear- and anxiety-like avoid-approach conflict behaviors.

The key difference between anxiety and fear on the predator-inhabited foraging arena can be operationalized by the reaction of the rat to the proximity of the predator – distal, approaching, and proximal – both pre- and post-attack. Post-attack, rats retreat and spend time hesitating at the opening of the enclosed nest-space (Amir et al., 2015) before deciding either to turn back into the nest-space or venture out and risk obtaining a food pellet. This conflict-associated hesitation is reminiscent of another risk-assessment behavior: the stretch-attend posture seen at the entry into novel, open spaces (Grewal et al., 1997) or into spaces laced with predator scent (Blanchard et al., 2001). Furthermore, Amir et al. reported that rats in the predator-inhabited foraging arena would occasionally leave the nest-space and begin their approach toward the ‘dangerous’ feeder site adjacent to the predator, and then, in what appeared to be a change-of-mind event, would turn around and retreat back into the nest-space, thus aborting the foraging attempt and failing to obtain the food pellet at the predator-occupied feeder site (Amir et al., 2015). Interestingly, data shows that rats are more likely to decide to leave the nest and forage for food if they have had amygdala lesions or intra-amygdalar infusions of muscimol (Choi and Kim, 2010). Furthermore, it has been shown that there are two subsets of basolateral amygdala neurons that ramp in activity at the nest-space choice-point: one population prior to retreating back into the nest (pause-retreat) and the other before deciding to initiate a foraging attempt (pause-approach) (Amir et al., 2015).

One challenge with rodent decision-making tasks is finding a balance between ethological validity and task complexity (Juavinett et al., 2018). The predator-inhabited foraging arena is a model of real-world foraging involving deliberation, approach incentives, and a sustained and tangible threat source that strikes a balance of low task complexity and high ethological validity, a valuable ratio for studying the interaction between fear, anxiety, and decision-making under naturalistic conditions. In contrast to many fear and anxiety assays, a benefit of this paradigm is that it is effectively one-dimensional. This allows for a clear-cut and continuous quantification of binary economic 'stay-or-go' decision-making along the full length of the track while providing access to similar circuitry and behavior involved in more complex, two-dimensional, real-world foraging scenarios. Furthermore, it has the advantage of evoking a variety of fear- and anxiety-like hesitation behaviors that neatly map onto the threat gradient continuum.

Although the design of the predator-inhabited foraging arena, with its spatially distinct distal-to-proximal threat gradient, allows for the differentiation of fear- and anxiety-like behaviors, no one has yet looked at what effects anxiolytics have on these behaviors. Verifying that anxiolytics do in fact reduce the anxiety-like behaviors seen on the predator-inhabited foraging arena would support the task's construct validity and serve as an important data point if it is to be used more widely in the study of fear and anxiety in rodents. Furthermore, it is known that there are sex differences in how males and females of various species (e.g., mice, rats, non-human primates, and humans) express fear and anxiety both neurophysiologically and behaviorally (Johnston and File, 1991; Crepeau and Newman, 1991; Maeng and Milad, 2015; Yokota et al., 2017), yet all experiments on the predator-inhabited foraging arena to date have used only male rats (Choi and Kim, 2010; Amir et al., 2015; Kim et al., 2015, 2018).

To explore these two questions, we acutely administered ethanol and diazepam, two pharmacological agents that have well-characterized anxiolytic effects (Wilson et al., 2004), to both male and female rats in the predator-inhabited foraging arena. We found that both ethanol and diazepam reduced approach time toward the threat source, indicating an attenuated fear response to proximal threat. This is consistent with the effect of ethanol and diazepam on other threat paradigms (Blanchard et al., 1990, 1993). Ethanol, however, had no effect on deliberative pausing behavior at the nest-space choice-point (an anxiety-like behavior) while diazepam increased the amount of deliberative pausing at the nest-space choice-point. Lastly, diazepam, but not ethanol, increased the probability of the rats making risky foraging decisions following choice-point deliberation.

MATERIALS AND METHODS

Subjects

Both male (n = 8) and female (n = 6) Brown Norway rats aged 8-10 months were used as subjects. All rats were maintained on a 12:12 hr light/dark cycle. Rats were food-restricted such that they had 1 hr per day to work for food in the foraging arena. Rats were always kept above 80% free-feeding weight and had unlimited access to water outside of the foraging arena. All procedures were approved by the University of Minnesota (UMN) Internal Animal Care and Use Committee (IACUC) and were performed in accordance with NIH guidelines.

Surgery

Following 7 d of linear track training, rats were chronically implanted with a light emitting diode (LED) fixed to the skull surface with metabond. Rodents were anesthetized throughout the duration of the surgery (0.5–2% isoflurane mixed with medical-grade O_2 via nosecone). To ensure rapid recovery, rats were given pre-surgery antibiotics (pennicillin G, 120 k units/kg) and post-surgery Baytril at 25mg/kg for three days post-surgery. Rats recovered from surgery in an incubator to maintain body temperature and they received Children's Tylenol post-surgery to alleviate discomfort. Rats were given 72 hrs to recover before resuming behavioral training.

Task and Data Collection

The foraging arena was 1.16m long and 33cm wide with walls 60cm tall. An overhead video camera tracked animal position from the head-LED at 30fps.

Behavioral Procedure

There were three phases of the predator-inhabited foraging arena: linear track training, injection habituation, and attack sessions (see Fig. 1). During all three phases of the task, sessions lasted 1 hr and rats began each session in the nest-space.

Phase 1: During the linear track phase, rats learned to shuttle back-and-forth on a linear track to receive food pellets at either end. Food pellets were only delivered at one of the feeders once the other feeder had been visited, thus rats needed to alternate between feeders to continually receive food. One end of the linear track had a partially enclosed ‘nest-space’, a high-walled room continuous with the linear track via an open doorway referred to as the ‘choice-point’. After ~10 days of linear track training, the rats transitioned to Phase 2 of the task.

Phase 2: During the injection habituation phase, rats received saline injections intraperitoneally (i.p.) prior to linear track training. Following 2 days of i.p. habituation, rats proceeded to Phase 3 of the task.

Phase 3: During the attack phase, rats received either drug or vehicle i.p. injections 5 min prior to being placed in the predator-inhabited foraging arena. Prior to the beginning of each session, a wall by the feeder site opposite the nest-space was removed and a robotic predator (SPIK3R, LEGO® MINDSTORMS® EV3) was placed in the open space near the feeder site. For the first 15 laps, the robot remained stationary. After the first 15 laps, the robotic predator would surge forward and attack the foraging rat with a 20% probability (i.e., on any given lap there was a 1/5 chance of the robot attacking) as it approached the feeder site adjacent to the predator.

Pharmacological Manipulation

Diazepam (2 mg/kg; Sigma) was dissolved in Tween 20 to prepare a stock solution, which was then diluted with 0.9% saline. The vehicle (10% Tween 20 in saline) was used as a control solution. Ethanol (1 g/kg) was prepared from 95% ethanol (Decon Labs) diluted with saline for a final concentration of 30% v/v to keep injection volumes below 5 mL/kg. Saline was used as a corresponding control solution. We chose these doses for both diazepam and ethanol due to their approximately matched anxiolytic efficacy (Wilson et al., 2004). All injections were administered i.p. 5 min prior to each session.

Data Analysis

All data was processed in MATLAB (MathWorks, Natick, MA) and statistically analyzed using JMP Pro 14 (SAS, Cary, NC). All figures depict the mean ± s.e.m. Statistical significance was assessed using an alpha value of 0.05. Matched-pairs or two-sample student's t-tests were used as indicated in each figure.

RESULTS

We trained food-deprived rats to forage in a linear track arena in which they had to leave an enclosed nest-space to receive food located at the opposite end of the track. Importantly, both the Zone 1 and Zone 3 food ports would only reset once the rat had visited the opposite feeder site. Once rats were sufficiently trained such that they were able to stay above 80% free feeding weight from daily 1 hr sessions on the linear track, we habituated them to i.p. injections for two days then we introduced a robotic predator (SPIK3R, LEGO® MINDSTORMS® EV3) to the arena situated near the feeder site opposite the nest-space. When the rat transitioned from Zone 2 to Zone 3, the robot would surge forward and attack the rat with a 20% probability (Fig. 4.1). During these attacks, the robotic scorpion charged forward toward the foraging rat and repeatedly snapped its pincers while emitting clicking sounds. Following predatory attack, rats typically fled to the nest and proceeded to exhibit a variety of fear- and anxiety-like hesitation behaviors for the duration of the session. For example, rats exhibited slower Zone 1 (‘safe’) to Zone 3 (‘dangerous’) approach times (i.e., slow ‘outbound’ laps, Fig. 4.2), more Zone 2 change-of-mind events (Fig. 4.3), more hesitation at the nest-space choice-point (Fig. 4.4), and heightened risk-aversion (Fig. 4.5). Slower ‘safe-to-dangerous’ outbound laps and an increase in the number of change-of-mind events occurred as threat became more proximal. The predator became fully visible to the rats when they were roughly midway down the linear track, which would coincide with the location where mid-track abort events were observed. Thus, in keeping with the threat gradient model, we classified these ‘visible-and-proximal’ threat-induced behaviors as being fear-like. Conversely, both choice-point hesitation and the initiation of a risky action occurred on this task when threat was distal, thus we categorized these two behaviors as being anxiety-like. In order to test the effects of anxiolytics on these fear and anxiety behaviors, we administered either ethanol or diazepam 5 min prior to an attack session. We interleaved anxiolytic injections with their vehicle controls. Drug administration was counterbalanced across all subjects using the following four permutations: Tween 20, diazepam, saline, ethanol; saline, ethanol, Tween 20, diazepam; diazepam, Tween 20, ethanol, saline; and ethanol, saline, diazepam, Tween 20.

Both diazepam and ethanol reduced the duration of approach to danger

To determine if rats distinguished between dangerous (i.e., 'outbound' laps when rats transitioned from Zone 1 to Zone 3 [Z1→Z3]) and safe directions (i.e., 'inbound' laps when the rats transitioned from Z3→Z1) on the task, we quantified the time it took rats to run risky, outbound laps (i.e., potential predatory attack) versus safe, inbound laps (i.e., no risk of predatory attack) during attack sessions. Inbound laps were significantly faster than outbound laps (Fig. 4.2B; matched pairs t-test, $t_{(13)} = 4.72$, $p = 0.0002$), and risky outbound laps were faster under ethanol (Fig. 4.2C; matched pairs t-test, $t_{(13)} = 3.24$, $p = 0.0032$) or diazepam (Fig. 4.2D; matched pairs t-test, $t_{(13)} = 2.80$, $p = 0.0075$) relative to their vehicle controls. We found no significant differences in lap duration between males and females (Fig. 4.2E; two-sample t-test, $t_{(12)} = 1.29$, $p = 0.2193$).

Both diazepam and ethanol reduced the number of change-of-mind events

As previously reported, we observed that rats that had recently experienced predatory attack would leave the nest-space, slowly approach the dangerous food source, pause, and then turn around and return back to the nest-space (i.e., Z1→Z2→Z1), thus giving up the opportunity to receive food (Amir et al., 2015). We quantified the number of these mid-track abort (MTA) 'change-of-mind' events as well as a control behavior, anti-MTA's (i.e., Z3→Z2 →Z3). MTAs were far more likely to occur than their control behavior, anti-MTAs (Fig. 4.3B; matched pairs t-test, $t_{(13)} = 3.24$, $p = 0.0032$). Interestingly, both ethanol (Fig. 4.3C; matched pairs t-test, $t_{(13)} = 2.12$, $p = 0.0268$) and diazepam (Fig. 4.3D; matched pairs t-test, $t_{(13)} = 2.17$, $p = 0.0245$) reduced the number of MTA events when compared against their vehicle controls. There was not a significant difference in MTA count between males and females (Fig. 4.3e; two-sample t-test, $t_{(12)}$ = 0.57, p = 0.5792).

Diazepam, but not ethanol, increased the amount of time spent hesitating at the choice-point

As noted in the introduction, hesitation at the exit of the enclosed nest-space is an anxiety-like behavior reminiscent of stretch-attend posture and open-space-entry-hesitation seen in a variety of anxiogenic tasks. Consistent with previous work (Amir et al., 2015), we found that rats would pause in the nest-space doorway before deciding to either approach the food source in Zone 3 or retreat back into the nest. We defined the deliberative pausing zone (DPZ) around the transition point from Zone 1 to Zone 2 and quantified epochs in which the rat entered the DPZ (the 'pause zone' in Fig. 4.4a) and stayed there for >2sec and <5min. Interestingly, while we found no difference in deliberative pausing between saline and ethanol (Fig. 4.4b; matched pairs t-test, $t_{(13)}$ = 1.58, p = 0.0689), rats spent more time pausing at the choice-point with diazepam than with Tween 20 (Fig. 4.4c; matched pairs t-test, $t_{(13)}$ = 2.19, p = 0.0234). There was no significant difference in deliberative pausing between males and females (Fig. 4.4d; two-sample t-test, $t_{(12)}$ = 0.93, p = 0.3676).

Diazepam, but not ethanol, increased the number of risky outbound journeys following choice-point hesitation

Following deliberative pausing at the choice-point, the rat must decide to either retreat back into the nest or to leave the nest and attempt a journey out to Zone 3 to get food. In order to explore how anxiolytics affected the rats' risk profiles, we quantified the number of these pause-then-approach (risk-taking) versus pause-then-retreat (risk-averse) events. After entering the nest doorway choice-point and pausing, rats were overall more likely to retreat than to approach (Fig. 4.5b; matched pairs t-test, $t_{(13)}$ = 4.46, p = 0.0006). Intriguingly, while there was no difference between ethanol and saline in the tendency to retreat or approach (Fig. 4.5c; matched pairs t-test, $t_{(13)}$ = 0.88, p = 0.8022), rats were more likely to approach and less likely to retreat under diazepam as opposed to its vehicle control (Fig. 4.5d; matched pairs t-test, $t_{(13)}$ = 2.65, p = 0.0100). We also found that males were more likely than females to retreat following deliberation at the choice-point (Fig. 4.5e; two-sample t-test, $t_{(12)}$ = 3.46, p = 0.0047).

DISCUSSION

We found that rats were slower on risky outbound journeys than on safe inbound journeys (Fig. 4.2), performed more MTAs on outbound rather than inbound journeys (Fig. 4.3), and retreated more often than they approached after pausing at the nest-space choice-point (Fig. 4.5). With both ethanol and diazepam, we found a decrease in the duration of risky outbound journeys (Fig. 4.2) and a reduction in the number of MTA events (Fig. 4.3). However, diazepam but not ethanol increased the amount of time rats spent pausing at the choice-point (Fig. 4.4) and increased the number of pause-then-approach events while decreasing the number of pause-then-retreat events (Fig. 4.5). We found no sex differences across any of the behaviors of interest except avoid-approach decisions following hesitation at the choice-point. However, it is possible that our sample size may be underpowered to reliably detect sex differences. Altogether, we found that two acutely administered anxiolytics, ethanol and diazepam, had different effects on decision-making behavior in a naturalistic avoid-approach conflict task. These data suggest that while both ethanol and diazepam dampened proximal-threat fear responses (Fig. 4.2-3), only diazepam increased distal-threat choice-point deliberation and, interestingly, this increased deliberation time resulted in more, not fewer, risky decisions to approach the threat source.

Our findings reveal that the anxiety-like behaviors seen in the predator-inhabited foraging arena are significantly diminished by anxiolytics, thus lending validity to the notion that this task can be used to model anxiety. The experiment presented here used single doses of ethanol and diazepam consistent with the dose range reported in similar behavioral pharmacology studies (Blanchard et al., 1990a; Blanchard et al., 1993; Treit et al., 1993). While there is an extensive literature describing the anxiety-related dose-response properties of both ethanol and diazepam (Blanchard et al., 1990b; Grewal et al., 1997; Kang-Park et al., 2004; Jimenez-Velazquez et al., 2010), it would be valuable for future work to investigate how the fear- and anxiety-like behaviors quantified on this task change with various doses of these two drugs in order to characterize the full dose-dependent response profile as has been done with other anxiogenic tasks (Blanchard et al., 1993; Wilson et al., 2004).

Interestingly, our data are at odds with some of the results reported in other anxiogenic tasks. For example, ethanol has been shown to increase formerly anxiogenic exploratory behavior in the open field arena and elevated plus maze (Prunell et al., 1994; Ferreira et al., 2000; Wscieklica et al., 2016) whereas ethanol did not increase rats' willingness to engage in risky foraging in our data (Fig. 4.4b). This is likely due to the different cognitive demands made by the two anxiogenic task types and highlights the fact that non-overlapping neural circuits may be engaged during such traditional anxiety assays (e.g., elevated plus maze and the open field arena) as opposed to avoid-approach conflict resulting from foraging in the face of a visible and active threat source (e.g., the predator-inhabited foraging arena).

It has been theorized that distal threat induces anxiety-like behaviors (e.g., hesitation), while proximal threat induces fear-like behaviors (e.g., freezing, fighting, fleeing) (Fanselow, 1994; Mobbs et al., 2015). Interpreted under this threat gradient model, both ethanol and diazepam appeared to attenuate the fear-response induced by proximal threat in our task (i.e., faster approach times and reduction of MTAs) while diazepam, but not ethanol, increased risk-assessment behavior seen when the threat was distal (i.e., increased choice-point hesitation). Current theories differentiate between multiple decision-making systems (i.e., action-selection systems), each of which is mediated by non-overlapping neural circuits. In the context of threat-processing, these theories postulate separate and dedicated decision-making algorithms for defensive reflexes, fear conditioning, innately aversive stimuli or contexts, and conflict (LeDoux and Daw, 2018). Our data suggests that ethanol and diazepam affect different components of these decision-making circuits.

A Markov model of risky decision-making

We posit that episodic future thinking and non-local cost-benefit analysis are neural algorithms central to certain forms of anxiety. Specifically, we posit that deliberative forms of anxiety such as those seen during motivational conflict requires a mental simulation of future scenarios, a representation of state-outcome contingencies, and a valuation of those expected outcomes in order for conflict to be resolved and an action to be selected. Furthermore, we argue that this process can be modeled as a Partially Observable Markov Decision Process wherein the agent iterates through a loop of belief-state updating until a state inference is made and decision threshold is passed resulting in commitment to one action over all other available actions.

A Markov model is a mathematical description characterizing the behavior of a system that probabilistically transitions through a series of states over time. There are four major types of Markov models: Markov Chains, Hidden Markov Models, Markov Decision Processes, and Partially Observable Markov Decision Processes. A Markov Chain (MC) describes a system in which there are a number of discrete observable states (S) with probabilistic transitions between those states (e.g., $S_1 \rightarrow S_1 = 0.2$ while $S_1 \rightarrow S_2 = 0.8$). A Hidden Markov Model (HMM) is related to a Markov Chain with the exception that the states themselves are unobservable, but the outcome of being in a given state is observable. As such, the present state can only be inferred from the observable outcomes and a probabilistic model of the parameters governing the unobservable states. Crucially, MCs and HMMs both involve systems lacking agency. A Markov Decision Process (MDP) accounts for a decision maker's ability to act and, in so doing, induce a state transition. Thus, in an MDP, each action (A) available to the agent has an associated set of observable states with associated state transition probabilities (e.g., $S_1,A_1 \rightarrow S_1 = 0.1$, $S_1,A_1 \rightarrow S_2 = 0.9$). Furthermore, in each state there can be an associated reward (R) or punishment (P) value. Lastly, a Partially Observable Markov Decision Process (POMDP) is like an HMM that accounts for a decision maker capable of taking actions, inducing state transitions, and receiving rewards and punishments. The agent in a POMDP attempts to infer the state it's in by not only having a functional model of the parameters governing the unobservable states and their transitions but also by means of exploratory actions that yield observations (O) which provide sensory evidence for the state the agent is currently occupying.

We propose that the motivational conflict underlying risky (i.e., costly and multivariate) decision-making can be modeled as a POMDP in which the agent performs exploratory behavior in the form of mental simulations (i.e., (Simulated State|Simulated Action) = (SS|SA)) to obtain observations of what is likely to happen in those simulated states (i.e., Expected(Outcomes|Simulated State) = E(O|SS)) (Fig. 4.6). Crucially, the values associated with these observations (i.e., V(O|SS)) are used to update the agent's belief about the state it is currently occupying, and this belief-state updating informs action selection (i.e., V(SA|E(O)) (Fig. 4.6). The agent is attempting to maximize R (e.g., access to food and safety) while minimizing P (e.g., exposure to danger and threat), thus

$$A(S_c) = \arg\max_v [V(O_n|SS_n)]$$

where $A(S_c)$ is the action taken in conflict state S_c, V is value (a weighted sum of R and P), O_n is the nth simulated outcome, $\arg\max_v$ is the maximal $V(O_n)$ for the nth iteration through the POMDP, and SS_n is the nth simulated state. When $A(S_c)$ exceeds some decision threshold (e.g., to either approach or avoid) the agent then selects that corresponding action (e.g., in the predator-inhabited foraging arena, either approach the food source, hesitate, or retreat back into the nest). Two possible models of how this can be achieved are a continuous integration to threshold (Fig. 4.6, lower left) or a series of non-additive, discrete simulation-space samples until one generated value passes a decision threshold (Fig. 4.6, lower right). Importantly, the baseline at which the value signal begins can be modeled as an incentive parameter that can start closer to or farther away from one of the decision thresholds depending on the internal state of the agent (e.g., if the rat is hungry, the baseline starts closer to Th_{App}). According to our POMDP algorithm, the belief-state updating loop repeats until a given cycle through the POMDP succeeds in generating a value signal V(SA|E(O)) that passes some confidence threshold for making a state inference I(S|V(SS)), thus resulting in a decision threshold being passed and an action being selected (Fig. 4.6).

In contrast to risky decision-making scenarios wherein the agent might have time to deliberate between options, there are also instances of explicit and immediate threat that require rapid action to ensure survival. This general 'detect-and-evade' algorithm can be modeled as a simple Markov Decision Process. In such situations, the agent transitions from a safe state (S_S) to a state of threat detection (S_T). The agent, upon detecting threat, mobilizes (e.g., changes in heart rate, attentional allocation, and circulating glucocorticoid levels) and executes an action (a) in the set of hard-wired, species-specific defense behaviors (A, such that $a \in A$). This threat evasion loop repeats until the agent either returns to a safe state or is captured and killed by the threat (Fig. 4.7).

Mapping avoid-approach conflict behavior to neuronal circuit computations

The threat-gradient framework posits that distal threat promotes anxiety-like behaviors (e.g., choice-point hesitation on the predator-inhabited foraging arena) while proximal threat promotes fear-like behaviors (e.g., mid-track aborts and slow approach to threat on the predator-inhabited foraging arena). Here, we argue that the underlying circuitry governing these two behavioral classes are not only computationally but neurophysiologically dissociable.

It is becoming increasingly clear that the hippocampus plays a central role in decision-making during avoid-approach conflict scenarios in both rodents and humans (Ito et al., 2016). The ventral hippocampus, but not the dorsal hippocampus, exhibits increased power in the theta range (4-10 Hz) during conflict in innately aversive contexts (Jacinto et al., 2016). In contrast, the dorsal hippocampus exhibits increased theta power during decisional uncertainty motivated by multiple competing tangible reinforcers such as is seen during both approach-approach conflict (Johnson and Redish, 2007) and avoid-approach conflict (Kim, 2015). Furthermore, there are marked differences between simultaneously recorded dorsal and ventral hippocampal theta power, frequency, and coherence during a place-response strategy switching task involving working memory and spatial planning (Schmidt et al., 2013). These data suggest that the dorsal and ventral hippocampus have non-overlapping roles in responding to uncertain environments and that their dynamics might be sensitive to specific forms of conflict (e.g., innate, contextual conflict versus tangible, external conflict) dictated by the cognitive demands of the task. Altogether, these data highlight theta power as a valuable marker for identifying task-dependent neural signatures during reward-based, threat-based, and conflict-based tasks. However, the larger question of how decisional conflict is represented and sequentially processed neurobiologically, from state-outcome encoding to outcome valuation and action selection, remains unclear.

During periods of conflict, we hypothesize that the dorsal hippocampus simulates states using its map of the task space as a substrate for spatial planning. State inferences and state-outcome contingencies have been shown to be represented in orbitofrontal and anterior cingulate cortices (Sharpe et al., 2015; Hillman and Bilkey, 2010; Cowen et al., 2012), and the ventral striatum and basolateral amygdala appear to be evaluating (and updating the stored value of) those contingency representations (Schoenbaum et al., 2003; Richard and Berridge, 2011; Sugam et al., 2014; Sharpe and Schoenbaum, 2016;

Zalocusky et al., 2016; Lichtenberg et al., 2017). In rats, the prelimbic and infralimbic cortices are thought to play an important role in recalling task-specific conditioned responses (CRs) for the maintenance of optimal behavior during probabilistic decision-making (St. Onge and Floresco, 2009; Zeeb et al., 2015), unlike orbitofrontal and anterior cingulate cortices which appear to be critical for learning the contingencies of a task and representing those environmental statistics for the purpose of cost-benefit calculations and conflict resolution (St. Onge and Floresco, 2009; Zeeb et al., 2015). Specifically, it has been argued that prelimbic ensembles are storing motor-inhibitory CRs (e.g., freezing) while infralimbic ensembles are storing motor-excitatory CRs (e.g., suppression of freezing), possibly through the use of a mixed selectivity encoding scheme which allows for a computationally efficient distributed representation of a multitude of task-relevant variables (Grunfeld and Likhtik, 2018). We suggest that it is this coordinated interaction between the dorsal hippocampus (state simulation), prefrontal cortices (contingencies and state-specific behaviors), and subcortical structures like the ventral striatum and basolateral amygdala (valuation) that underlies the cascade of representations triggered by decisional conflict.

Ethanol and diazepam: differences in pharmacokinetics and pharmacodynamics

Ethanol and diazepam both act as positive allosteric modulators at the GABA$_A$ receptor benzodiazepine binding-site. Unlike diazepam however, ethanol targets a variety of ion channel types and signaling systems (Crews et al., 1996; Lobo and Harris, 2008), providing a possible explanation for why ethanol did not affect choice-point hesitation in the same way as diazepam. For example, ethanol is known to disrupt the hypothalamic–pituitary–adrenal axis as well as

GABAergic, glutamatergic, opioidergic, and cholinergic neurotransmission (Deitrich et al., 1989), any one of which could explain the behavioral differences seen between the two anxiolytics in our data. The finding that ethanol and diazepam had an anxiolytic effect on mid-track aborts and slow approach to threat is likely because both these behaviors rely on Pavlovian systems implemented by structures like the central nucleus of the amygdala and the periaqueductal gray (LeDoux and Daw, 2018), both of which have been shown to be affected by ethanol and diazepam (Kang-Park et al., 2004; Jimenez-Velazquez et al., 2010; Roberto et al., 2012; Li et al., 2013). Interpreting our findings in the framework of our model, MTAs and slow approach to threat does not likely involve hippocampal-dependent state simulations whereas conflicted choice-point hesitation likely does.

This is supported by data showing an increase in the power of hippocampal theta oscillations during spatial planning (Johnson and Redish, 2007) in addition to the well-documented ability of diazepam to attenuate the power of hippocampal theta (Yeung et al., 2012). Therefore, we suggest that diazepam is likely impairing the rats' ability to utilize their cognitive map of the task space for spatial planning resulting in prolonged indecision and an increase in risk-taking behavior resulting from a compromised ability to represent potential future threat, both of which are consistent with our data (Fig. 4.4-5).

Conclusion

While both ethanol and diazepam attenuated proximal-threat fear behavior, diazepam exclusively increased distal-threat hesitation and risky decision-making. Taken together, these data suggest that ethanol and diazepam act on non-overlapping threat-processing circuits during avoid-approach conflict involving naturalistic threat and reward incentives. It is important for future research to be sensitive to the structure of the behavioral paradigms being used, how that structure influences which neural circuits are recruited to successfully navigate the task, and how that affects the generalizability of results obtained from tasks with differing behavioral and cognitive demands.

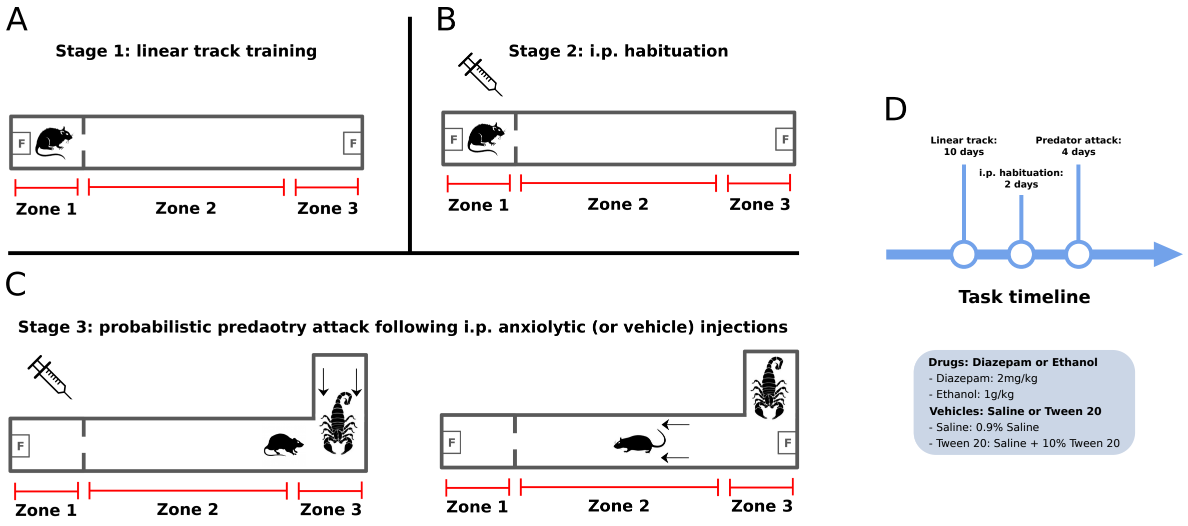

A
Stage 1: linear track training
F
F
Zone 1
Zone 2
Zone 3
B
Stage 2: i.p. habituation
F
F
Zone 1
Zone 2
Zone 3
C
Stage 3: probabilistic predaotry attack following i.p. anxiolytic (or vehicle) injections
F
Zone 1
Zone 2
Zone 3
F
F
Zone 1
Zone 2
Zone 3
D
Linear track: 10 days
i.p. habituation: 2 days
Predator attack: 4 days
Task timeline
Drugs: Diazepam or Ethanol
- Diazepam: 2mg/kg
- Ethanol: 1g/kg
Vehicles: Saline or Tween 20
- Saline: 0.9% Saline
- Tween 20: Saline + 10% Tween 20

Figure 4.1 Task design. A: In the linear track training stage, food-deprived rats learn to move from Z1→Z3→Z1 etc. to receive food at food ports denoted ‘F’. B: Same as in A with the exception that rats receive saline i.p. injections 5 min prior to session. C: A robotic predator is introduced to the arena. Now when the rat crosses from Z2→Z3 there is a 20% chance of the predator surging forward and attacking the rat. During these attack epochs, the rat typically freezes and retreats back to the nest-space (Z1) without retrieving the food in Z3. D: A timeline depicting the course of the experiment for each subject and the four drug conditions that were counterbalanced across the four attack days for each subject.

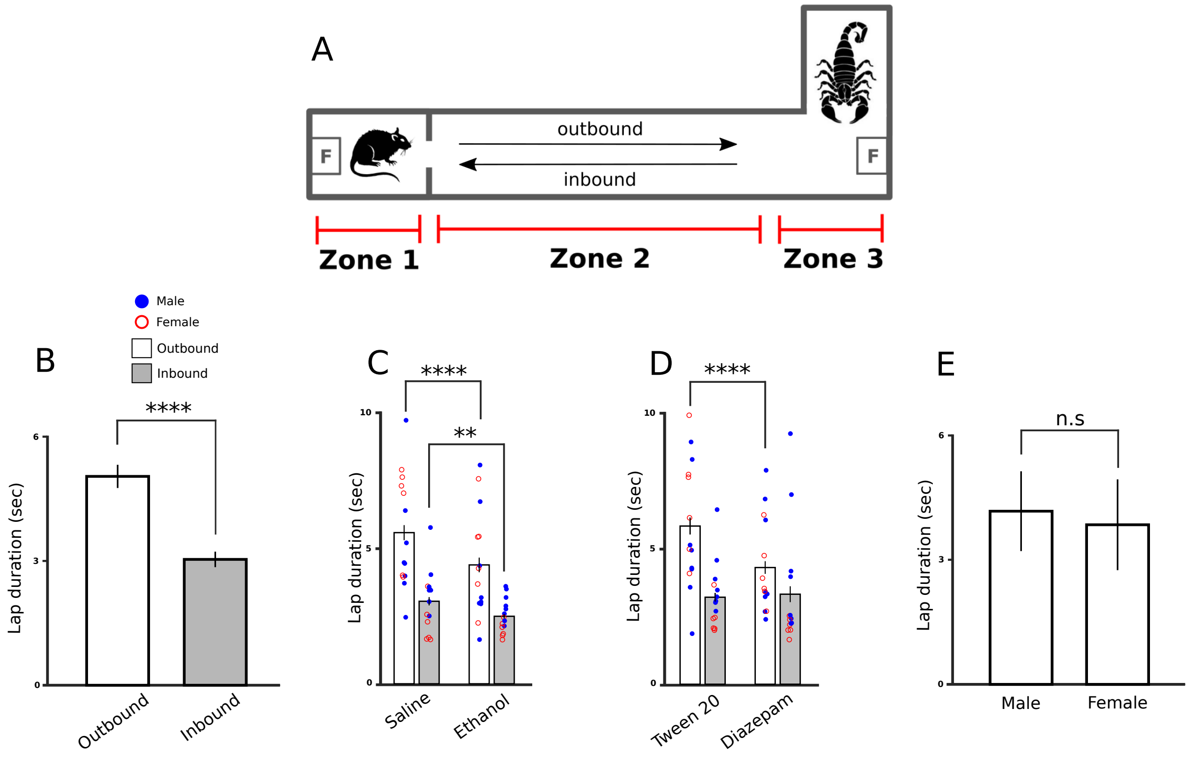

A
outbound
inbound
F
F
Zone 1
Zone 2
Zone 3
Male
Female
Outbound
Inbound
B
****
Lap duration (sec)
6
3
0
Outbound
Inbound
C
****
**
10
5
0
Lap duration (sec)
Saline
Ethanol
D
****
10
5
0
Lap duration (sec)
Tween 20
Diazepam
E
n.s
6
3
0
Lap duration (sec)
Male
Female

Figure 4.2 Inbound versus outbound lap duration. A: A schematic demonstrating the directionality of an inbound versus an outbound lap. B: Inbound laps were faster than outbound laps (matched pairs t-test, $t_{(13)} = 4.72$, ****p = 0.0002). C: Outbound laps were faster under ethanol than its vehicle control saline (matched pairs t-test, $t_{(13)} = 3.24$, ***p = 0.0032). Inbound laps were also found to be slightly faster under ethanol when compared to saline (matched pairs t-test, t(13) = 2.60, *p = 0.0218). D: Risky outbound laps were faster under diazepam than its vehicle control Tween 20 (matched pairs t-test, $t_{(13)} = 2.80$, **p = 0.0075), but no significant difference was found for inbound laps. E: There was not a significant difference in lap duration (outbound and inbound laps pooled) between males and females (two-sample t-test, $t_{(12)} = 1.29$, p = 0.2193). Blue circles = males, red circles = females. Data are mean ± s.e.m. *P < 0.05, **p < 0.01, ***p < 0.005, ****p < 0.001.

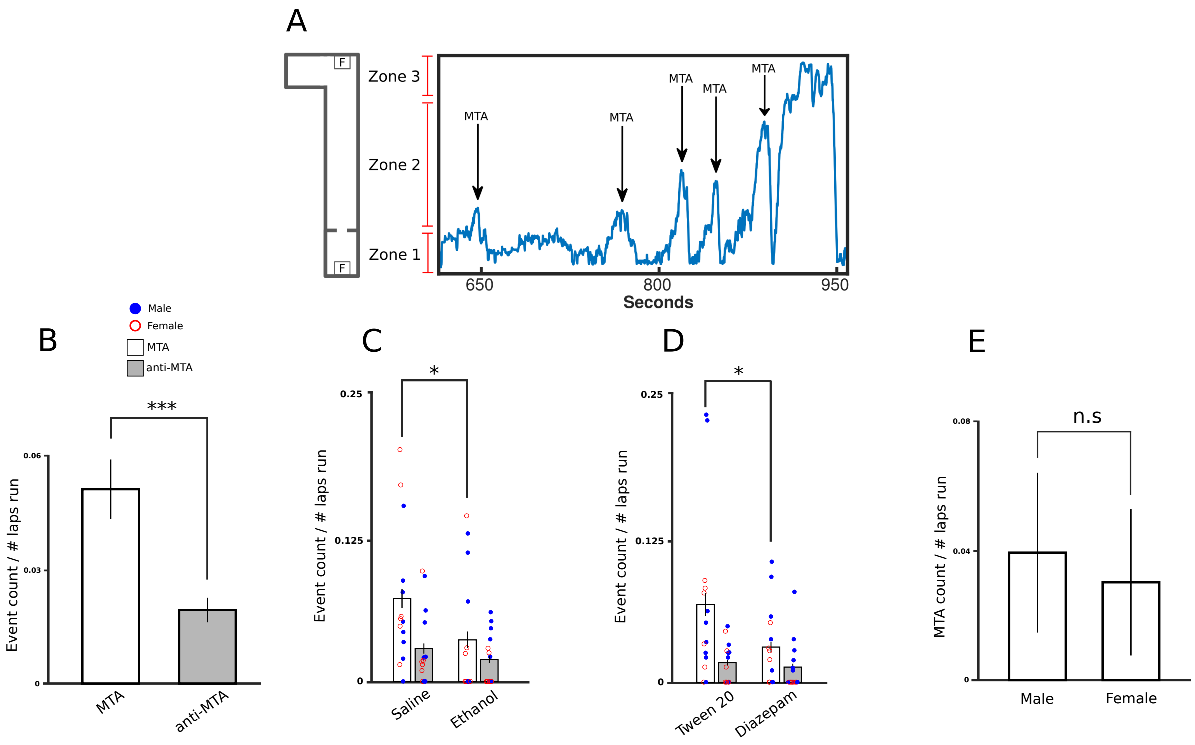

A
F
F
Zone 3
Zone 2
Zone 1
MTA
MTA
MTA
MTA
MTA
650
800
950
Seconds
Male
Female
MTA
anti-MTA
B
***
0.06
0.03
0
Event count / # laps run
MTA
anti-MTA
C
*
0.25
0.125
0
Event count / # laps run
Saline
Ethanol
D
*
0.25
0.125
0
Event count / # laps run
Tween 20
Diazepam
E
n.s
0.08
0.04
0
MTA count / # laps run
Male
Female

Figure 4.3 Mid-track ‘change of mind’ events. A: A representative trace of five consecutive mid-track aborts (MTAs) from one animal in the span of 300s. MTAs were operationalized as the rat leaving Z1, entering Z2, then returning back to Z1 without having entered Z3 to receive food (i.e., Z1→Z2→Z1). The anti-MTA control behavior measured when the rat left Z3, entered Z2, and returned to Z3 without having entered Z1 to receive food (i.e., Z3→Z2→Z3). B: MTAs were more likely to occur than their control behavior, anti-MTAs (matched pairs t-test, $t_{(13)} = 3.24$, ***p = 0.0032). C: Ethanol reduced the number of MTA events when compared against saline (matched pairs t-test, $t_{(13)} = 2.12$, *p = 0.0268). D: Diazepam reduced the number of MTA events when compared against Tween 20 (matched pairs t-test, $t_{(13)} = 2.17$, *p = 0.0245). E: There was not a significant difference in MTA count between males and females (two-sample t-test, $t_{(12)} = 0.57$, p = 0.5792).

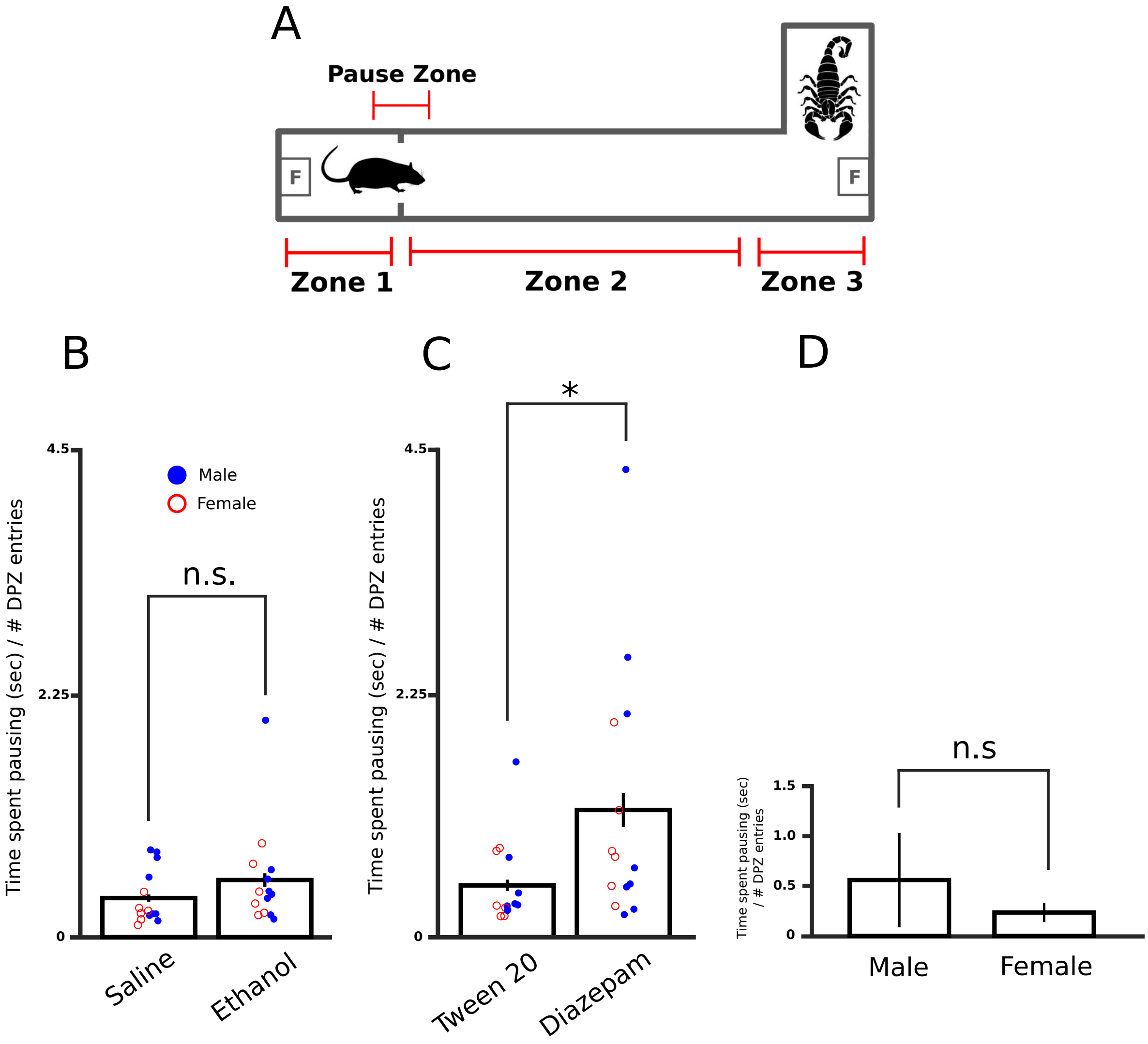

A
Pause Zone
F
F
Zone 1
Zone 2
Zone 3
B
C
D
Time spent pausing (sec) / # DPZ entries
4.5
2.25
0
● Male
○ Female
n.s.
Saline
Ethanol
Time spent pausing (sec) / # DPZ entries
4.5
2.25
0
*
Tween 20
Diazepam
Time spent pausing (sec) / # DPZ entries
1.5
1.0
0.5
0
n.s
Male
Female

Figure 4.4 Hesitation at the choice-point. A: Schematic of a rat pausing at the nest-space choice-point before either deciding to leave the nest and risk foraging for food or retreat back into the nest. B: There was no significant difference in deliberative pausing between ethanol and saline (matched pairs t-test, $t_{(13)} = 1.58$, p = 0.0689). C: Rats spent more time pausing at the choice-point under diazepam than under Tween 20 (matched pairs t-test, $t_{(13)} = 2.19$, *p = 0.0234). D: There was not a significant difference in deliberative pausing between males and females (two-sample t-test, $t_{(12)} = 0.94$, p = 0.3676).

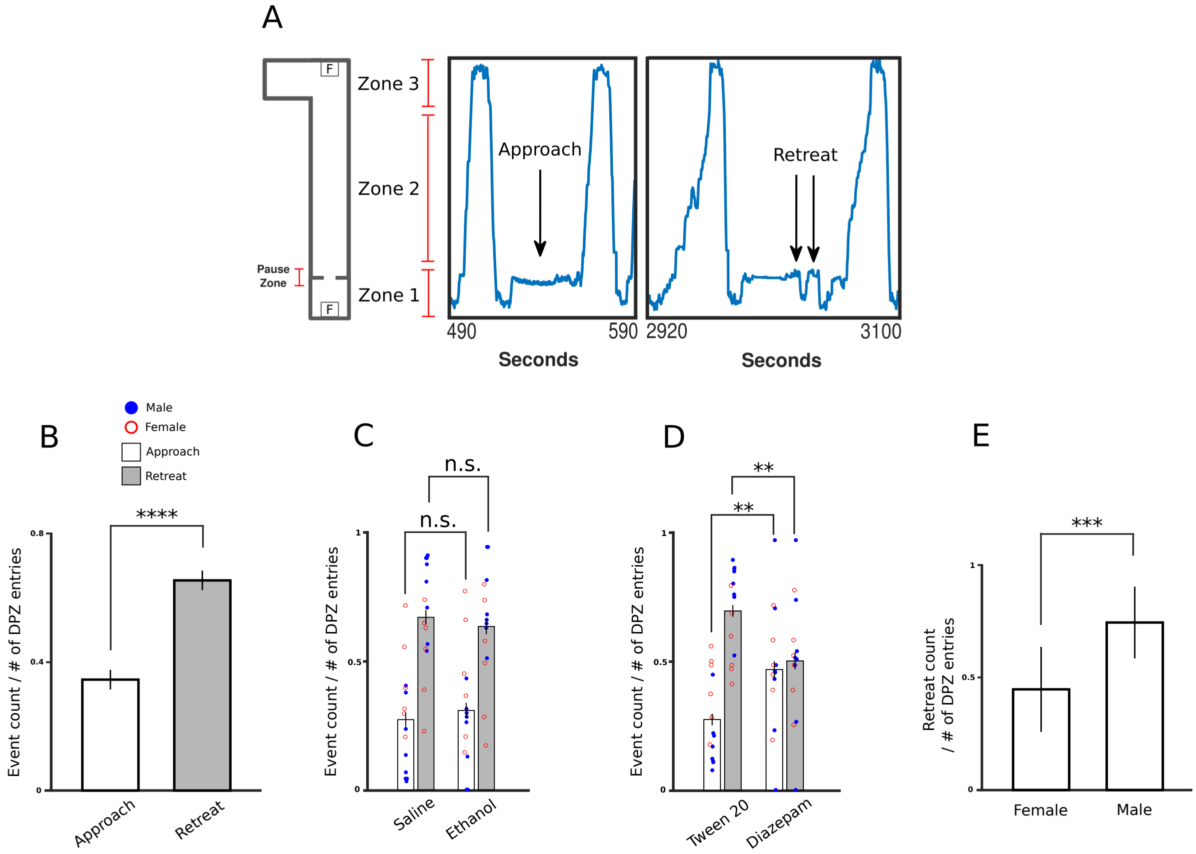
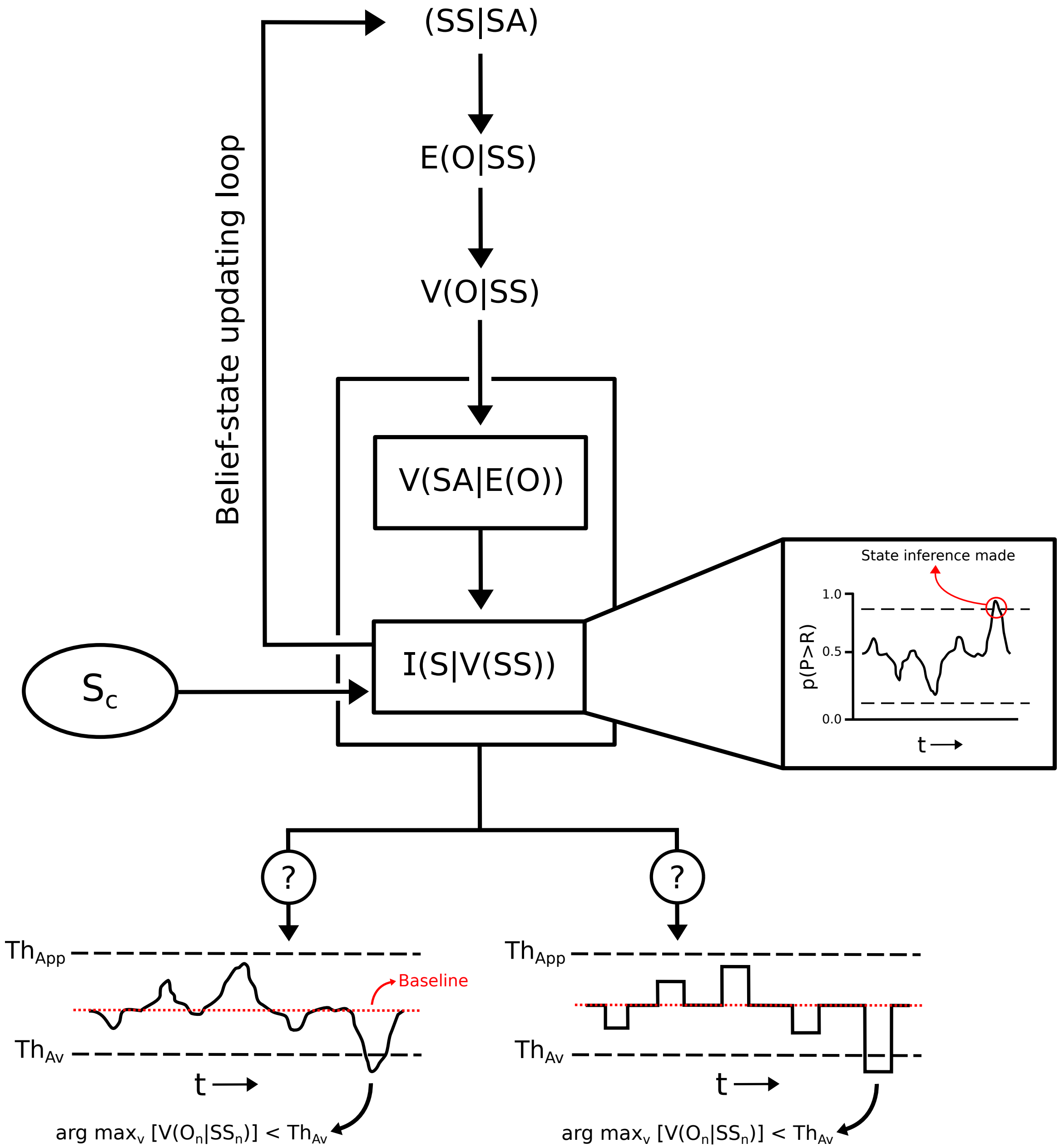

A
F
Zone 3
Zone 2
Pause Zone
Zone 1
F
Approach
Retreat
490
590
2920
3100
Seconds
Seconds
● Male
○ Female
□ Approach
▨ Retreat
B
****
Event count / # of DPZ entries
0.8
0.4
0
Approach
Retreat
C
n.s.
n.s.
Event count / # of DPZ entries
1
0.5
0
Saline
Ethanol
D
**
**
Event count / # of DPZ entries
1
0.5
0
Tween 20
Diazepam
E
***
Retreat count / # of DPZ entries
1
0.5
0
Female
Male

Figure 4.5 Pause-then-approach versus pause-then-retreat behavior. A: A representative trace of a pause-approach event (middle panel) and a pause-retreat event (right panel) obtained from one animal in the same session but at different time points in that session. B: After entering the nest doorway choice-point and pausing, rats were more likely to retreat than to approach (matched pairs t-test, $t_{(13)}$ = 4.46, ****p = 0.0006). C: There was no difference between ethanol and saline in the tendency to retreat or approach (matched pairs t-test, $t_{(13)}$ = 0.88, p = 0.8022). D: Rats were more likely to approach and less likely to retreat under diazepam when compared to vehicle control (matched pairs t-test, $t_{(13)}$ = 2.65, *p = 0.0100). E: Males are more likely than females to retreat following deliberation at the choice-point (two-sample t-test, $t_{(12)}$ = 3.46, ***p = 0.0047).



$$A(S_c)=\begin{cases} \text{approach} & \text{if } \arg\max_v [V(O_n|SS_n)] > Th_{App} \\ \text{hesitate} & \text{if } Th_{Av} < \arg\max_v [V(O_n|SS_n)] < Th_{App} \\ \text{avoid} & \text{if } \arg\max_v [V(O_n|SS_n)] < Th_{Av} \end{cases}$$

Figure 4.6 Markov model of risky decision-making. During conflict, agents simulate action-state transitions, the expected contingencies of those states, the values associated with those contingencies, and ultimately the values associated with performing the simulated action that lead to that simulated state (SS). If the V(A|E(O)) for a given SS fails to exceed a confidence threshold for making a state inference (see expanded view of I(S|V(SS)) which depicts threat-state monitoring, i.e., p(P>R) eventually passing a state inference threshold), the algorithm iterates through the belief-state updating loop until a state inference threshold is passed resulting in commitment to a decision. S_c = conflict state. S = state. A = action. SS = simulated state. SA = simulated action. E = expected. O = outcome. V = value. I = inferred. Th_{Av} = threshold for avoid decision. Th_{App} = threshold for approach decision.

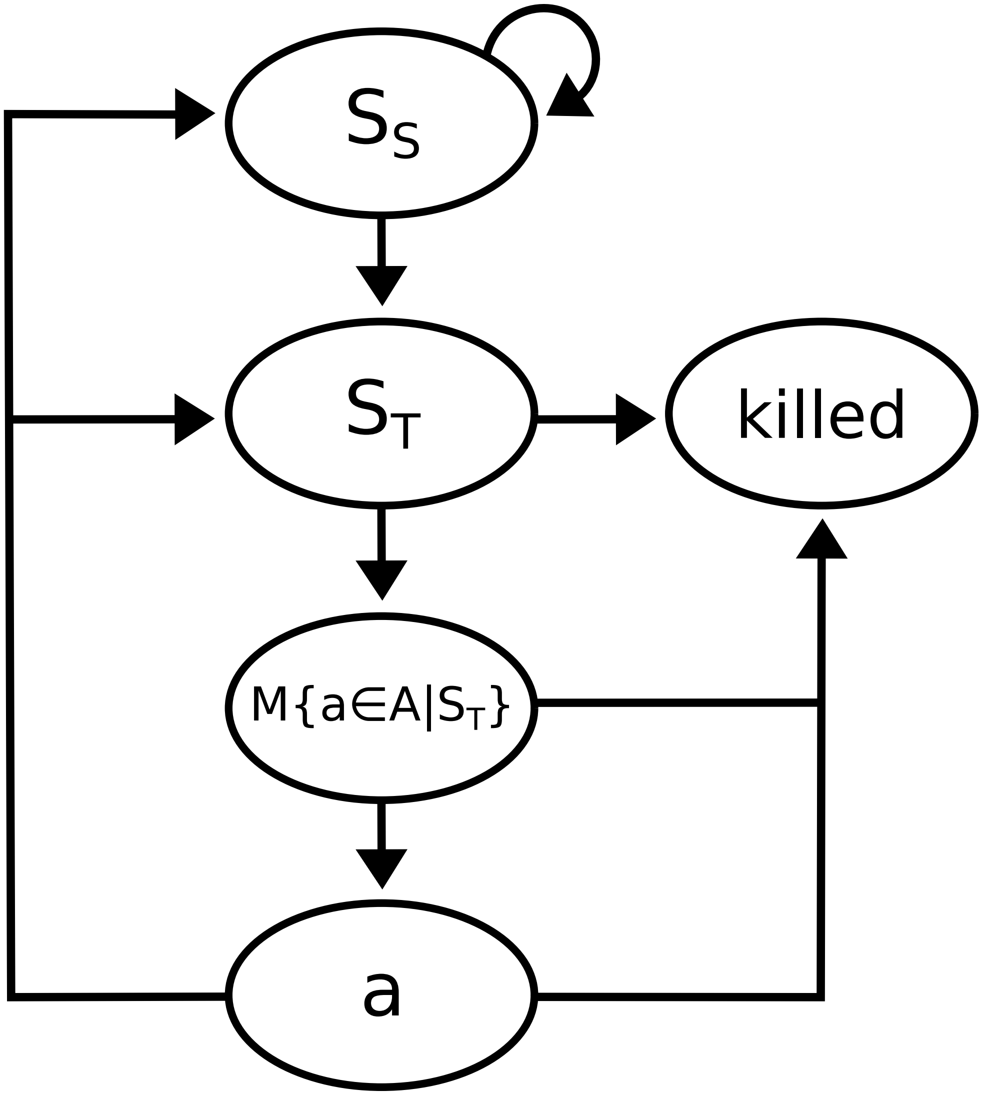

S_S
S_T
killed
M{a∈A|S_T}
a

Figure 4.7 Markov model of threat detection and evasion. The agent begins in a safe state (S_S) with no threat present. When the agent detects threat, it transitions from a safe state to a threat state (S_T). Following threat detection, the agent mobilizes for defensive behavior ($M\{a \in A|S_T\}$). An action (a) in the set of hard-wired, species-specific actions (A) is executed and the agent then transitions to one of three possible states: it is either captured and killed by the threat (killed), it remains in a state of threat, or it evades the threat and returns to a safe state. SS = safe state. ST = threat state. a = action. A = set of defensive behaviors. $a \in A$ = action that exists in the set of defensive behaviors. $M\{a \in A|S_T\}$ = mobilize for an action that exists in the set of defensive behaviors given the threat state.

— Original research —

Chapter 5

Dorsal hippocampus represents locations to avoid as well as locations to approach during avoid-approach conflict

Abstract

Various forms of anxiety are thought to involve mental representations of negative future outcomes, yet it remains unclear how the brain imagines these hypothetical scenarios. The hippocampus offers a candidate mechanism for resolving this dilemma. In the rat, hippocampal representations of space can be both local, reflecting the present location of the animal, and non-local, reflecting regions of space that the animal is not currently occupying. In reward-based decision-making, there are hippocampal sequences that extend from the location of the animal toward conflicting goal locations. These non-local 'fictive' representations are known to occur during approach-approach decision-making, but it is unknown whether there are similar hippocampal fictive representations during avoid-approach decision-making. We recorded hippocampal neural ensembles in rats behaving in an avoid-approach conflict

foraging task in order to investigate the role of negative future thinking in anxiety. We found distinct hippocampal fictive representations that co-occurred with two anxiety-like behaviors: (1) forward sequences during choice-point hesitation that shifted from representing reward in a safe environment to representing threat in a dangerous environment, and (2) discrete representations of threat during a change-of-mind behavior. Our results support the view that anxiety resulting from avoid-approach conflict involves representations of hypothetical scenarios, and that these fictive representations are, at least in part, neurally encoded in the hippocampus. These findings highlight hippocampal fictive representations as a potential target for the treatment of anxiety disorders.

Chapter taken from:
Walters, C. J. and Redish, A. D. (2021), 'Dorsal hippocampus represents locations to avoid as well as locations to approach during avoid-approach conflict', (Manuscript in preparation)

Introduction

Theories have long hypothesized that anxiety involves imagination, specifically the ability to 'mentally simulate' negative future outcomes. Indeed, the psychiatry literature has largely assumed this to be the case (Beck et al., 2005), positing that anxiety disorders result from aberrant cognitive schemas that distort one's beliefs and expectations about the future. Much theorizing has revolved around the notion of the 'prospective brain', a framework which emphasizes the importance of episodic future thinking in planning and goal-directed decision-making (Suddendorf and Corballis, 2007; Schacter et al.,2008). Behavioral data, for example, suggests that anxious individuals exhibit abnormalities in their episodic future thinking (MacLeod and Byrne,1996; Wu et al., 2015). Despite the prevalence of this cognitive framing of anxiety, neurophysiological studies have concentrated on the emotional dimension of anxiety while largely ignoring the role of prospection (Calhoon and Tye, 2015; Tovote et al., 2015). One reason for this lack of neurophysiological data on negative future thinking is that it has been unclear how to study prospection at the neural level.

Studies building on the spatial navigation literature suggest that the dorsal hippocampus encodes mental simulations of hypothetical scenarios during planning (Redish, 2016). Specifically, hippocampal representations of space can be both local, reflecting the present location of the animal, and non-local, reflecting regions of space that the animal is not currently occupying (Johnson and Redish, 2007; Johnson et al., 2009; Redish, 2016). In rodents, these dorsal hippocampal fictive representations have been implicated in approach-approach conflict (Johnson and Redish, 2007; Pfeiffer and Foster, 2013; Kay et al., 2020) as well as appetitive (Shin et al., 2019) and aversive (Wu et al., 2017) memory recall. In the human literature, the hippocampus has similarly been implicated in episodic future recall (Buckner and Carroll, 2007; Hassabis et al., 2007; Hassabis and Maguire, 2009; Martin et al., 2011) and reward-based decision-making (Peters and Büchel, 2010; Lebreton et al., 2013). However, it remains unclear what role, if any, hippocampal fictive representations play in threat-based avoid-approach conflict.

In non-human animals, there has been an emerging interest in ethologically-grounded tasks that allow for the study of anxiety in situations that more closely approximate natural conditions, thus providing access to more nuanced and potentially clinically-relevant manifestations of anxiety (Pellman and Kim, 2016; Paré and Quirk, 2017). One such task is the predator-inhabited foraging arena, an avoid-approach conflict task that requires a rat to forage for food in the presence of a hostile robotic predator designed to mimic the hazardous environmental conditions faced by rodents in the wild (Choi and Kim, 2010). The predator-inhabited foraging arena elicits a variety of anxiety-like behaviors in rats (Choi and Kim, 2010; Amir et al., 2015) such as choice-point hesitation (which is analogous to stretch-attend posture, a well-characterized risk assessment behavior (Grewal et al., 1997; Holly et al., 2016)) and mid-track retreat decisions (i.e., change-of-mind events), both of which are modulated by anti-anxiety drugs (Amir et al., 2015; Walters et al., 2019). Importantly, studies investigating human clinical anxiety have supported the translational validity of findings obtained from rodent conflict paradigms (Kirlic et al., 2017), suggesting that avoid-approach assays are well-suited for cross-species investigations of anxiety.

We sought to bring the anxiety, prospection, and ethology literatures together in order to interrogate the neural basis of negative future thinking. To do this, we recorded neural activity from dorsal hippocampal ensembles as hungry rats freely behaved in the predator-inhabited foraging arena to determine whether hippocampal fictive representations co-occur with anxious behaviors. We found both sequential and discrete non-local hippocampal representations that co-occured with anxiety-like behaviors. These findings suggest that the dorsal hippocampus shifts from conveying reward-related representations in safe environments to reward- and threat-related representations in dangerous environments. This implies that the hippocampus plays a role in negative future thinking, thus providing a novel neural mechanism that could underlie a long-hypothesized psychological description of anxiety.

RESULTS

We used a task that requires rats to confront a hostile robotic predator in order to obtain food. We trained four Brown Norway rats to navigate in this task while we recorded from layer CA1 of the dorsal hippocampus using 24-tetrode microdrive arrays. Rats were food-deprived to 80% of their free-feeding weight and had 1 hr a day in the arena to forage for food. The experiment had three phases: linear track training (Fig. 5.1A), robot present (Fig. 5.1B), and robot attack (Fig. 5.1C). During linear track training sessions, rats had to shuttle between Zone 1 (Z1) and Zone 3 (Z3), both of which contained a feeder site, in order to receive food pellets. The Z1 feeder site was primed to deliver a food pellet only after the rat visited the Z3 feeder site (and vice versa). After rats learned to navigate the linear track, an immobile robot was introduced near the

Z3 feeder site (robot present phase, Fig. 5.1B). This was done to habituate rats to the presence of the robot.

These training sessions (Fig. 5.1A-B) continued until rats ran enough laps to start gaining weight (80-100 laps per 1 hr session). Once this criterion was met, rats graduated to the attack sessions (Fig. 5.1C-E). Attack sessions were identical to phase two (Fig. 5.1B) except that once rats crossed the unmarked threshold separating Z2 from Z3, there was a 20% probability that the robot would aggressively surge forward toward the rat while contracting its pincers and making clicking noises (Fig. 5.1D). Following these mock attacks, rats typically fled to Z1 (Fig. 5.1E) and exhibited a variety of anxiety-like behaviors for the remainder of the session (Choi and Kim, 2010; Amir et al., 2015). We investigated the neurophysiological underpinnings of the anxiety-like behaviors (choice-point hesitation, Fig. 5.1F; and mid-track abort, Fig. 5.1G) seen during attack sessions following confrontation with the robotic predator.

We decoded hippocampal population activity using 50 millisecond time bins during choice-point hesitation and observed sequential non-local representations extending toward the Z3 feeder site (Fig. 5.2A, C) and the attack zone (Fig. 5.2B, C). These sequences occurred during periods of high sharp-wave ripple power (Fig 2F, G). Interestingly, these sharp-wave ripple sequences terminated in the attack zone more frequently during post-attack choice-point hesitation relative to pre-attack choice-point hesitation (Fig. 5.2C). Rats also spent more time hesitating at the choice-point post-attack (Fig. 5.2D; p<0.001, Wilcoxon rank sum), and there were more sharp-wave ripple events post-attack relative to pre-attack (Fig. 5.2E; p<0.001, Wilcoxon rank sum). These results are consistent with previous work showing that ripple-associated sequences terminate at aversive locations in the environment (Wu et al., 2017). However, in this experiment we were able to observe a shift away from the noisier pre-attack sharp-wave ripple sequences that favored the reward zone (Fig. 5.2C, blue) toward the less noisy post-attack representations that primarily ended in either the reward zone or the attack zone (Fig. 5.2C, red), thus reflecting the avoid-approach conflict. Altogether, these data suggest that there is a shift away from noisy non-local representations of reward and toward non-local representations of both reward and threat during choice-point hesitation following aversive encounters with the robotic predator.

In predator-inhabited environments, rats more slowly approach a risky food source than a safe food source (Walters et al., 2019), suggesting lap speed as a useful behavioral metric of anxious conflict. We found that outbound lap speed was slower during post-attack laps relative to pre-attack laps (Fig. 5.3A; p<0.001, Wilcoxon rank sum), whereas inbound lap speed was less affected (Fig. 5.3B; p=0.041, Wilcoxon rank sum). Furthermore, there was an overrepresentation of the attack zone on outbound laps post-attack relative to pre-attack (Fig. 5.3C-D). This difference was significant both as rats approached the attack zone (Fig. 5.4D, p=0.016, Wilcoxon rank sum) and as rats traveled through the attack zone (Fig. 5.4C, p=0.013, Wilcoxon rank sum). Interestingly, this difference was significant only when pre-attack and post-attack were temporally aligned (i.e., lap-aligned to the same point in time) but not spatially aligned (i.e., taking into account that rats approached Z3 more slowly on post-attack laps relative to pre-attack laps, and so spent more time in and around the attack zone on post-attack outbound laps) (Fig. 5.4), suggesting that the reason rats overrepresented the attack zone on outbound laps post-attack was because of their behavioral reluctance to approach the robotic predator.

Consistent with previous experiments (Amir et al., 2015), rats exhibited a change-of-mind behavior on a subset of outbound laps during which they turned around mid-journey and returned to the nest space (Fig. 5.1G, Fig. 5.5A). Interestingly, there were non-local representations of the attack zone at the moment rats performed these mid-track aborts (MTAs; Fig. 5.5A, D). As a control for MTAs (Z1→Z2→Z3), we examined the number of times rats performed anti-MTAs (Z3→Z2→Z3), the identical behavior as MTAs but in the opposite direction (i.e., leaving the predator-inhabited feeder site, entering the middle of the track, and then returning to the predator-inhabited feeder site without obtaining food in the nest space). Consistent with prior experiments (Walters et al., 2019), we found that MTAs occurred more frequently than anti-MTAs (Fig. 5.5B; $p<0.001$, Wilcoxon rank sum), suggesting that MTAs were a behavioral marker of anxiety and not just random reorientations. Additionally, MTAs were more common during attack sessions relative to non-attack sessions (Fig. 5.5C; $p=0.002$, Wilcoxon rank sum), suggesting that outbound laps were more anxiogenic when the foraging environment was unsafe. These results suggest that MTAs are a behavioral readout of anxious conflict, and that at the moment rats change their minds and forgo food in favor of safety there are hippocampal non-local representations of threat.

DISCUSSION

Anxiety has long been hypothesized to involve negative beliefs about the future. Indeed, individuals with anxiety disorders have a negativity bias when imagining future scenarios (MacLeod and Byrne, 1996; Wu et al., 2015). This begs the question: what are the neural mechanisms that support negative future thinking? Recent work has underscored the hippocampus as an important component of imagination and prospective cognition. While the hippocampus is thought to encode a cognitive map of the environment that serves as a substrate for planning trajectories through space (O'Keefe and Nadel, 1978; Redish, 2016), research in this domain has largely focused on reward-based planning; as a result, little is known about how threat-based planning occurs and whether it involves similar neuronal dynamics. Previous evidence has implicated goal-directed hippocampal fictive representations in approach-approach conflict. Our data find that similar hippocampal fictive representations exist during avoid-approach conflict, that they correspond to both threat and reward stimuli, and that the representations of threat develop after the animal learns that the foraging arena is unsafe.

Theta-associated hippocampal sequences have been shown to project toward appetitive locations on reward-based tasks (Johnson and Redish, 2007; Pfeiffer and Foster, 2013; Shin et al., 2019; Kay et al., 2020), while ripple-associated hippocampal sequences have been shown to project toward aversive locations on threat-based tasks (Wu et al., 2017). However, natural environments are complex and contain a variety of both appetitive and aversive stimuli that must be accounted for in order to inform optimal behavior. It has remained unclear how the hippocampus represents and arbitrates between conflicting goals of opposing valence. Addressing this gap in the literature, we found ripple-associated hippocampal sequences during choice-point hesitation that shift from noisy representations of reward in a safe environment to less noisy representations of both reward and threat after learning that the environment is unsafe. Additionally, we discovered threat-related non-local representations appearing in theta sequences that co-occurred with a risk-averse change-of-mind behavior.

Kim et al. showed that pharmacological inhibition (or excitation) of the amygdala increased (or decreased) risky approach decisions in the predator-inhibited foraging arena task (Choi and Kim, 2010). Using the same task, Kim et al. observed that post-attack place cell remapping occurs at locations near the robotic predator (Kim et al., 2015) and that this remapping was amygdala-dependent. These data suggest that aversive stimuli disrupt hippocampal representations of space, and that this disruption is mediated by the amygdala. Consistent with Kim et al., our data suggest that there is a trend toward place field instability in place cells on the robot side of the track when comparing the first half of each session to the second half (Kim et al., 2015) (Fig. 5.6; p=0.064, Wilcoxon rank sum). Additionally, we observed a shift toward non-local hippocampal representations of threat after the environment transitions from safe to unsafe. Amir et al. found two populations of basolateral amygdala neurons that exhibited behaviorally-relevant activity patterns: one population that ramped up in firing rate as rats hesitated at the choice-point prior to approach decisions, and another population that ramped up as rats hesitated at the choice-point prior to retreat decisions (Amir et al., 2015). An intriguing open question is whether these amygdala signals relate to the hippocampal sequences we have identified during choice-point hesitation. For instance, are there aversive (or appetitive) representations in the amygdala during conflict that co-occur with the hippocampal sequences, thus serving to infuse the fictive representations of space with negative (or positive) value? If so, how does the amygdala-mediated salience landscape interact with the hippocampal-mediated cognitive map, and do these amygdala representations of value appear before, during, or after the non-local hippocampal representations?

Threat processing frameworks have long proposed that there is a threat gradient strategy that human and non-human animals use to successfully navigate threatening situations (Fanselow et al., 1988). This gradient is a continuum that spans from most safe (no threat present) to least safe (active predatory pursuit), with intermediate stages corresponding to varying degrees of vulnerability and threat visibility. Each stage in this hypothesized survival algorithm has associated behaviors ranging from mental simulation, foraging, and environment building to defensive behaviors such as freezing, fleeing, and fighting. Fear and anxiety are differentiated along this threat continuum, with fear behaviorally corresponding to strategies for managing proximal and immediate threats (freezing, fleeing, and fighting) while anxiety behaviorally corresponds to strategies for managing distal and potential threats (passive avoidance, niche construction, and mental simulation) (Mobbs et al., 2015). Interestingly, our data support that mental simulations occur for both distal (i.e., the ripple-associated hippocampal sequences during choice-point hesitation) and proximal (i.e., the theta-associated attack zone representation prior to change-of-mind events), thus suggesting a direct role for mental simulations of threat during both passive and active avoidance. Our data therefore both complement and serve to further refine the gradient model of threat processing.

Finally, our data have potential implications for our understanding of anxiety and its various disorders. The function of anxiety is to preemptively mobilize for threat, and the ability to mentally simulate fictive scenarios confers the ability to weigh the pros and cons associated with taking specific actions (Beck et al., 2005; Schacter et al., 2008; Redish, 2016; Miloyan et al., 2016; Heller and Bagot, 2020). However, if these mental simulations become excessive or skewed toward simulating aversive fictive outcomes, then they could result in some of the defining symptoms of many anxiety disorders such as indecision, rumination, and negativity bias (Beck et al., 2005). Indeed, individuals with generalized anxiety disorder are known to possess a negativity bias in their episodic future thinking (Wu et al., 2015). An interesting question for future research will be to determine whether the fictive hippocampal representations reported here occur excessively or are negatively biased in animal models of anxiety disorders. Importantly, non-human animal avoid-approach conflict tasks have been shown to be translationally relevant (Kirlic et al., 2017). Many avoid-approach conflict paradigms carried out in non-human animals have been adapted for and validated in human subjects at both the behavioral and neurophysiological level, highlighting these assays as a powerful tool to better understand the neural basis of human anxiety disorders (Kirlic et al., 2017). In line with this, there has been a recent surge in studies investigating the neural basis of anxiety in human subjects using avoid-approach conflict foraging tasks involving artificial predators (Mobbs et al., 2009; Qi et al., 2018; Bach et al., 2019; Fung et al., 2019; Korn and Bach, 2019; Abivardi et al., 2020). An exciting direction for future research will be to continue translating such semi-naturalistic anxiety paradigms across species and identifying areas of overlap and divergence in terms of the behavioral and neurophysiological underpinnings of anxiety in healthy and clinical populations.

MATERIALS AND METHODS

Subjects

Four Brown Norway rats aged 8-10 months served as the experimental subjects (2 male, 2 female). All rats were maintained on a 12:12 hr light/dark cycle. Rats were food-restricted such that they had 1 hr per day to work for food pellets (45 mg each, Test Diet, Richmond, IN) in the foraging arena. Rats were always kept above 80% free-feeding weight and had unlimited access to water outside of the foraging arena. All procedures were approved by the University of Minnesota (UMN) Institutional Animal Care and Use Committee (IACUC) and were performed in accordance with NIH guidelines.

Foraging arena task procedure

The foraging arena was 1.16 m long and 33 cm wide with walls 60 cm tall. An overhead video camera tracked animal position from a head-affixed LED at 30 fps.

There were three phases of the predator-inhabited foraging arena: linear track training, robot habituation, and attack sessions (see Fig. 5.1A-C). During all three phases of the task, sessions lasted 1 hr and rats began each session in the nest-space.

Phase 1: During the linear track phase, rats learned to shuttle back-and-forth to receive food pellets at either end of the track. Food pellets were only delivered at one of the feeders once the other feeder had been visited; thus, rats needed to alternate between feeders to continually receive food. One end of the linear track had a partially enclosed “nest-space”, a high-walled room continuous with the linear track via an open door-way referred to as the “choice-point”.

Phase 2: The robot habituation phase was identical to Phase 1 with the exception that the robotic predator (SPIK3R, LEGO® MINDSTORMS® EV3) was placed next to the Z3 feeder site. Following 2 days of robot habituation, rats proceeded to Phase 3 of the task.

Phase 3: For the first 15 laps during Phase 3 sessions, the robot remained stationary. After the first 15 laps, the robotic predator would surge forward and attack the foraging rat with a 20% probability (i.e., on any given lap there was a 1/5 chance of the robot attacking the rat) as it approached the feeder site adjacent to the predator.

Pre-attack analyses included portions of Phase 3 sessions prior to robot attack as well as control sessions after Phase 3 where there was either a LEGO® pyramid or the robotic predator (which was turned off and thus in a non-attack setting) situated near the Zone 3 feeder site. Post-attack analyses included portions of Phase 3 sessions following the first robot attack.

<u>Surgery and electrode targets</u>

Rats had ad libitum access to food (Teklad pellets) for at least 3 days prior to surgery. Rats were then chronically implanted with a hyperdrive (built in-house) containing 24 individually moveable tetrodes. Hyperdrives consisted of two bundles (12 tetrodes each) targeting dorsal hippocampal layer CA1 bilaterally (3.8 mm posterior and ± 3.0 mm lateral from bregma). Protective shrouds were printed on a Form 2 3D printer (Formlabs, Somerville, MA) to cover the drives and amplifier boards.

Rats were anesthetized throughout the duration of the surgery (0.5–2% isoflurane vaporized in medical-grade O^2 via nosecone). Rats were placed in a stereotaxic instrument (Kopf, Tujunga, CA) and given carprofen (Rimadyl) subcutaneously and penicillin (Combi-Pen-48) intramuscularly. 3–5 jewlers' screws were used to anchor the drives to the skull, one of which was used as ground for the tetrodes.

Two craniotomies were opened for the bilateral tetrode bundles using a surgical trephine. Dura was removed using forceps, and the tetrode drives were positioned using the stereotax. Silicone gel (Dow Corning, Midland, MI) was applied to the craniotomies after the bundles were positioned at the surface of the brain. We applied a layer of Metabond (Parkell, Edgewood, NY) followed by a layer of dental acrylic (The Hygenic Corporation, Cuyahoga, OH) to secure the drives to the skull. After surgery, all tetrodes were lowered by 640 μm. Post-surgery, rats were subcutaneously injected with carprofen for 3 days as well as enrofloxacin (Enroflox) for 6 days. Rats began behavioral training 5 days after hyperdrive surgery. A subset of the cohort (n=2; 1 male, 1 female) also had a silicon probe implanted in prelimbic cortex. The prelimbic cell yields were not sufficiently large to warrant analysis, however, and as a result that data is not included in this manuscript.

Histology

After the last recording session was obtained from each rat, tetrode recording locations were marked with electrolytic lesions. A 10 μA was passed through a channel on each tetrode for 10 s. At least two days after the lesions were made, rats were anesthetized with a pentobarbital sodium solution (150 mg/kg, Fatal-Plus) and then perfused transcardially with saline followed by 10% formalin. Brains were stored in a 30% sucrose formalin solution until slicing. Coronal slices were made through the hippocampus of 3 rats (with sagittal slices being made in 1 rat) using a cryostat, and the slices were stained with cresyl violet and imaged to determine tetrode placement (Fig. 5.7, aligned to Paxinos and Watson, 2006).

Statistics and data analysis

All data were processed and statistically analyzed in MATLAB (MathWorks, Natick, MA). Unless stated otherwise, all figures depict the mean ± s.e.m.

Behavioral quantification

Mid-track aborts were defined as anytime the rat left the nest space (i.e., Zone 1), entered the middle of the track (i.e., Zone 2), and then turned around and re-entered the nest space without having activated the Zone 3 feeder site. Choice point hesitation was defined as anytime the rat entered the nest space doorway and remained stationary there for >2 s. Lap speed was quantified as the time it took a rat to traverse Zone 2 (i.e., how long it took to reach Zone 3 after leaving Zone 1 or vice versa).

Bayesian decoding

A Bayesian approach (Zhang, Ginzburg, McNaughton, & Sejnowski, 1998) was used to decode spatial information represented by hippocampal layer CA1 ensembles. For the decoding, we segmented the track into a 1x24 grid and used 50 ms time bins. In brief, the aim of Bayesian decoding is to use the instantaneous activity (i.e., the neuronal firing rate over 50 ms increments) of the neural ensemble to assign a probability to each position in space (each of the 24 bins along the length of the track) that reflects the extent to which the ensemble spiking is representing that location. Applying Bayes' rule to this encoding model, we can generate a mapping from ensemble spiking patterns to a decoded spatial position (i.e., one of the 24 spatial bins). We took the decoded position for each 50 ms time increment to be the spatial bin with the largest posterior probability.

Spectral analysis and sharp-wave ripple detection

Spectral analyses were performed using the MATLAB 'spectrogram' function. The spectrogram parameters were a 1 s window, an overlap window of 250 ms, and a frequency range from 1-300 Hz. Power spectral density plots were generated by temporally averaging spectrograms. Frequency-by-frequency power correlation plots (Masimore et al., 2004) were generated by correlating across spectra within a given spectrogram. To detect sharp-wave ripples, the local field potential was band passed in the 140-200 Hz range. If the filtered local field potential signal exceeded a 4 standard deviation threshold, it was classified as a sharp-wave ripple complex.

Quantifying hippocampal sequences

Sequences were calculated by determining the location of the largest posterior probability for each spatial bin. Using a 0.15 posterior threshold, we assessed whether there were 3 decoding bins where that threshold was surpassed. If the largest posterior exceeded the threshold in all 3 bins, and each bin's largest posterior was in front of the previous bin's largest posterior (with the first bin's largest posterior being in front of the true position of the rat, the second bin's largest posterior being in front of the first bin's largest posterior, and so on), we defined that 150 ms segment as a sequence.

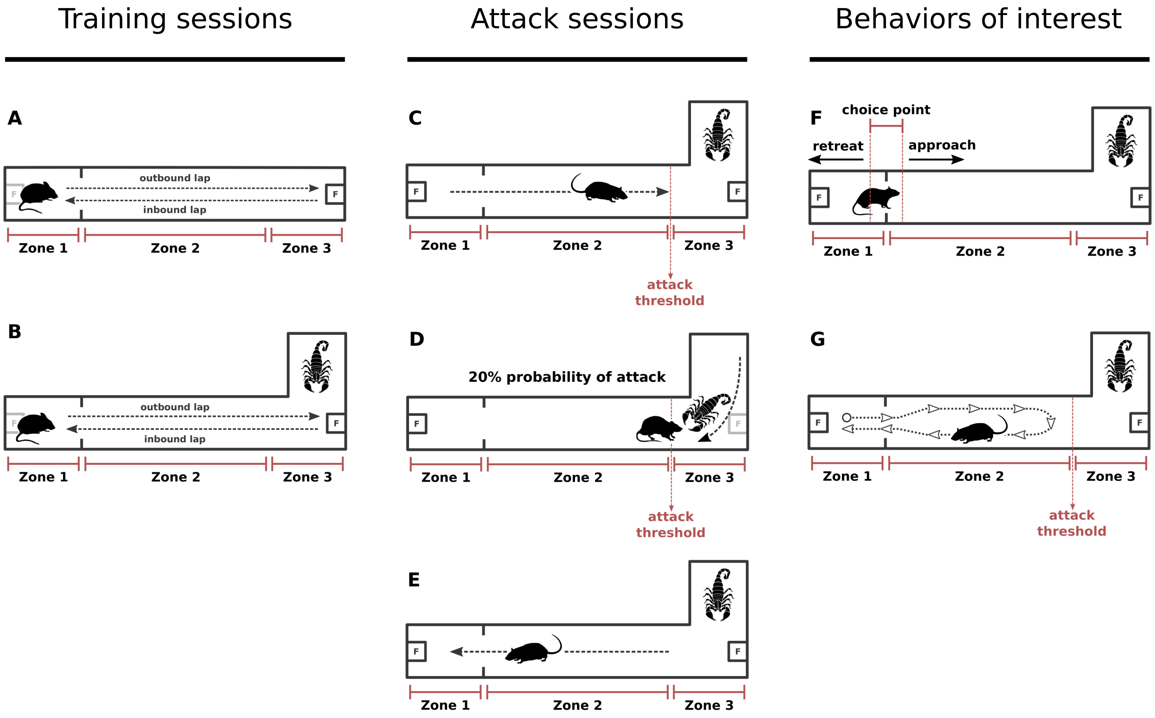

Training sessions
Attack sessions
Behaviors of interest
A
outbound lap
inbound lap
F
F
Zone 1
Zone 2
Zone 3
B
outbound lap
inbound lap
F
F
Zone 1
Zone 2
Zone 3
C
F
F
Zone 1
Zone 2
Zone 3
attack
threshold
D
20% probability of attack
F
F
Zone 1
Zone 2
Zone 3
attack
threshold
E
F
F
Zone 1
Zone 2
Zone 3
F
choice point
retreat
approach
F
F
Zone 1
Zone 2
Zone 3
G
F
F
Zone 1
Zone 2
Zone 3
attack
threshold

Figure 5.1 Task design

(**A**) During training sessions, rats shuttle back-and-forth between feeder sites on opposite ends of a linear track. (**B**) After the rat reaches behavioral criteria on the linear track, an immobile robotic predator is introduced in Zone 3 (Z3) to habituate rats to its presence. (**C** to **E**) After two sessions of robot habituation, rats transition to the attack phase. During attack sessions, the robot attacks the foraging rat with a 20% probability once it transitions from Z2→Z3. Following predatory attack, rats typically flee to the Z1 nest space. (**F**, **G**) Rats then exhibit choice-point hesitation behavior prior to either retreating back into the nest space or risking a foraging attempt. After being attacked, rats will enter the doorway and hesitate (**F**), or, on a subset of approach decisions, change their mind mid-approach and retreat back to the nest space before reaching the Z3 attack threshold (**G**). These mid-track abort decisions are defined as any time the rat leaves Z1, enters Z2, then returns back to Z1 without having entered Z3 (i.e., Z1→Z2→Z1).

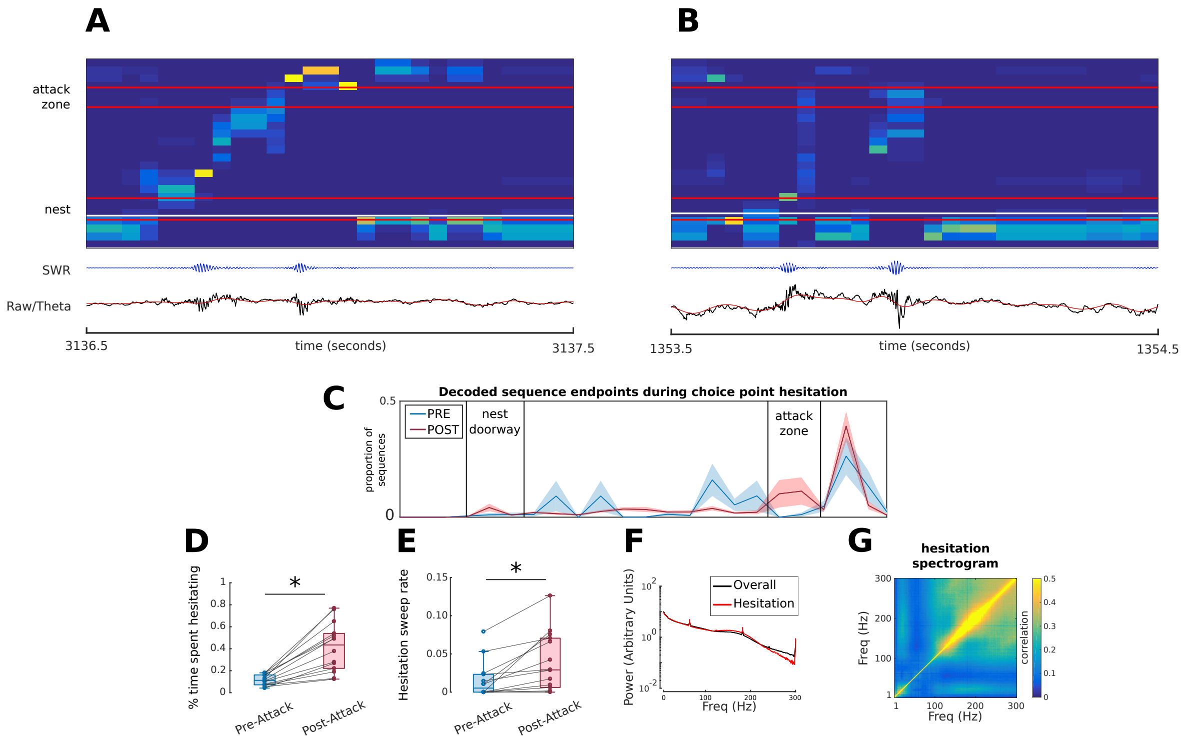

A
B
attack
zone
nest
SWR
Raw/Theta
3136.5
time (seconds)
3137.5
1353.5
time (seconds)
1354.5
C
Decoded sequence endpoints during choice point hesitation
0.5
0
proportion of
sequences
PRE
POST
nest
doorway
attack
zone
D
% time spent hesitating
1
0.8
0.6
0.4
0.2
0
*
Pre-Attack
Post-Attack
E
Hesitation sweep rate
0.15
0.10
0.05
0
*
Pre-Attack
Post-Attack
F
Power (Arbitrary Units)
10^2
10^0
10^{-2}
0
100
200
300
Freq (Hz)
Overall
Hesitation
G
hesitation
spectrogram
Freq (Hz)
300
200
100
1
1
100
200
300
Freq (Hz)
correlation
0.5
0.4
0.3
0.2
0.1
0

Figure 5.2 Forward sequences during choice-point hesitation

(**A**) Representative example of a sharp-wave ripple sequence extending from the position of the rat in the choice-point (white line) to the Z3 feeder. Decoding (top panel) and local field potential (bottom panel; 140-200 Hz bandpass in blue, 5-10 Hz bandpass in red, raw local field potential in black). (**B**) Representative example of a sharp-wave ripple sequence extending from the position of the rat in the choice-point (white line) to the attack zone. Decoding (top panel) and local field potential (bottom panel). (**C**) Proportion of forward sequences both pre- and post-attack (blue and red, respectively). (**D**) Percent of time spent hesitating in the choice-point pre- and post-attack ($p<0.001$, Wilcoxon rank sum). (**E**) Hesitating sequence rate in the choice-point pre- and post-attack ($p<0.001$, Wilcoxon rank sum). (**F**) Power spectral density during choice-point hesitation. (**G**) Frequency-by-frequency power correlation during choice-point hesitation. Note the strong sharp wave power and correlation signals indicating that these hesitation sequences are related to sharp wave ripple complexes.

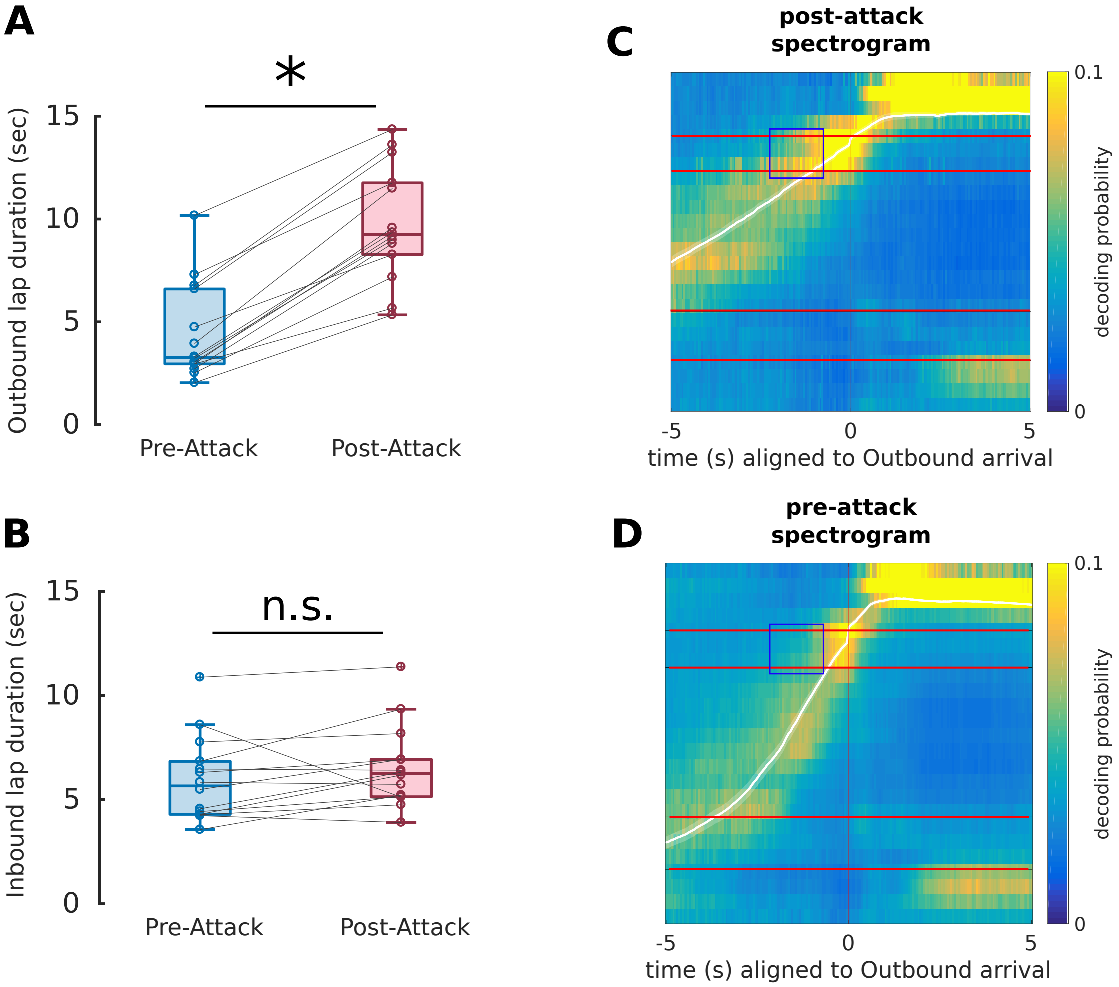

A
*
Outbound lap duration (sec)
15
10
5
0
Pre-Attack
Post-Attack
B
n.s.
Inbound lap duration (sec)
15
10
5
0
Pre-Attack
Post-Attack
C
post-attack spectrogram
0.1
0
decoding probability
-5
0
5
time (s) aligned to Outbound arrival
D
pre-attack spectrogram
0.1
0
decoding probability
-5
0
5
time (s) aligned to Outbound arrival

Figure 5.3 After being attacked, rats are slower on outbound laps and exhibit heightened attack zone representations

(**A**) Outbound lap duration in seconds (p<0.001, Wilcoxon rank sum). (**B**) Inbound lap duration in seconds (p=0.041, Wilcoxon rank sum). (**C**) Averaged post-attack outbound lap decoding. (**D**) Averaged pre-attack outbound lap decoding. Note the increased attack zone decoding (blue box) relative to the pre-attack outbound laps.

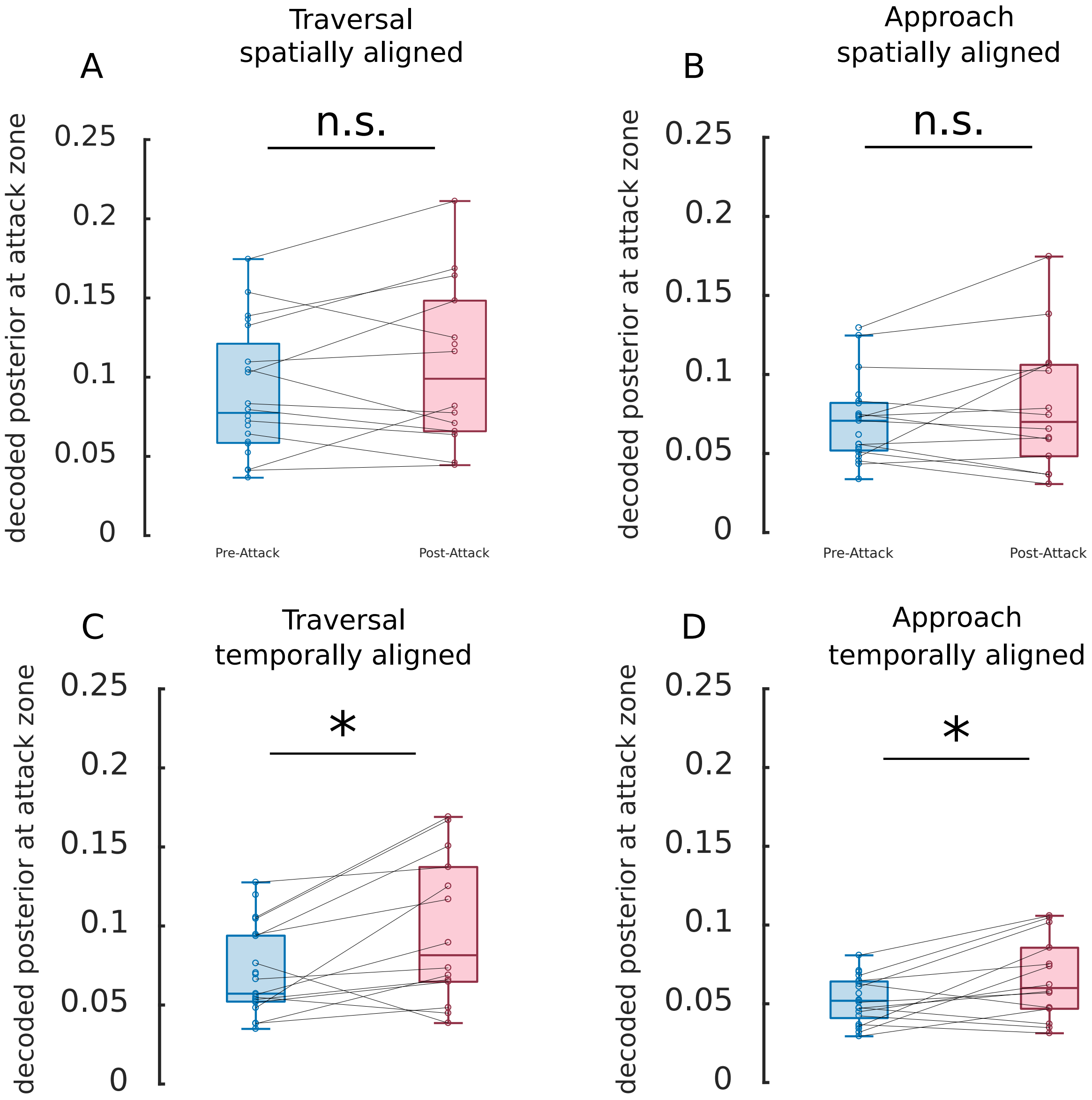

A
Traversal
spatially aligned
n.s.
decoded posterior at attack zone
0.25
0.2
0.15
0.1
0.05
0
Pre-Attack
Post-Attack
B
Approach
spatially aligned
n.s.
decoded posterior at attack zone
0.25
0.2
0.15
0.1
0.05
0
Pre-Attack
Post-Attack
C
Traversal
temporally aligned
*
decoded posterior at attack zone
0.25
0.2
0.15
0.1
0.05
0
Pre-Attack
Post-Attack
D
Approach
temporally aligned
*
decoded posterior at attack zone
0.25
0.2
0.15
0.1
0.05
0
Pre-Attack
Post-Attack

Figure 5.4 Attack zone approach and traversal

Each panel shows the decoded hippocampal representation of the attack zone, comparing outbound laps before and after attack. (**A**) Spatially aligned traversals of the actual attack zone (n.s., p=0.454, Wilcoxon rank sum). (**B**) Spatially aligned approaches to the attack zone (n.s., p=0.903, Wilcoxon rank sum). (**C**) Temporally aligned traversals of the actual attack zone (p=0.013, Wilcoxon rank sum). (**D**) Temporally aligned approaches to the attack zone (p=0.016, Wilcoxon rank sum).

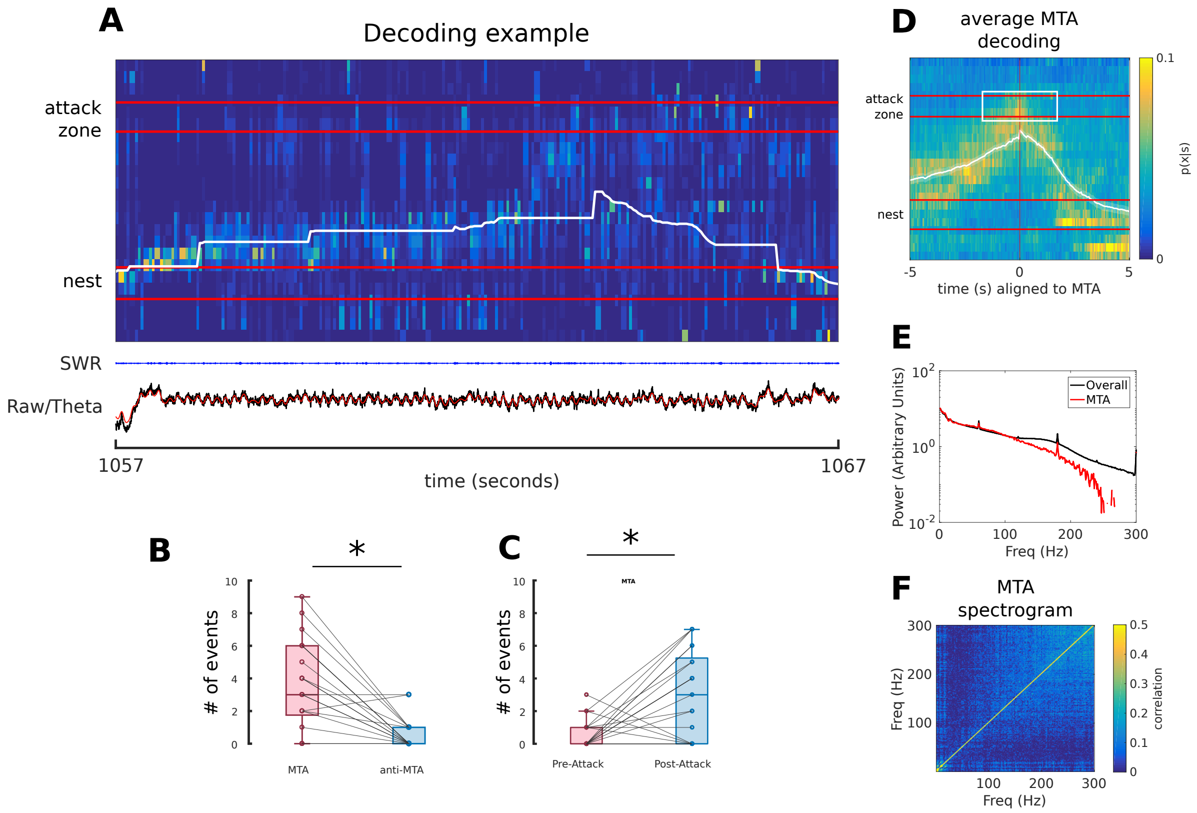

A
Decoding example
attack
zone
nest
SWR
Raw/Theta
1057
1067
time (seconds)
B
*
10
8
6
4
2
0
# of events
MTA
anti-MTA
C
*
MTA
10
8
6
4
2
0
# of events
Pre-Attack
Post-Attack
D
average MTA
decoding
attack
zone
nest
-5
0
5
time (s) aligned to MTA
0.1
0
p(x|s)
E
Power (Arbitrary Units)
10^2
10^0
10^{-2}
Overall
MTA
0
100
200
300
Freq (Hz)
F
MTA
spectrogram
300
200
100
Freq (Hz)
100
200
300
Freq (Hz)
0.5
0.4
0.3
0.2
0.1
0
correlation

Figure 5.5 Attack zone representation during mid-track abort 'change-of-mind' decisions

(**A**) Representative example of a mid-track abort decision. White line = rat position. Decoding (top panel) and local field potential (bottom panel). (**B**) Number of MTA and anti-MTA decisions ($p<0.001$, Wilcoxon rank sum). (**C**) Number of MTA decisions pre- and post-attack ($p=0.002$, Wilcoxon rank sum). (**D**) MTA-peak time-aligned decoding. White line = rat position. Note the decoding to the attack zone at the peak of the MTA (white box). (**E**) Power spectral density during MTA. (**F**) Frequency-by-frequency power correlation during MTA. Note the lack of sharp wave power correlation, indicating that the rats were in theta during MTA events.

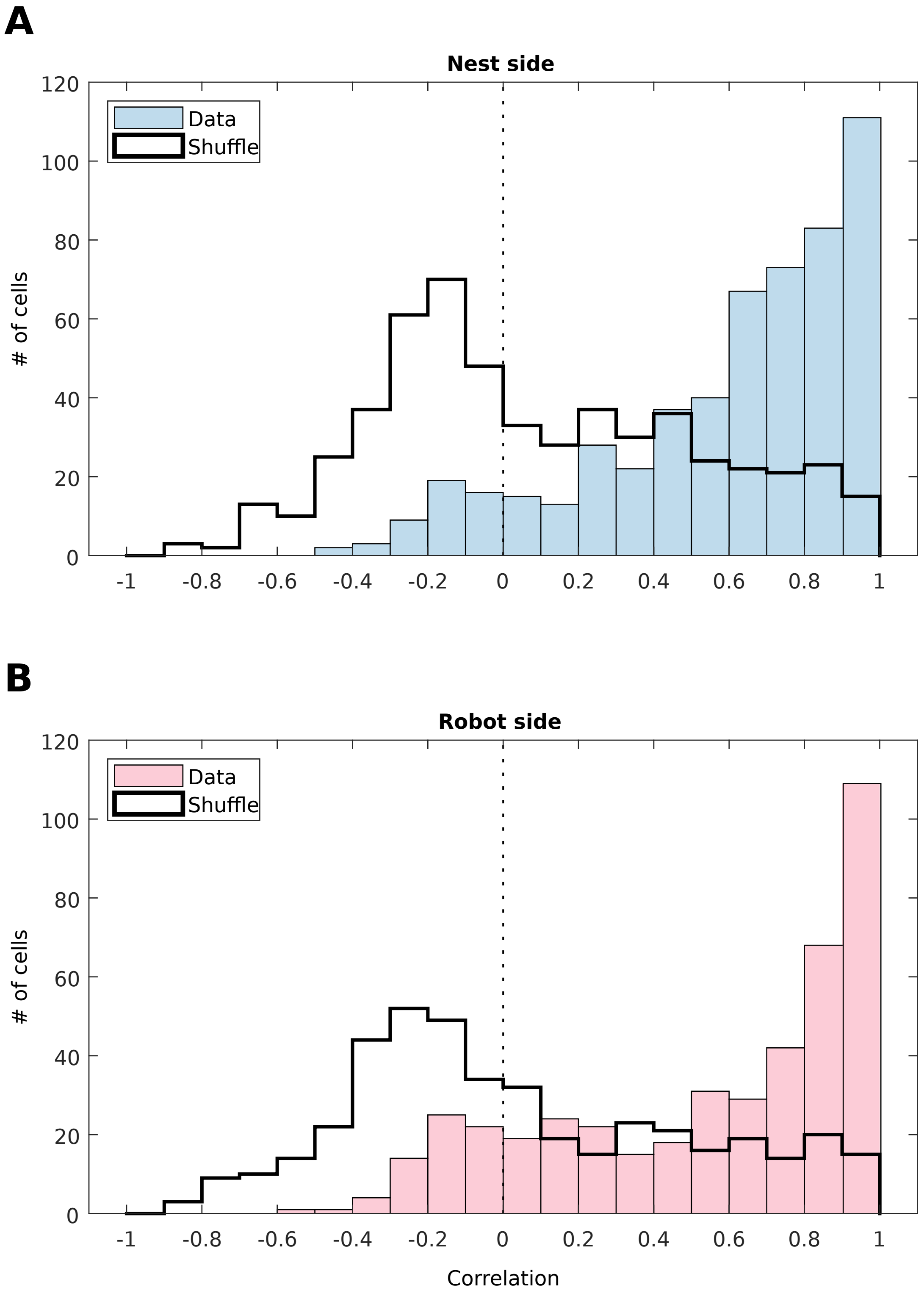

A
Nest side
Data
Shuffle
# of cells
0
20
40
60
80
100
120
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
B
Robot side
Data
Shuffle
# of cells
0
20
40
60
80
100
120
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Correlation

Figure 5.6 Nest side and robot side place field stability

(**A**) Nest side place field stability relative to shuffles (comparison of nest side place field stability relative to shuffles: $p<0.001$, Wilcoxon rank sum). Place fields from place cells corresponding to the nest side of the foraging arena were autocorrelated during the first half (i.e., safe environment) and second half (i.e., unsafe environment) of the session. To obtain shuffles, the activity of a given nest side place cell during the first half of the session was correlated with the activity of a randomly selected place cell during the second half of the session. (**B**) Same as (A), but for robot side place cells (comparison of robot side place field stability relative to shuffles: $p<0.001$, Wilcoxon rank sum). The difference between robot side place field stability and nest side place field stability was not statistically significant ($p=0.064$, Wilcoxon rank sum), but indicated that there was a trend toward less place field stability in robot side place cells.

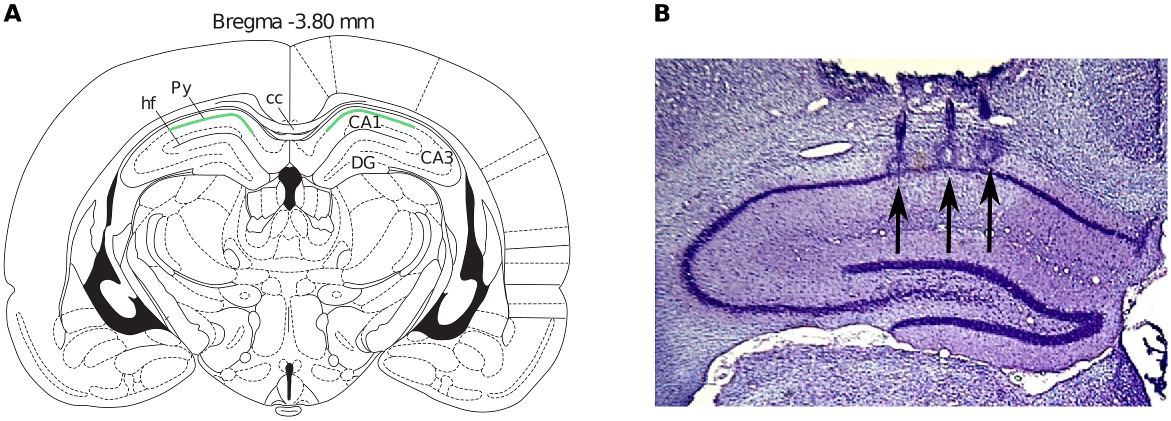

A
Bregma -3.80 mm
Py
hf
cc
CA1
DG
CA3
B

Figure 5.7 Histology and tetrode targeting

(**A**) Coronal section indicating location of tetrodes in dorsal hippocampal layer CA1 (highlighted in green). Schematic from Paxinos and Watson (2006). (**B**) Representative cresyl violet staining of a coronal slice through the hippocampus. Arrows indicating electrolytic lesions.

— Conclusion —

Chapter 6

The neuroscience of negative thinking: from behavior to physiology, physiology to cognition, and cognition to disease

In this thesis we have examined anxiety through multiple lenses spanning historical, behavioral, and pharmacological as well as ethological, theoretical, and neural lines of inquiry.

Historically, we have reviewed attempts by thinkers ranging from Seneca to Sigmund Freud, as well as more recent efforts by psychologists and neuroscientists, to understand anxiety (Chapter 1). Theoretically, I argued that anxiety should not be viewed as a unitary phenomenon, but rather as a range of threat-processing neural algorithms for responding to motivational conflict or potential threats (Chapter 3). Specifically, I argued that there are at least three distinct neural algorithms: hard-wired defensive behaviors, generalization, and prospection. Furthermore, I more broadly examined what happens when these (and other) neural algorithms go awry, and how dysfunction at specific steps in neural information processing produce specific psychiatric diseases (Chapter 2). Importantly, I argue that aberrant negative future thinking is a likely source of many anxiety disorders.

Behaviorally, I presented evidence arguing that, as is seen during reward-based decision-making (Chapter 3), conflict-induced hesitation is a marker of anxiety (Chapters 3, 4, and 5). Ethologically, I demonstrated that rats exhibit anxiety-like hesitation during a semi-naturalistic foraging task, and pharmacologically, I demonstrated that anxiolytic drugs reduce the frequency of these behaviors (Chapter 4).

Altogether, I used modeling and a recently developed anxiogenic avoid-approach conflict task to study the behavioral, pharmacological, and computational foundations of several anxiety-like behaviors (choice-point hesitation and mid-track aborts) in a semi-naturalistic setting, then I tested those theoretical predictions by investigating the neural population dynamics underpinning anxious behavior (Chapter 5). My data support the conclusion that the anxiety-like behaviors I observed during naturalistic avoid-approach conflict co-occur with non-local representations of both rewarding and threatening locations in the environment, suggesting a close relationship between anxiety and future thinking. These results open up several exciting questions for future research in fields ranging from fear and anxiety to decision-making and prospection.

I have argued that hesitation during conflict is a behavioral marker of anxiety during which animals are engaging in prospection. However, recent machine learning technologies allow researchers to interrogate behavior at a fine-grain scale (Arac et al., 2019). An interesting direction for future research will be to quantify behavior at this more nuanced level and look for more novel behavioral markers of anxiety. For instance, recent studies have demonstrated the ability to track the posture of individual body parts during a variety of tasks (Mathis et al., 2018; Pereira et al., 2018). Investigating the neural representations that co-occur with these more subtle bodily expressions of anxiety will provide a more complete picture of anxiety.

Another avenue for future research will be to design targeted, non-habit forming pharmaceuticals for treating anxiety disorders. While both ethanol and diazepam are habit-forming, ethanol is non-specific while diazepam is more restricted in its mechanism of action. I have proposed a computational mechanism of action for diazepam, namely that it might be indirectly disrupting fictive representations generated by the hippocampus due to its ability to suppress hippocampal theta oscillations. A deeper understanding of the computational basis of various types of anxiety must be used to guide the synthesis of increasingly targeted, safer drugs. It is important to emphasize that, due to the multi-faceted nature of anxiety, it is unlikely that there will ever be a single treatment for anxiety; instead, behavioral and pharmacological remedies will have to be personalized and tailored to each patient's specific type of anxiety and therefore their specific category of circuit-dysfunction.

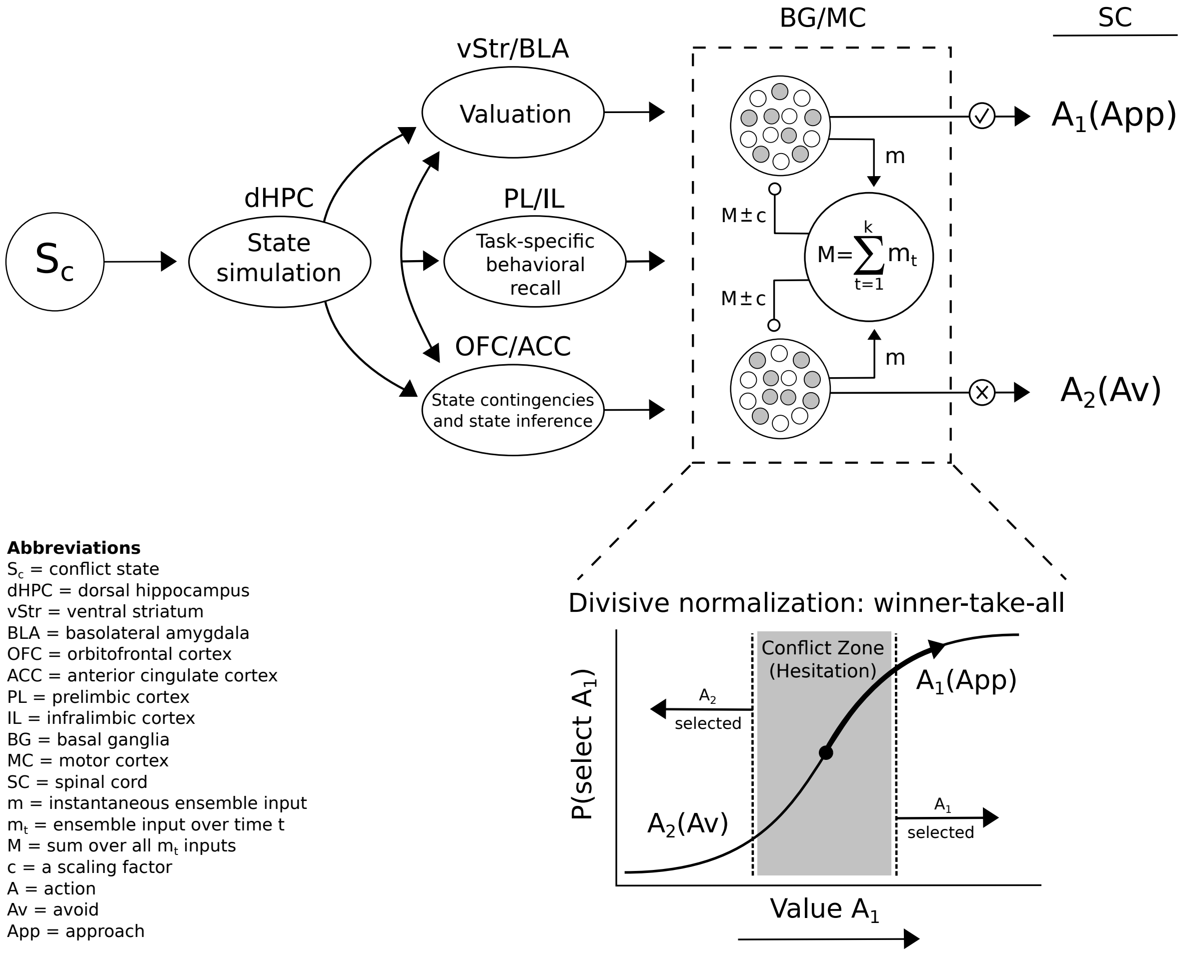
It is well-known that the ventral hippocampus and amygdala are critical to certain forms of anxiety (Maren, 1999; Richmond et al., 1999; LeDoux, 2000; Maren and Holt, 2004; McHugh et al., 2004; LeDoux, 2007; Sierra-Mercado et al., 2011; Jin and Maren, 2015; Beyeler et al., 2016; Xu et al., 2016; Kim and Cho, 2017). In this dissertation, I showed that the dorsal hippocampus contains anxiety-relevant representations during periods of motivational conflict (Chapter 5). This raises the question: what other structures are involved in anxiety (and what computations they might be performing)? The available data suggest that there is a multi-structure anxiety network including the prefrontal cortex, amygdala, hippocampus, bed nucleus of the stria terminalis, hypothalamus, and brainstem (LeDoux, 2000; LeDoux, 2007; Janak and Tye, 2015; Mobbs et al., 2019). It will be the work of future research to examine the dynamics of this network **(1)** during different types of anxiety, **(2)** during healthy and pathological expressions of anxiety, and **(3)** in a cross-species manner.

Though this anxiety network will be challenging to untangle in its various modes and instantiations, I would speculate that **(1)** the dorsal hippocampus is involved in episodic future thinking, **(2)** the ventral hippocampus signals whether an environment and/or context is safe or unsafe, **(3)** the amygdala plays a role in rapid species-specific defensive behaviors as well as infusing dorsal hippocampal representations with value, and **(4)** the prefrontal cortex releases context-appropriate behaviors (see Fig. 6.1). It is important to keep in mind that this network is densely interconnected, and any one of these computations could be a concerted effort between multiple interacting structures. The downstream targets of this anxiety network are structures like the hypothalamus, bed nucleus of the stria terminalis, and brainstem. These 'effector' structures are functionally more well-understood and appear to be involved in hormone regulation and autonomic function.

This framework generates a host of follow-up experiments and predictions. Firstly, it could be experimentally investigated whether there are hippocampal representations of reward and conflict during other anxiogenic paradigms, demonstrating that prospection is a more general computational phenomenon seen across a variety of anxiety-inducing scenarios. Secondly, future research could explore whether there are similar representations in humans (e.g., using multi-voxel pattern analysis) during periods of anxious conflict, thus providing evidence for a conserved, cross-species system. Thirdly, it would be fascinating to record from both the amygdala and the dorsal hippocampus during an anxiogenic task to determine whether there are amygdalar representations of value (either positively-, negatively-, or bi-valenced) that co-occur with the hippocampal non-local representations of space. Additionally, it would be worth knowing whether these amygdalar representations occur near the beginning (and thus potentially initiate the non-local hippocampal activity) or end (and thus are potentially initiated by the non-local hippocampal activity) of these fictive spatial representations. Finally, if there are amygdalar representations of value found during periods of hippocampal non-local representation, are these similarly seen in rodents and humans?

While much remains to be explored, much of this puzzle has already been solved. For instance, in keeping with the above outline of the anxiety network, **(1)** the ventral hippocampus does appear to play a role in signalling safety (Meyer et al., 2019; Çavdaroğlu et al., 2020), **(2)** the prefrontal cortex does appear to releases context-appropriate behaviors (St. Onge and Floresco, 2009; Zeeb et al., 2015; Grunfeld and Likhtik, 2018), and **(3)** the amygdala does appear to represent value and drive species-specific defensive behaviors (LeDoux, 2007; Janak and Tye, 2015).

As we continue to examine the neural foundations of threat-processing, we will not only inch closer to solving the riddle of anxiety, but we will ultimately develop lasting therapeutics for this major class of psychiatric disease. Importantly, through the use of tools that seek to bridge the gap between neurophysiology and behavior, we will undoubtedly shed light on more general computational principles that structure how the brain functions in states of health and disease.



Abbreviations

S_c = conflict state
dHPC = dorsal hippocampus
vStr = ventral striatum
BLA = basolateral amygdala
OFC = orbitofrontal cortex
ACC = anterior cingulate cortex
PL = prelimbic cortex
IL = infralimbic cortex
BG = basal ganglia
MC = motor cortex
SC = spinal cord
m = instantaneous ensemble input
m_t = ensemble input over time t
M = sum over all m_t inputs
c = a scaling factor
A = action
Av = avoid
App = approach

Figure 6.1 A proposed model of a conflict detection-and-resolution network that tracks information flow from the initial recognition of a conflict scenario all the way to the selection of one action among competing alternatives. During periods of conflict, I hypothesize that the dorsal hippocampus simulates states using its map of the task space as a substrate for spatial planning (Addis and Schacter, 2007; Buckner and Carroll, 2007; Gilbert and Wilson, 2007; Suddendorf and Corbaillis, 2007; Schacter and Addis, 2011; Redish, 2016; Kay, 2020). State-outcome contingencies (i.e., the environmental statistics) are then represented by orbitofrontal, prelimbic, and infralimbic cortices (St. Onge and Floresco, 2009; Hillman and Bilkey, 2010; Cowen et al., 2012; Sharpe et al., 2015; Zeeb et al., 2015), while the ventral striatum and basolateral amygdala evaluate (and update the stored value of) those contingencies (Schoenbaum et al., 2003; Richard and Berridge, 2011; Sugam et al., 2014; Sharpe and Schoenbaum, 2016; Zalocusky et al., 2016; Lichtenberg et al., 2017). It has been suggested that the basal ganglia is well situated to integrate incoming contingency/value information and gate access to action representations in motor cortex and the spinal cord that are indexed by basal ganglia ensembles (Mink, 1996; Hazy et al., 2007). Current motor control theories propose that a single action representation stored in motor cortex is then selected from an array of potential actions, possibly by means of a winner-take-all network architecture between basal ganglia and motor cortex ensembles (Lisman, 2014; Baston and Ursino, 2015). Finally, spinal cord networks execute the appropriate muscle synergies corresponding to the selected action (in the case of this figure, selecting and executing an approach behavior).

Bibliography

Abivardi, A., Khemka, S. and Bach, D. R. (2020), ‘Hippocampal representation of threat features and behavior in a human approach–avoidance conflict anxiety task’, Journal of Neuroscience 40(35), 6748–6758.

Adams, C. D. and Dickinson, A. (1981), ‘Instrumental responding following reinforcer devaluation’, The Quarterly Journal of Experimental Psychology Section B 33(2b), 109–121.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. and Friston, K. J. (2013), ‘The computational anatomy of psychosis’, Frontiers in Psychiatry 4, 47.

Addis, D. R. and Schacter, D. L. (2008), ‘Constructive episodic simulation: Temporal distance and detail of past and future events modulate hippocampal engagement’, Hippocampus 18(2), 227–237.

Adhikari, A., Topiwala, M. A. and Gordon, J. A. (2010), ‘Synchronized activity between the ventral hippocampus and the medial prefrontal cortex during anxiety’, Neuron 65(2), 257–269.

Amemori, K.-i. and Graybiel, A. M. (2012), ‘Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making’, Nature Neuroscience 15(5), 776.

Amir, A., Lee, S.-C., Headley, D. B., Herzallah, M. M. and Pare, D. (2015), ‘Amygdala signaling during foraging in a hazardous environment’, Journal of Neuroscience 35(38), 12994–13005.

Anagnostaras, S. G., Maren, S. and Fanselow, M. S. (1999), ‘Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: Within-subjects examination’, Journal of Neuroscience 19(3), 1106–1114.

Arac, A., Zhao, P., Dobkin, B. H., Carmichael, S. T. and Golshani, P. (2019), ‘Deep behavior: A deep learning toolbox for automated analysis of animal and human behavior imaging data’, Frontiers in Systems Neuroscience 13, 20.

Aronov, D., Nevers, R. and Tank, D. W. (2017), ‘Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit’, Nature 543(7647), 719–722.

Bach, D. R. (2015), ‘Anxiety-like behavioural inhibition is normative under environmental threat-reward correlations’, PLoS Computational Biology 11(12), e1004646.

Bach, D. R., Hoffmann, M., Finke, C., Hurlemann, R. and Ploner, C. J. (2019), ‘Disentangling hippocampal and amygdala contribution to human anxiety-like behavior’, Journal of Neuroscience 39(43), 8517–8526.

Balban, M. Y., Cafaro, E., Saue-Fletcher, L., Washington, M. J., Bijanzadeh, M., Lee, A. M., Chang, E. F. and Huberman, A. D. (2020), ‘Human responses to visually evoked threat’, Current Biology.

Bannerman, D., Rawlins, J., McHugh, S., Deacon, R., Yee, B., Bast, T., Zhang, W.-N., Pothuizen, H. and Feldon, J. (2004), 'Regional dissociations within the hippocampus—memory and anxiety', Neuroscience & Biobehavioral Reviews 28(3), 273–283.

Barker, J. M., Taylor, J. R. and Chandler, L. J. (2014), 'A unifying model of the role of the infralimbic cortex in extinction and habits', Learning & Memory 21(9), 441–448.

Barrett, L. F. (2017), 'The theory of constructed emotion: An active inference account of interoception and categorization', Social Cognitive and Affective Neuroscience 12(1), 1–23.

Baston, C., & Ursino, M. (2015), 'A biologically inspired computational model of basal ganglia in action selection'. Computational Intelligence and Neuroscience, 2015.

Bechara, A. and Damasio, A. R. (2005), 'The somatic marker hypothesis: A neural theory of economic decision', Games and Economic Behavior 52(2), 336–372.

Beck, A. T., Emery, G. and Greenberg, R. L. (2005), 'Anxiety disorders and phobias: A cognitive perspective', Basic Books.

Beeman, C. L., Bauer, P. S., Pierson, J. L. and Quinn, J. J. (2013), 'Hippocampus and medial prefrontal cortex contributions to trace and contextual fear memory expression over time’, Learning & Memory 20(6), 336–343.

Benoit, R. G., Davies, D. J. and Anderson, M. C. (2016), ‘Reducing future fears by suppressing the brain mechanisms underlying episodic simulation’, Proceedings of the National Academy of Sciences 113(52), E8492–E8501.

Benoit, R. G., Gilbert, S. J. and Burgess, P. W. (2011), ‘A neural mechanism mediating the impact of episodic prospection on farsighted decisions’, Journal of Neuroscience 31(18), 6771–6779.

Bergstrom, D., Carlson, J., Chase, T., Braun, A. (1987), ‘D1 dopamine receptor activation required for postsynaptic expression of D2 agonist effects’, Science 236(4802), 719–722.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S. and Salzman, C. D. (2020), ‘The geometry of abstraction in the hippocampus and prefrontal cortex’, Cell 183(4), 954–967.

Berns, G. S. and Sejnowski, T. J. (1998), ‘A computational model of how the basal ganglia produce sequences’, Journal of Cognitive Neuroscience 10(1), 108–121.

Beyeler, A., Namburi, P., Glober, G. F., Simonnet, C., Calhoon, G. G., Conyers, G. F., Luck, R., Wildes, C. P. and Tye, K. M. (2016), ‘Divergent routing of positive and negative information from the amygdala during memory retrieval’, Neuron 90(2), 348–361.

Bischof-Köhler, D. (1985), ‘Zur phylogenese menschlicher motivation’.

Blanchard, D. C. and Blanchard, R. J. (2008), ‘4. defensive behaviors, fear, and anxiety’, Handbook of Behavioral Neuroscience 17, 63–79.

Blanchard, D. C., Blanchard, R. J., Tom, P. and Rodgers, R. J. (1990), ‘Diazepam changes risk assessment in an anxiety/defense test battery’, Psychopharmacology 101(4), 511–518.

Blanchard, R. J. and Blanchard, D. C. (1989), ‘Antipredator defensive behaviors in a visible burrow system’, Journal of Comparative Psychology 103(1), 70.

Blanchard, R. J., Blanchard, D. C., Rodgers, J. and Weiss, S. M. (1990), ‘The characterization and modelling of antipredator defensive behavior’, Neuroscience & Biobehavioral Reviews 14(4), 463–472.

Blanchard, R. J., Blanchard, D. C., Weiss, S. M. and Meyer, S. (1990), ‘The effects of ethanol and diazepam on reactions to predatory odors’, Pharmacology Biochemistry and Behavior 35(4), 775–780.

Blanchard, R. J., Magee, L. K., Veniegas, R. and Blanchard, D. C. (1993), ‘Alcohol and anxiety: ethopharmacological approaches.’, Progress in Neuropsychopharmacology & Biological Psychiatry .

Blanchard, R. J., Yang, M., Li, C.-I., Gervacio, A. and Blanchard, D. C. (2001), ‘Cue and context conditioning of defensive behaviors to cat odor stimuli’, Neuroscience & Biobehavioral Reviews 25(7-8), 587–595.

Blanchard, R. J., Yudko, E. B., Rodgers, R. J. and Blanchard, D. C. (1993), ‘Defense system psychopharmacology: an ethological approach to the pharmacology of fear and anxiety’, Behavioural Brain Research 58(1-2), 155–165.

Bolles, R. C. (1970), ‘Species-specific defense reactions and avoidance learning’, Psychological Review 77(1), 32.

Borsboom, D., Cramer, A. O. and Kalis, A. (2019), ‘Brain disorders? Not really: Why network structures block reductionism in psychopathology research’, Behavioral and Brain Sciences 42.

Buckner, R. L. and Carroll, D. C. (2007), ‘Self-projection and the brain’, Trends in Cognitive Sciences 11(2), 49–57.

Burgos-Robles, A., Kimchi, E. Y., Izadmehr, E. M., Porzenheim, M. J., Ramos-Guasp, W. A., Nieh, E. H., Felix-Ortiz, A. C., Namburi, P., Leppla, C. A., Presbrey, K. N. et al. (2017), ‘Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment’, Nature Neuroscience 20(6), 824–835.

Buzsáki, G. (2015), ‘Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning’, Hippocampus 25(10), 1073–1188.

Bzdok, D. and Meyer-Lindenberg, A. (2018), ‘Machine learning for precision psychiatry: Opportunities and challenges’, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 3(3), 223–230.

Calhoon, G. G. and Tye, K. M. (2015), ‘Resolving the neural circuits of anxiety’, Nature Neuroscience 18(10), 1394–1404.

Carr, M. F., Jadhav, S. P. and Frank, L. M. (2011), ‘Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval’, Nature Neuroscience 14(2), 147.

Çavdaroğlu, B., Toy, J., Schumacher, A., Carvalho, G., Patel, M. and Ito, R. (2020), ‘Ventral hippocampus inactivation enhances the extinction of active avoidance responses in the presence of safety signals but leaves discrete trial operant active avoidance performance intact’, Hippocampus 30(9), 913–925.

Chekroud, A. M. (2015), ‘Unifying treatments for depression: An application of the free energy principle’, Frontiers in Psychology 6, 153.

Choi, J.-S. and Kim, J. J. (2010), ‘Amygdala regulates risk of predation in rats foraging in a dynamic fear environment’, Proceedings of the National Academy of Sciences 107(50), 21773–21777.

Ciocchi, S., Passecker, J., Malagon-Vina, H., Mikus, N. and Klausberger, T. (2015), ‘Selective information routing by ventral hippocampal CA1 projection neurons’, Science 348(6234), 560–563.

Clayton, N. S., Bussey, T. J. and Dickinson, A. (2003), ‘Can animals recall the past and plan for the future?’, Nature Reviews Neuroscience 4(8), 685–691.

Conceicao, V. A., Dias, A., Farinha, A. C. and Maia, T. V. (2017), ‘Premonitory urges and tics in tourette syndrome: Computational mechanisms and neural correlates’, Current Opinion in Neurobiology 46, 187–199.

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C. and Ioannidis, J. P. (2008), ‘Life cycle of translational research for medical interventions’, Science 321.5894, 1298-1299.

Corcoran, K. A. and Maren, S. (2001), ‘Hippocampal inactivation disrupts contextual retrieval of fear memory after extinction’, Journal of Neuroscience 21(5), 1720–1726.

Corcoran, K. A. and Quirk, G. J. (2007), ‘Activity in prelimbic cortex is necessary for the expression of learned, but not innate, fears’, Journal of Neuroscience 27(4), 840–844.47.

Cowen, S. L., Davis, G. A. and Nitz, D. A. (2012), ‘Anterior cingulate neurons in the rat map anticipated effort and reward to their associated action sequences’, Journal of Neurophysiology 107(9), 2393–2407.

Crepeau, L. and Newman, J. (1991), ‘Gender differences in reactivity of adult squirrel monkeys to short-term environmental challenges’, Neuroscience & Biobehavioral Reviews 15(4), 469–471.

Crews, F. T., Morrow, A. L., Criswell, H. and Breese, G. (1996), ‘Effects of ethanol on ion channels’, International Review of Neurobiology 39, 283–367.

Daw, N. D., Kakade, S. and Dayan, P. (2002), ‘Opponent interactions between serotonin and dopamine’, Neural Networks 15(4-6), 603–616.

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.

Dayan, P. and Abbott, L. (2001), ‘Theoretical neuroscience: Computational and mathematical modeling of neural systems’, Computational Neuroscience Series.

Dayan, P. and Huys, Q. J. (2008), ‘Serotonin, inhibition, and negative mood’, PLoS Computational Biology 4(2), e4.

Dayan, P. and Huys, Q. J. (2009), ‘Serotonin in affective control’, Annual Review of Neuroscience 32, 95–126.

Decker, M. W., Curzon, P. and Brioni, J. D. (1995), ‘Influence of separate and combined septal and amygdala lesions on memory, acoustic startle, anxiety, and locomotor activity in rats’, Neurobiology of Learning and Memory 64(2), 156–168.

Declercq, M., De Houwer, J. and Baeyens, F. (2008), ‘Evidence for an expectancy-based theory of avoidance behaviour’, Quarterly Journal of Experimental Psychology 61(12), 1803–1812.

Degroot, A. and Treit, D. (2003), ‘Septal GABAergic and hippocampal cholinergic systems interact in the modulation of anxiety’, Neuroscience 117(2), 493–501.

Degroot, A. and Treit, D. (2004), ‘Anxiety is functionally segregated within the septo-hippocampal system’, Brain Research 1001(1-2), 60–71.

Deitrich, R. A., Dunwiddie, T. V., Harris, R. A. and Erwin, V. G. (1989), ‘Mechanism of action of ethanol: Initial central nervous system actions’, Pharmacological Reviews 41(4), 489–537.

Dias, B. G., Banerjee, S. B., Goodman, J. V. and Ressler, K. J. (2013), ‘Towards new approaches to disorders of fear and anxiety’, Current Opinion in Neurobiology 23(3), 346–352.

Dougherty, D. D., Brennan, B. P., Stewart, S. E., Wilhelm, S., Widge, A. S. and Rauch, S. L. (2018), ‘Neuroscientifically informed formulation and treatment planning for patients with obsessive-compulsive disorder: A review’, JAMA Psychiatry 75(10), 1081–1087.

Echeveste, R., Aitchison, L., Hennequin, G. and Lengyel, M. (2020), ‘Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference’, Nature Neuroscience 23(9), 1138–1149.

Ehlers, A., Margraf, J., Roth, W. T., Taylor, C. B. and Birbaumer, N. (1988), ‘Anxiety induced by false heart rate feedback in patients with panic disorder’, Behaviour Research and Therapy 26(1), 1–11.

Einarsson, E. Ö., Pors, J. and Nader, K. (2015), ‘Systems reconsolidation reveals a selective role for the anterior cingulate cortex in generalized contextual fear memory expression’, Neuropsychopharmacology 40(2), 480–487.

Everitt, B., Morris, K., O’Brien, A. and Robbins, T. (1991), ‘The basolateral amygdala-ventral striatal system and conditioned place preference: Further evidence of limbic-striatal interactions underlying reward-related processes’, Neuroscience 42(1), 1–18.

Fanselow, M., Lester, L., Bolles, R. and Beecher, M. (1988), ‘Evolution and learning’.

Fanselow, M. S. (1994), ‘Neural organization of the defensive behavior system responsible for fear’, Psychonomic Bulletin & Review 1(4), 429–438.

Fanselow, M. S. and Dong, H.-W. (2010), ‘Are the dorsal and ventral hippocampus functionally distinct structures?’, Neuron 65(1), 7–19.

Fanselow, M. S. and Lester, L. S. (1988), ‘A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior’, in R. C. Bolles & M. D. Beecher (Eds.), Evolution and learning (p. 185–212), Lawrence Erlbaum Associates, Inc.

Ferreira, V., Takahashi, R. and Morato, G. (2000), ‘Dexamethasone reverses the ethanol-induced anxiolytic effect in rats’, Pharmacology Biochemistry and Behavior 66(3), 585–590.

Fiser, J., Berkes, P., Orbán, G. and Lengyel, M. (2010), ‘Statistically optimal perception and learning: From behavior to neural representations’, Trends in Cognitive Sciences 14(3), 119–130.

Flagel, S., Pine, D., Ahmari, S., First, M., Friston, K., Mathys, C. D., Redish, A., Schmack, K., Smoller, J., Thapar, A. et al. (2016), ‘A novel framework for improving psychiatric diagnostic nosology’, Computational Psychiatry: New perspectives on mental illness.

Frank, M. J., Santamaria, A., O’Reilly, R. C. and Willcutt, E. (2007), ‘Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder’, Neuropsychopharmacology 32(7), 1583.

Frankland, P. W., Bontempi, B., Talton, L. E., Kaczmarek, L. and Silva, A. J. (2004), ‘The involvement of the anterior cingulate cortex in remote contextual fear memory’, Science 304(5672), 881–883.

Frankland, P. W., Cestari, V., Filipkowski, R. K., McDonald, R. J. and Silva, A. J. (1998), ‘The dorsal hippocampus is essential for context discrimination but not for contextual conditioning’, Behavioral Neuroscience 112(4), 863.

Freeman, J., Garcia, A., Benito, K., Conelea, C., Sapyta, J., Khanna, M., March, J. and Franklin, M. (2012), ‘The pediatric obsessive compulsive disorder treatment study for young children (pots jr): Developmental considerations in the rationale, design, and methods', Journal of Obsessive-compulsive and Related Disorders 1(4), 294–300.

Freud, S. (1917), 'Introductory lectures on psychoanalysis'.

Friedman, A., Homma, D., Gibb, L. G., Amemori, K.-i., Rubin, S. J., Hood, A. S., Riad, M. H. and Graybiel, A. M. (2015), 'A corticostriatal path targeting striosomes controls decision-making under conflict', Cell 161(6), 1320–1333.

Friston, K. (2010), 'The free-energy principle: A unified brain theory?', Nature Reviews Neuroscience 11(2), 127.

Friston, K. J., Stephan, K. E., Montague, R. and Dolan, R. J. (2014), 'Computational psychiatry: The brain as a phantastic organ', The Lancet Psychiatry 1(2), 148–158.

Frith, U. (2003), 'Autism: Explaining the enigma', Blackwell Publishing.

Fung, B. J., Qi, S., Hassabis, D., Daw, N. and Mobbs, D. (2019), 'Slow escape decisions are swayed by trait anxiety', Nature Human Behaviour 3(7), 702–708.

Garfinkel, S. N., Tiley, C., O'Keeffe, S., Harrison, N. A., Seth, A. K. and Critchley, H. D. (2016), 'Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety', Biological Psychology 114, 117–126.

Gazzaniga, M. S., Bogen, J. E. and Sperry, R. W. (1965), ‘Observations on visual perception after disconnexion of the cerebral hemispheres in man’, Brain 88(2), 221–236.

George, M. S., Trimble, M. R., Ring, H. A., Sallee, F. and Robertson, M. M. (1993), ‘Obsessions in obsessive-compulsive disorder with and without gilles de la tourette’s syndrome’, The American Journal of Psychiatry.

Gilbert, D. L., Budman, C. L., Singer, H. S., Kurlan, R. and Chipkin, R. E. (2014), ‘A D1 receptor antagonist, ecopipam, for treatment of tics in tourette syndrome’, Clinical Neuropharmacology 37(1), 26–30.

Gilbert, D. T. and Wilson, T. D. (2007), ‘Prospection: Experiencing the future’, Science 317(5843), 1351–1354.

Gilbert, D. T. and Wilson, T. D. (2009), ‘Why the brain talks to itself: Sources of error in emotional prediction’, Philosophical Transactions of the Royal Society B: Biological Sciences 364(1521), 1335–1341.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W. and de Wit, S. (2011), ‘Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder’, American Journal of Psychiatry 168(7), 718–726.

Gillan, C. M. and Robbins, T. W. (2014), ‘Goal-directed learning and obsessive–compulsive disorder’, Philosophical Transactions of the Royal Society B: Biological Sciences 369(1655), 20130475.

Giustino, T. F. and Maren, S. (2015), ‘The role of the medial prefrontal cortex in the conditioning and extinction of fear’, Frontiers in Behavioral Neuroscience 9, 298.

Gómez, C., Lizier, J. T., Schaum, M., Wollstadt, P., Grützner, C., Uhlhaas, P., Freitag, C. M., Schlitt, S., Bölte, S., Hornero, R. et al. (2014), ‘Reduced predictable information in brain signals in autism spectrum disorder’, Frontiers in Neuroinformatics 8, 9.

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J. and Steele, J. D. (2011), ‘Expected value and prediction error abnormalities in depression and schizophrenia’, Brain 134(6), 1751–1764.

Gray, J. A. (1982), ‘Précis of the neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system’, Behavioral and Brain Sciences 5(3), 469–484.

Gray, J. A. and McNaughton, N. (2003), ‘The neuropsychology of anxiety: An enquiry into the function of the septo-hippocampal system’, New York: Oxford University.

Graybiel, A. M. (1995), ‘Building action repertoires: Memory and learning functions of the basal ganglia’, Current Opinion in Neurobiology 5(6), 733–741.

Graybiel, A. M. and Rauch, S. L. (2000), ‘Toward a neurobiology of obsessive-compulsive disorder’, Neuron 28(2), 343–347.

Grewal, S. S., Shepherd, J. K., Bill, D. J., Fletcher, A. and Dourish, C. T. (1997), ‘Behavioural and pharmacological characterisation of the canopy stretched attend posture test as a model of anxiety in mice and rats’, Psychopharmacology 133(1), 29–38.

Grunfeld, I. S. and Likhtik, E. (2018), ‘Mixed selectivity encoding and action selection in the prefrontal cortex during threat assessment’, Current Opinion in Neurobiology 49, 108–115.

Guise, K. G. and Shapiro, M. L. (2017), ‘Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding’, Neuron 94(1), 183–192.

Happé, F. and Frith, U. (2006), ‘The weak coherence account: Detail-focused cognitive style in autism spectrum disorders’, Journal of Autism and Developmental Disorders 36(1), 5–25.

Hassabis, D., Kumaran, D., Vann, S. D. and Maguire, E. A. (2007), ‘Patients with hippocampal amnesia cannot imagine new experiences’, Proceedings of the National Academy of Sciences 104(5), 1726–1731.

Hassabis, D. and Maguire, E. A. (2009), ‘The construction system of the brain’, Philosophical Transactions of the Royal Society B: Biological Sciences 364(1521), 1263–1271.

Hauser, T. U., Fiore, V. G., Moutoussis, M. and Dolan, R. J. (2016), ‘Computational psychiatry of ADHD: Neural gain impairments across marrian levels of analysis’, Trends in Neurosciences 39(2), 63–73.

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S.and Brem, S. (2014), ‘Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit / hyperactivity disorder’, JAMA Psychiatry 71(10), 1165–1173.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007), ‘Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system’, Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1485), 1601-1613.

Hebb, D. (1957), 0.(1949), ‘The organization of behavior’.

Heilbronner, S. R., Rodriguez-Romaguera, J., Quirk, G. J., Groenewegen, H. J. and Haber, S. N. (2016), ‘Circuit-based corticostriatal homologies between rat and primate’, Biological Psychiatry 80(7), 509–521.

Heller, A. S. and Bagot, R. C. (2020), ‘Is hippocampal replay a mechanism for anxiety and depression?’, JAMA Psychiatry 77(4), 431–432.

Hertz, J., Krogh, A., Palmer, R. G. and Horner, H. (1991), ‘Introduction to the theory of neural computation’, Physics Today 44, 70.

Hillman, K. L. and Bilkey, D. K. (2010), 'Neurons in the rat anterior cingulate cortex dynamically encode cost–benefit in a spatial decision-making task', Journal of Neuroscience 30(22), 7705–7713.

Hobin, J. A., Ji, J. and Maren, S. (2006), 'Ventral hippocampal muscimol disrupts context-specific fear memory retrieval after extinction in rats', Hippocampus 16(2), 174–182.

Holly, K. S., Orndorff, C. O. and Murray, T. A. (2016), 'Matsap: An automated analysis of stretch-attend posture in rodent behavioral experiments', Scientific Reports 6(1), 1–9.

Hopfield, J. J. (1982), 'Neural networks and physical systems with emergent collective computational abilities', Proceedings of the National Academy of Sciences 79(8), 2554–2558.

Huang, Y. and Rao, R. P. (2011), 'Predictive coding', Wiley Interdisciplinary Reviews: Cognitive Science 2(5), 580–593.

Huys, Q. J., Daw, N. D. and Dayan, P. (2015), 'Depression: A decision-theoretic analysis', Annual Review of Neuroscience 38, 1–23.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P. and Roiser, J. P. (2012), 'Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees', PLoS Computational Biology 8(3), e1002410.

Huys, Q. J., Maia, T. V. and Frank, M. J. (2016), ‘Computational psychiatry as a bridge from neuroscience to clinical applications’, Nature Neuroscience 19(3), 404.

Huys, Q. J., Pizzagalli, D. A., Bogdan, R. and Dayan, P. (2013), ‘Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis’, Biology of Mood & Anxiety Disorders 3(1), 12.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I. and Moser, M.-B. (2015), ‘A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation’, Nature 522(7554), 50–55.

Ito, R. and Lee, A. C. (2016), ‘The role of the hippocampus in approach-avoidance conflict decision-making: Evidence from rodent and human studies’, Behavioural Brain Research 313, 345–357.54.

Iwata, J., LeDoux, J. E., Meeley, M. P., Arneric, S. and Reis, D. J. (1986), ‘Intrinsic neurons in the amygdaloid field projected to by the medial geniculate body mediate emotional responses conditioned to acoustic stimuli’, Brain Research 383(1-2), 195–214.

Jacinto, L. R., Cerqueira, J. J. and Sousa, N. (2016), ‘Patterns of theta activity in limbic anxiety circuit preceding exploratory behavior in approach-avoidance conflict’, Frontiers in Behavioral Neuroscience 10, 171.

Jadhav, S. P. and Frank, L. M. (2009), ‘Reactivating memories for consolidation’, Neuron 62(6), 745–746.

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

Jay, T. M., Glowinski, J. and Thierry, A.-M. (1989), ‘Selectivity of the hippocampal projection to the prelimbic area of the prefrontal cortex in the rat’, Brain Research 505(2), 337–340.

Jiménez-Velázquez, G., López-Muñoz, F. J. and Fernández-Guasti, A. (2010), ‘Parallel anxiolytic-like and antinociceptive actions of diazepam in the anterior basolateral amygdala and dorsal periaqueductal gray’, Brain Research 1349, 11–20.

Jin, J. and Maren, S. (2015), ‘Fear renewal preferentially activates ventral hippocampal neurons projecting to both amygdala and prefrontal cortex in rats’, Scientific Reports 5(1), 1–6.

Johansen, J. P., Fields, H. L. and Manning, B. H. (2001), ‘The affective component of pain in rodents: Direct evidence for a contribution of the anterior cingulate cortex’, Proceedings of the National Academy of Sciences 98(14), 8077–8082.

Johnson, A., Fenton, A. A., Kentros, C. and Redish, A. D. (2009), ‘Looking for cognition in the structure within the noise’, Trends in Cognitive Sciences 13(2), 55–64.

Johnson, A. and Redish, A. D. (2007), ‘Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point’, Journal of Neuroscience 27(45), 12176–12189.

Johnston, A. L. and File, S. E. (1991), ‘Sex differences in animal tests of anxiety’, Physiology & Behavior 49(2), 245–250.

Jones, M. W. and Wilson, M. A. (2005), ‘Theta rhythms coordinate hippocampal–prefrontal interactions in a spatial memory task’, PLoS Biology 3(12), e402.

Juavinett, A. L., Erlich, J. C. and Churchland, A. K. (2018), ‘Decision-making behaviors: weighing ethology, complexity, and sensorimotor compatibility’, Current Opinion in Neurobiology 49, 42–50.

Kahneman, D. (2011), ‘Thinking, fast and slow’, Macmillan.

Kalanithi, P. S., Zheng, W., Kataoka, Y., DiFiglia, M., Grantz, H., Saper, C. B., Schwartz, M. L., Leckman, J. F. and Vaccarino, F. M. (2005), ‘Altered parvalbumin-positive neuron distribution in basal ganglia of individuals with tourette syndrome’, Proceedings of the National Academy of Sciences 102(37), 13307–13312.

Kang-Park, M.-H., Wilson, W. and Moore, S. (2004), ‘Differential actions of diazepam and zolpidem in basolateral and central amygdala nuclei’, Neuropharmacology 46(1), 1–9.

Karlsson, M. P. and Frank, L. M. (2009), ‘Awake replay of remote experiences in the hippocampus’, Nature Neuroscience 12(7), 913–918.

Kay, K., Chung, J. E., Sosa, M., Schor, J. S., Karlsson, M. P., Larkin, M. C., Liu, D. F. and Frank, L. M. (2020), ‘Constant sub-second cycling between representations of possible futures in the hippocampus’, Cell 180(3), 552–567.

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J. and Rush-worth, M. F. (2006), ‘Optimal decision making and the anterior cingulate cortex’, Nature Neuroscience 9(7), 940–947.

Kierkegaard, S. (1844), ‘The concept of anxiety’.

Killcross, S. and Coutureau, E. (2003), ‘Coordination of actions and habits in the medial prefrontal cortex of rats’, Cerebral Cortex 13(4), 400–408.

Kim, E. J., Kong, M.-S., Park, S. G., Mizumori, S. J., Cho, J. and Kim, J. J. (2018), ‘Dynamic coding of predatory information between the prelimbic cortex and lateral amygdala in foraging rats’, Science Advances 4(4), eaar7328.

Kim, E. J., Park, M., Kong, M.-S., Park, S. G., Cho, J. and Kim, J. J. (2015), ‘Alterations of hippocampal place cells in foraging rats facing a “predatory” threat’, Current Biology 25(10), 1362–1367.

Kim, J. J. and Fanselow, M. S. (1992), ‘Modality-specific retrograde amnesia of fear’, Science 256(5057), 675–677.

Kim, W. B. and Cho, J.-H. (2017), ‘Synaptic targeting of double-projecting ventral ca1 hippocampal neurons to the medial prefrontal cortex and basal amygdala’, Journal of Neuroscience 37(19), 4868–4882.

Kirlic, N., Young, J. and Aupperle, R. L. (2017), ‘Animal to human translational paradigms relevant for approach avoidance conflict decision making’, Behaviour Research and Therapy 96, 14–29.

Kjelstrup, K. G., Tuvnes, F. A., Steffenach, H.-A., Murison, R., Moser, E. I. and Moser, M.-B. (2002), ‘Reduced fear expression after lesions of the ventral hippocampus’, Proceedings of the National Academy of Sciences 99(16), 10825–10830.

Knill, D. C. and Pouget, A. (2004), ‘The bayesian brain: the role of uncertainty in neural coding and computation’, TRENDS in Neurosciences 27(12), 712–719.

Komorowski, R. W., Garcia, C. G., Wilson, A., Hattori, S., Howard, M. W. and Eichenbaum, H. (2013), ‘Ventral hippocampal neurons are shaped by experience to represent behaviorally relevant contexts’, Journal of Neuroscience 33(18), 8079–8087.

Koob, G. F. and Volkow, N. D. (2010), ‘Neurocircuitry of addiction’, Neuropsychopharmacology 35(1), 217.

Korn, C. W. and Bach, D. R. (2019), ‘Minimizing threat via heuristic and optimal policies recruits hippocampus and medial prefrontal cortex’, Nature Human Behaviour 3(7), 733–745.

Laplane, D., Levasseur, M., Pillon, B., Dubois, B., Baulac, M., Mazoyer,B., Dinh, S. T., Sette, G., Danze, F. and Baron, J. (1989), ‘Obsessive-compulsive and other behavioural changes with bilateral basal ganglia lesions: A neuropsychological, magnetic resonance imaging and positron tomography study’, Brain 112(3), 699–725.

Lashley, K. S. (1951), ‘The problem of serial order in behavior’, Vol. 21, Bobbs-Merrill.

Lebreton, M., Bertoux, M., Boutet, C., Lehericy, S., Dubois, B., Fossati,P. and Pessiglione, M. (2013), ‘A critical role for the hippocampus in the valuation of imagined outcomes’, PLoS Biology 11(10), e1001684.

LeDoux, J. (2007), ‘The amygdala’, Current Biology 17(20), R868–R874.

LeDoux, J. (2012), ‘Rethinking the emotional brain’, Neuron 73(4), 653–676.

LeDoux, J. and Daw, N. D. (2018), ‘Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour’, Nature Reviews Neuroscience.

LeDoux, J. E. (2000), ‘Emotion circuits in the brain’, Annual Review of Neuroscience 23(1), 155–184.

LeDoux, J. E. (2015), ‘Anxious: Using the brain to understand and treat fear and anxiety’, Penguin.

Lee, D. K., Itti, L., Koch, C. and Braun, J. (1999), ‘Attention activates winner-take-all competition among visual filters’, Nature Neuroscience 2(4), 375.

Lewin, K. (1931), ‘Environmental forces in child behavior and development’, in C. Murchison (Ed.), The International University series in psychology. A handbook of child psychology (p. 94–127), Clark University Press.

Li, C., McCall, N. M., Lopez, A. J. and Kash, T. L. (2013), ‘Alcohol effects on synaptic transmission in periaqueductal gray dopamine neurons’, Alcohol 47(4), 279–287.

Lichtenberg, N. T., Pennington, Z. T., Holley, S. M., Greenfield, V. Y.,Cepeda, C., Levine, M. S. and Wassum, K. M. (2017), ‘Basolateral amygdala to orbitofrontal cortex projections enable cue-triggered reward expectations’, Journal of Neuroscience 37(35), 8374–8384.

Lieberman, J. A. (2015), ‘Shrinks: The untold story of psychiatry’, Hachette UK.

Lisman, J. (2014), ‘Two-phase model of the basal ganglia: Implications for discontinuous control of the motor system’, Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1655), 20130489.

Lobo, I. A. and Harris, R. A. (2008), ‘$GABA_A$ receptors and alcohol’, Pharmacology Biochemistry and Behavior 90(1), 90–94.

Loh, M., Rolls, E. T. and Deco, G. (2007), ‘A dynamical systems hypothesis of schizophrenia’, PLoS Computational Biology 3(11), e228.

Lopresto, D., Schipper, P. and Homberg, J. R. (2016), ‘Neural circuits and mechanisms involved in fear generalization: Implications for the pathophysiology and treatment of posttraumatic stress disorder’, Neuroscience & Biobehavioral Reviews 60, 31–42.

Lovibond, P. F., Saunders, J. C., Weidemann, G. and Mitchell, C. J. (2008), ‘Evidence for expectancy as a mediator of avoidance and anxiety in a laboratory model of human avoidance learning’, The Quarterly Journal of Experimental Psychology 61(8), 1199–1216.

Lynn, C. W. and Bassett, D. S. (2019), ‘The physics of brain network structure, function and control’, Nature Reviews Physics p. 1.

MacDonald, A. W., Zick, J. L., Chafee, M. V. and Netoff, T. I. (2016), ‘Integrating insults: Using fault tree analysis to guide schizophrenia research across levels of analysis’, Frontiers in Human Neuroscience 9, 698.

MacLeod, A. K. and Byrne, A. (1996), ‘Anxiety, depression, and the anticipation of future positive and negative experiences’, Journal of Abnormal Psychology 105(2), 286.

Maeng, L. Y. and Milad, M. R. (2015), ‘Sex differences in anxiety disorders:interactions between fear, stress, and gonadal hormones’, Hormones and Behavior 76, 106–117.

Maia, T. V. and Conceicao, V. A. (2017), ‘The roles of phasic and tonic dopamine in tic learning and expression’, Biological Psychiatry 82(6), 401–412.

Maia, T. V. and Frank, M. J. (2011), ‘From reinforcement learning models to psychiatric and neurological disorders’, Nature Neuroscience 14(2), 154.

Maia, T. V. and McClelland, J. L. (2012), ‘A neurocomputational approach to obsessive-compulsive disorder’, Trends in Cognitive Sciences 16(1), 14–15.

Maren, S. (1999), ‘Neurotoxic or electrolytic lesions of the ventral subiculum produce deficits in the acquisition and expression of pavlovian fear conditioning in rats’, Behavioral Neuroscience 113(2), 283.

Maren, S., Aharonov, G. and Fanselow, M. S. (1997), ‘Neurotoxic lesions of the dorsal hippocampus and pavlovian fear conditioning in rats’, Behavioural Brain Research 88(2), 261–274.

Maren, S. and Holt, W. G. (2004), ‘Hippocampus and pavlovian fear conditioning in rats: muscimol infusions into the ventral, but not dorsal, hippocampus impair the acquisition of conditional freezing to an auditory conditional stimulus’, Behavioral Neuroscience 118(1), 97.

Martin, V. C., Schacter, D. L., Corballis, M. C. and Addis, D. R. (2011), ‘A role for the hippocampus in encoding simulations of future events’, Proceedings of the National Academy of Sciences 108(33), 13858–13863.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W. and Bethge, M. (2018), 'DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning', Nature Neuroscience 21(9), 1281–1289.

Mathys, C. et al. (2016), 'How could we get nosology from computation?', in Computational psychiatry: new perspectives on mental illness (Redish AD, Gordon JA, eds). Strüngmann Forum Reports, Vol. 20.

Mattar, M. G. and Daw, N. D. (2018), 'Prioritized memory access explains planning and hippocampal replay', Nature Neuroscience 21(11), 1609–1617.

McHugh, S., Deacon, R., Rawlins, J. and Bannerman, D. M. (2004), 'Amygdala and ventral hippocampus contribute differentially to mechanisms of fear and anxiety', Behavioral Neuroscience 118(1), 63.

McNaughton, N., Kocsis, B. and Hajos, M. (2007), 'Elicited hippocampal theta rhythm: A screen for anxiolytic and procognitive drugs through changes in hippocampal function?', Behavioural Pharmacology 18(5-6), 329–346.

Meyer, H. C., Odriozola, P., Cohodes, E. M., Mandell, J. D., Li, A., Yang, R., Hall, B. S., Haberman, J. T., Zacharek, S. J., Liston, C. et al. (2019), 'Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans', Proceedings of the National Academy of Sciences 116(52), 26970–26979.

Miller, N., Brown, J., Lipofsky, H. and Miller, N. (1943), ‘A theoretical and experimental analysis of conflict behavior: III. Approach-avoidance conflict as a function of strength of drive and strength of shock’, Personality and the Behavior Disorders.

Miller, N. E. (1944), ‘Experimental studies of conflict’.

Miloyan, B., Bulley, A. and Suddendorf, T. (2016), ‘Episodic foresight and anxiety: Proximate and ultimate perspectives’, British Journal of Clinical Psychology 55(1), 4–22.

Mink, J. W. (1996). ‘The basal ganglia: Focused selection and inhibition of competing motor programs’, Progress in Neurobiology, 50(4), 381-425.

Mobbs, D., Adolphs, R., Fanselow, M. S., Barrett, L. F., LeDoux, J. E., Ressler, K. and Tye, K. M. (2019), ‘Viewpoints: Approaches to defining and investigating fear’, Nature Neuroscience 22(8), 1205–1216.

Mobbs, D., Hagan, C. C., Dalgleish, T., Silston, B. and Prévost, C. (2015), ‘The ecology of human fear: Survival optimization and the nervous system’, Frontiers in Neuroscience 9, 55.

Mobbs, D. and Kim, J. J. (2015), ‘Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans’, Current Opinion in Behavioral Sciences 5, 8–15.

Mobbs, D., Marchant, J. L., Hassabis, D., Seymour, B., Tan, G., Gray, M., Petrovic, P., Dolan, R. J. and Frith, C. D. (2009), ‘From threat to fear: The neural organization of defensive fear systems in humans’, Journal of Neuroscience 29(39), 12236–12243.

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.

Mobbs, D., Yu, R., Rowe, J. B., Eich, H., FeldmanHall, O. and Dalgleish, T. (2010), ‘Neural activity associated with monitoring the oscillating threat value of a tarantula’, Proceedings of the National Academy of Sciences 107(47), 20582–20586.

Molewijk, H., Van der Poel, A. and Olivier, B. (1995), ‘The ambivalent behaviour “stretched approach posture” in the rat as a paradigm to characterize anxiolytic drugs’, Psychopharmacology 121(1), 81–90.

Monosov, I. E. (2017), ‘Anterior cingulate is a source of valence-specific information about value and uncertainty’, Nature Communications 8(1), 1–12.

Montague, P. R., Dolan, R. J., Friston, K. J. and Dayan, P. (2012), ‘Computational psychiatry’, Trends in Cognitive Sciences 16(1), 72–80.

Moutoussis, M., Shahar, N., Hauser, T. U. and Dolan, R. J. (2018), ‘Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies’, Computational Psychiatry 2, 50–73.

Muenzinger, K. F. and Gentry, E. (1931), ‘Tone discrimination in white rats’, Journal of Comparative Psychology 12(2), 195.

Narayanan, R. T., Seidenbecher, T., Kluge, C., Bergado, J., Stork, O. and Pape, H.-C. (2007), ‘Dissociated theta phase synchronization in amygdalo-hippocampal circuits during various stages of fear memory’, European Journal of Neuroscience 25(6), 1823–1831.

Narayanan, R. T., Seidenbecher, T., Sangha, S., Stork, O. and Pape, H.-C. (2007), ‘Theta resynchronization during reconsolidation of remote contextual fear memory’, Neuroreport 18(11), 1107–1111.

Niedenthal, P. M. (2007), ‘Embodying emotion’, Science 316(5827), 1002–1005.

NIMH (2019a), ‘National Institute of Mental Health: Anxiety disorders’, https://www.nimh.nih.gov/health/topics/anxiety-disorders/index.shtml.

NIMH (2019b), ‘National Institute of Mental Health: The research domain criteria’, https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/index.shtml.

Niv, Y. (2019), ‘Learning task-state representations’, Nature Neuroscience 22(10), 1544–1553.

Nolen-Hoeksema, S. (2000), ‘The role of rumination in depressive disorders and mixed anxiety/depressive symptoms’, Journal of Abnormal Psychology 109(3), 504.

O’Keefe, J. and Nadel, L. (1978), ‘The hippocampus as a cognitive map’, Oxford: Clarendon Press.

Orsini, C. A., Kim, J. H., Knapska, E. and Maren, S. (2011), ‘Hippocampal and prefrontal projections to the basal amygdala mediate contextual regulation of fear after extinction’, Journal of Neuroscience 31(47), 17269–17277.

Paré, D. and Quirk, G. J. (2017), ‘When scientific paradigms lead to tunnel vision: Lessons from the study of fear’, npj Science of Learning 2(1), 6.

Paulus, M. P. and Yu, A. J. (2012), ‘Emotion and decision-making: Affect-driven belief systems in anxiety and depression’, Trends in Cognitive Sciences 16(9), 476–483.

Pellicano, E. and Burr, D. (2012), ‘When the world becomes “too real”: A Bayesian explanation of autistic perception’, Trends in Cognitive Sciences 16(10), 504–510.

Pellman, B. A. and Kim, J. J. (2016), ‘What can ethobehavioral studies tell us about the brain’s fear system?’, Trends in Neurosciences 39(6), 420–431.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M. and Shaevitz, J. W. (2019), ‘Fast animal pose estimation using deep neural networks’, Nature Methods 16(1), 117–125.

Perusini, J. N. and Fanselow, M. S. (2015), ‘Neurobehavioral perspectives on the distinction between fear and anxiety’, Learning & Memory 22(9), 417–425.

Peters, J. and Büchel, C. (2010), ‘Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions’, Neuron 66(1), 138–148.

Peterson, B. S., Skudlarski, P., Anderson, A. W., Zhang, H., Gatenby, J. C., Lacadie, C. M., Leckman, J. F. and Gore, J. C. (1998), ‘A functional magnetic resonance imaging study of tic suppression in tourette syndrome’, Archives of General Psychiatry 55(4), 326–333.

Pfeiffer, B. E. and Foster, D. J. (2013), ‘Hippocampal place-cell sequences depict future paths to remembered goals’, Nature 497(7447), 74–79.

Phillips, R. G. and LeDoux, J. E. (1994), ‘Lesions of the dorsal hippocampal formation interfere with background but not foreground contextual fear conditioning’, Learning & Memory 1(1), 34–44.

Phillips, R. and LeDoux, J. (1992), ‘Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning’, Behavioral Neuroscience 106(2), 274.

Pinel, J. P., Mana, M. J. and Ward, J. A. (1989), ‘Stretched-approach sequences directed at a localized shock source by rattus norvegicus’, Journal Of Comparative Psychology 103(2), 140.

Place, R., Farovik, A., Brockmann, M. and Eichenbaum, H. (2016), ‘Bidirectional prefrontal-hippocampal interactions support context-guided memory’, Nature Neuroscience 19(8), 992–994.

Popa, D., Duvarci, S., Popescu, A. T., Léna, C. and Paré, D. (2010), ‘Coherent amygdalocortical theta promotes fear memory consolidation during paradoxical sleep’, Proceedings of the National Academy of Sciences 107(14), 6516–6519.

Preston, A. R. and Eichenbaum, H. (2013), ‘Interplay of hippocampus and prefrontal cortex in memory’, Current Biology 23(17), R764–R773.

Prunell, M., Escorihuela, R., Fernandez-Teruel, A., Nunez, J. and Tobena, A. (1994), ‘Anxiolytic profiles of alprazolam and ethanol in the elevated plus maze test and the early acquisition of shuttlebox avoidance’, Pharmacological Research 29(1), 37–46.

Qi, S., Hassabis, D., Sun, J., Guo, F., Daw, N. and Mobbs, D. (2018), ‘How cognitive and reactive fear circuits optimize escape decisions in humans’, Proceedings of the National Academy of Sciences 115(12), 3186–3191.64.

Quirk, G. J. (2002), ‘Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery’, Learning & Memory 9(6), 402–407.

Ramachandran, V. S., Blakeslee, S. and Shah, N. (1998), ‘Phantoms in the brain: Probing the mysteries of the human mind’, William Morrow New York.

Raymond, J. G., Steele, J. D. and Seriès, P. (2017), ‘Modeling trait anxiety: From computational processes to personality’, Frontiers in Psychiatry 8, 1.

Redish, A. D. (1999), ‘Beyond the cognitive map: From place cells to episodic memory’, MIT press.

Redish, A. D. (2004), ‘Addiction as a computational process gone awry’, Science 306(5703), 1944–1947.

Redish, A. D. (2013), ‘The mind within the brain: How we make decisions and how those decisions go wrong’, Oxford University Press.

Redish, A. D. (2016), ‘Vicarious trial and error’, Nature Reviews Neuroscience 17(3), 147.

Redish, A. D. and Gordon, J. A. (2016), ‘Computational psychiatry: New perspectives on mental illness’, Vol. 20, MIT Press.

Redish, A. D., Jensen, S. and Johnson, A. (2008), ‘Addiction as vulnerabilities in the decision process’, Behavioral and Brain Sciences 31(4), 461–487.

Redish, A. D., Kummerfeld, E., Morris, R. L. and Love, A. C. (2018), ‘Opinion: Reproducibility failures are essential to scientific inquiry’, Proceedings of the National Academy of Sciences 115(20), 5042–5046.

Reynolds, J. H. and Heeger, D. J. (2009), ‘The normalization model of attention’, Neuron 61(2), 168–185.

Richard, J. M. and Berridge, K. C. (2011), ‘Nucleus accumbens dopamine/glutamate interaction switches modes to generate desire versus dread: D1 alone for appetitive eating but D1 and D2 together for fear’, Journal of Neuroscience 31(36), 12866–12879.

Richmond, M., Yee, B., Pouzet, B., Veenman, L., Rawlins, J., Feldon, J. and Bannerman, D. (1999), ‘Dissociating context and space within the hippocampus: Effects of complete, dorsal, and ventral excitotoxic hippocampal lesions on conditioned freezing and spatial learning’, Behavioral Neuroscience 113(6), 1189.

Rivier, C., Bruhn, T. and Vale, W. (1984), ‘Effect of ethanol on the hypothalamic-pituitary-adrenal axis in the rat: Role of corticotropin-releasing factor (CRF)’, Journal of Pharmacology and Experimental Therapeutics 229(1), 127–131.

Roberto, M., Gilpin, N. W. and Siggins, G. R. (2012), ‘The central amygdala and alcohol: Role of γ-aminobutyric acid, glutamate, and neuropeptides’, Cold Spring Harbor Perspectives in Medicine 2(12), a012195.

Robinson, T. E. and Berridge, K. C. (2001), ‘Incentive-sensitization and addiction’, Addiction 96(1), 103–114.

Rolls, E. T., Loh, M. and Deco, G. (2008), ‘An attractor hypothesis of obsessive–compulsive disorder’, European Journal of Neuroscience 28(4), 782–793.

Rothschild, G., Eban, E. and Frank, L. M. (2017), ‘A cortical–hippocampal–cortical loop of information processing during memory consolidation’, Nature Neuroscience 20(2), 251–259.

Rudy, J. W., Barrientos, R. M. and O’Reilly, R. C. (2002), ‘Hippocampal formation supports conditioning to memory of a context’, Behavioral Neuroscience 116(4), 530.

Sagvolden, T. and Sergeant, J. A. (1998), ‘Attention deficit / hyperactivity disorder: From brain dysfunctions to behaviour’, Behavioural Brain Research, 94(1), 1–10.

Saint-Cyr, J. A., Taylor, A. and Nicholson, K. (1995), ‘Behavior and the basal ganglia’, Advances in Neurology 65, 1–28.

Schacter, D. L., Addis, D. R. and Buckner, R. L. (2008), ‘Episodic simulation of future events: concepts, data, and applications’, in A. Kingstone & M. B. Miller (Eds.), Annals of the New York Academy of Sciences: Vol. 1124. The Year in Cognitive Neuroscience 2008 (p. 39–60). Blackwell Publishing.

Schmidt, B., Hinman, J. R., Jacobson, T. K., Szkudlarek, E., Argraves, M., Escabı´, M. A. and Markus, E. J. (2013), ‘Dissociation between dorsal and ventral hippocampal theta oscillations during decision-making', Journal of Neuroscience 33(14), 6212–6224.

Schmitz, T. W. and Duncan, J. (2018), 'Normalization and the cholinergic microcircuit: A unified basis for attention', Trends in Cognitive Sciences 22(5), 422–437.

Schoenbaum, G., Setlow, B., Saddoris, M. P. and Gallagher, M. (2003), 'Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala', Neuron 39(5), 855–867.

Schultz, W., Dayan, P. and Montague, P. R. (1997), 'A neural substrate of prediction and reward', Science 275(5306), 1593–1599.

Scoville, W. B. and Milner, B. (1957), 'Loss of recent memory after bilateral hippocampal lesions', Journal of Neurology, Neurosurgery, and Psychiatry 20(1), 11.

Seamans, J. K., Lapish, C. C. and Durstewitz, D. (2008), 'Comparing the prefrontal cortex of rats and primates: Insights from electrophysiology', Neurotoxicity Research 14(2), 249–262.

Seamans, J. K. and Yang, C. R. (2004), 'The principal features and mechanisms of dopamine modulation in the prefrontal cortex', Progress in Neurobiology 74(1), 1–58.

Seidenbecher, T., Laxmi, T. R., Stork, O. and Pape, H.-C. (2003), ‘Amygdalar and hippocampal theta rhythm synchronization during fear memory retrieval’, Science 301(5634), 846–850.

Seligman, M. E. (1972), ‘Learned helplessness’, Annual Review of Medicine 23(1), 407–412.

Seneca, L. A. (65 CE), Letters from a Stoic.

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P. and Dolan, R. (2012), ‘Serotonin selectively modulates reward value in human decision-making’, Journal of Neuroscience 32(17), 5833–5842.

Shannon, C. E. (1948), ‘A mathematical theory of communication’, Bell System Technical Journal 27(3), 379–423.

Sharpe, M. J. and Schoenbaum, G. (2016), ‘Back to basics: Making predictions in the orbitofrontal–amygdala circuit’, Neurobiology of Learning and Memory 131, 201–206.

Sharpe, M. J., Wikenheiser, A. M., Niv, Y. and Schoenbaum, G. (2015), ‘The state of the orbitofrontal cortex’, Neuron 88(6), 1075–1077.

Shin, J. D., Tang, W. and Jadhav, S. P. (2019), ‘Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making’, Neuron 104(6), 1110–1125.

Sierra-Mercado, D., Padilla-Coreano, N. and Quirk, G. J. (2011), ‘Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear’, Neuropsychopharmacology 36(2), 529–538.

Sierra, R. O., Pedraza, L. K., Zanona, Q. K., Santana, F., Boos, F. Z., Crestani, A. P., Haubrich, J., de Oliveira Alvares, L., Calcagnotto, M. E. and Quillfeldt, J. A. (2017), ‘Reconsolidation-induced rescue of a remote fear memory blocked by an early cortical inhibition: Involvement of the anterior cingulate cortex and the mediation by the thalamic nucleus reuniens’, Hippocampus 27(5), 596–607.

Singer, A. C. and Frank, L. M. (2009), ‘Rewarded outcomes enhance reactivation of experience in the hippocampus’, Neuron 64(6), 910–921.

Smith, A., Li, M., Becker, S. and Kapur, S. (2006), ‘Dopamine, prediction error and associative learning: A model-based account’, Network: Computation in Neural Systems 17(1), 61–84.

Smith, K. S. and Graybiel, A. M. (2013), ‘A dual operator view of habitual behavior reflecting cortical and striatal dynamics’, Neuron 79(2), 361–374.

Sommer, W., Möller, C., Wiklund, L., Thorsell, A., Rimondini, R., Nissbrandt, H. and Heilig, M. (2001), ‘Local 5, 7-dihydroxytryptamine lesions of rat amygdala: Release of punished drinking, unaffected plus maze behavior and ethanol consumption’, Neuropsychopharmacology 24(4), 430–440.

St. Onge, J. R. and Floresco, S. B. (2010), ‘Prefrontal cortical contribution to risk-based decision making’, Cerebral Cortex 20(8), 1816–1828.

Stephan, K. E., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Friston, K. J., Heinz, A., Huys, Q. J., Owen, M. J., Binder, E. B. et al. (2016), ‘Charting the landscape of priority problems in psychiatry, part 1: Classification and diagnosis’, The Lancet Psychiatry 3(1), 77–83.

Stern, C. A., Gazarini, L., Vanvossen, A. C., Hames, M. S. and Bertoglio, L. J. (2014), ‘Activity in prelimbic cortex subserves fear memory reconsolidation over time’, Learning & Memory 21(1), 14–20.

Suddendorf, T. (2013), ‘The gap: The science of what separates us from other animals’, Constellation.

Suddendorf, T., Addis, D. R. and Corballis, M. C. (2009), ‘Mental time travel and the shaping of the human mind’, Philosophical Transactions of the Royal Society B: Biological Sciences 364(1521), 1317–1324.

Suddendorf, T. and Corballis, M. C. (2007), ‘The evolution of foresight: What is mental time travel, and is it unique to humans?’, Behavioral and Brain Sciences 30(3), 299–313.

Sugam, J. A., Saddoris, M. P. and Carelli, R. M. (2014), ‘Nucleus accumbens neurons track behavioral preferences and reward outcomes during risky decision making’, Biological Psychiatry 75(10), 807–816.

Sun, C., Yang, W., Martin, J. and Tonegawa, S. (2020), ‘Hippocampal neurons represent events as transferable units of experience’, Nature Neuroscience 23(5), 651–663.

Sutton, R. S., Barto, A. G. et al. (1998), ‘Introduction to reinforcement learning, Vol. 2’, MIT Press Cambridge.

Swain, J. E., Scahill, L., Lombroso, P. J., King, R. A. and Leckman, J. F. (2007), ‘Tourette syndrome and tic disorders: A decade of progress’, Journal of the American Academy of Child & Adolescent Psychiatry 46(8), 947–968.

Takahashi, S. (2015), ‘Episodic-like memory trace in awake replay of hippocampal place cell activity sequences’, Elife 4, e08105.

Tolman, E. C. (1939), ‘Prediction of vicarious trial and error by means of the schematic sowbug’, Psychological Review 46(4), 318.

Tovote, P., Fadok, J. P. and Lüthi, A. (2015), ‘Neuronal circuits for fear and anxiety’, Nature Reviews Neuroscience 16(6), 317–331.

Treit, D. and Menard, J. (1997), ‘Dissociations among the anxiolytic effects of septal, hippocampal, and amygdaloid lesions’, Behavioral Neuroscience 111(3), 653.

Treit, D., Menard, J. and Royan, C. (1993), ‘Anxiogenic stimuli in the elevated plus maze’, Pharmacology Biochemistry and Behavior 44(2), 463–469.

Treit, D., Robinson, A., Rotzinger, S. and Pesold, C. (1993), ‘Anxiolytic effects of serotonergic interventions in the shock-probe burying test and the elevated plus maze test’, Behavioural Brain Research 54(1), 23–34.

Tripp, G. and Wickens, J. R. (2008), ‘Research review: Dopamine transfer deficit: A neurobiological theory of altered reinforcement mechanisms in ADHD’, Journal of Child Psychology and Psychiatry 49(7), 691–704.

Trivedi, M. A. and Coover, G. D. (2004), ‘Lesions of the ventral hippocampus,but not the dorsal hippocampus, impair conditioned fear expression and inhibitory avoidance on the elevated t-maze’, Neurobiology of Learning and Memory 81(3), 172–184.

Tsibulsky, V. L. and Norman, A. B. (1999), ‘Satiety threshold: a quantitative model of maintained cocaine self-administration’, Brain Research 839(1), 85–93.

Tulving, E. (1985), ‘Memory and consciousness’, Canadian Psychology/Psychologie Canadienne 26(1), 1.

Van Boxtel, J. J. and Lu, H. (2013), ‘A predictive coding perspective on autism spectrum disorders’, Frontiers in Psychology 4, 19.

Van der Poel, A. (1979), ‘A note on “stretched attention”, a behavioural element indicative of an approach-avoidance conflict in rats’, Animal Behaviour 27, 446–450.

Vanvossen, A. C., Portes, M. A., Scoz-Silva, R., Reichmann, H. B., Stern, C. A. and Bertoglio, L. J. (2017), ‘Newly acquired and reactivated contextual fear memories are more intense and prone to generalize after activation of prelimbic cortex nmda receptors’, Neurobiology of Learning and Memory 137, 154–162.

Verduzco-Flores, S., Ermentrout, B. and Bodner, M. (2012), ‘Modeling neuropathologies as disruption of normal sequence generation in working memory networks’, Neural Networks 27, 21–31.

Vinogradov, S. (2017), ‘The golden age of computational psychiatry is within sight’, Nature Human Behaviour 1.2 (2017): 1-3.

Walker, D. L. and Davis, M. (2008), ‘Role of the extended amygdala in short-duration versus sustained fear: A tribute to Dr. Lennart Heimer’, Brain Structure and Function 213(1-2), 29–42.

Walters, C. J., Jubran, J., Sheehan, A., Erickson, M. T. and Redish, A. D. (2019), ‘Avoid-approach conflict behaviors differentially affected by anxiolytics: Implications for a computational model of risky decision-making’, Psychopharmacology 236(8), 2513–2525.

Walters, C. J. and Redish, A. (2018), ‘A case study in computational psychiatry: Addiction as failure modes of the decision-making system’, in Computational Psychiatry, Academic Press, 199-217.

Walton, M. E., Rudebeck, P. H., Bannerman, D. M. and Rushworth, M. F. (2007), ‘Calculating the cost of acting in frontal cortex’, Annals of the NewYork Academy of Sciences 1104, 340.

Wang, Q., Jin, J. and Maren, S. (2016), ‘Renewal of extinguished fear activates ventral hippocampal neurons projecting to the prelimbic and infralimbic cortices in rats’, Neurobiology of Learning and Memory 134, 38–43.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N. and Behrens, T. E. (2020), ‘The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation’, Cell 183(5), 1249–1263.

Williams, J. and Dayan, P. (2005), ‘Dopamine, learning, and impulsivity: A biological account of attention-deficit/hyperactivity disorder’, Journal of Child & Adolescent Psychopharmacology 15(2), 160–179.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N. and O’Keefe, J. (2005), ‘Attractor dynamics in the hippocampal representation of the local environment’, Science 308(5723), 873–876.

Wilson, M. A., Burghardt, P. R., Ford, K. A., Wilkinson, M. B. and Primeaux, S. D. (2004), ‘Anxiolytic effects of diazepam and ethanol in two behavioral models: Comparison of males and females’, Pharmacology Biochemistry and Behavior 78(3), 445–458.

Wscieklica, T., de Barros Viana, M., Maluf, L. L. S., Pouza, K. C. P., Spadari, R. C. and Céspedes, I. C. (2016), ‘Alcohol consumption increases locomotion in an open field and induces fos-immunoreactivity in reward and approach / withdrawal-related neurocircuitries’, Alcohol 50, 73–82.

Wu, C.-T., Haggerty, D., Kemere, C. and Ji, D. (2017), ‘Hippocampal awake replay in fear memory retrieval’, Nature Neuroscience 20(4), 571–580.

Wu, J. Q., Szpunar, K. K., Godovich, S. A., Schacter, D. L. and Hofmann, S. G. (2015), ‘Episodic future thinking in generalized anxiety disorder’, Journal of Anxiety Disorders 36, 1–8.

Xu, C., Krabbe, S., Gründemann, J., Botta, P., Fadok, J. P., Osakada, F., Saur, D., Grewe, B. F., Schnitzer, M. J., Callaway, E. M. et al. (2016), ‘Distinct hippocampal pathways mediate dissociable roles of context in memory retrieval’, Cell 167(4), 961–972.

Xu, W. and Südhof, T. C. (2013), ‘A neural circuit for memory specificity and generalization’, Science 339(6125), 1290–1295.

Yeung, M., Treit, D. and Dickson, C. T. (2012), ‘A critical test of the hippocampal theta model of anxiolytic drug action’, Neuropharmacology 62(1), 155–160.

Yokota, S., Suzuki, Y., Hamami, K., Harada, A. and Komai, S. (2017), ‘Sex differences in avoidance behavior after perceiving potential risk in mice’, Behavioral and Brain Functions 13(1), 9.72.

Zalocusky, K. A., Ramakrishnan, C., Lerner, T. N., Davidson, T. J., Knutson, B. and Deisseroth, K. (2016), ‘Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making’, Nature 531(7596), 642–646.

Zeeb, F. D., Baarendse, P., Vanderschuren, L. and Winstanley, C. A. (2015), ‘Inactivation of the prelimbic or infralimbic cortex impairs decision-making in the rat gambling task’, Psychopharmacology 232(24), 4481–4491.

Zentall, T. R. (2006), ‘Mental time travel in animals: A challenging question’, Behavioural Processes 72(2), 173–183.

Zick, J. L., Blackman, R. K., Crowe, D. A., Amirikian, B., DeNicola, A. L., Netoff, T. I. and Chafee, M. V. (2018), ‘Blocking NMDAR disrupts spike timing and decouples monkey prefrontal circuits: Implications for activity-dependent disconnection in schizophrenia’, Neuron 98(6), 1243–1255.

Appendices

Appendix A

The basics of computational modeling

Organisms are typically modeled as open systems; that is, they are systems that take inputs (e.g., environmental stimuli) and produce outputs (e.g., behavior). Between the environmental inputs and the behavioral outputs, there is a critical intermediate step: namely, the system must process the information it receives and transform it in such a way so as to guide behavior. Crucially, we (the observer) do not have direct access to this information-processing step – the system states (which change over the period of observation) and their parameter values (which do not change over the period of observation) are hidden from the observer. Critically, these hidden states (often referred to as latent states or latent variables) influence the nature of the system's outputs (i.e., the behavior that we can observe and measure, Fig. A.1).

As modelers, we want to develop methods by which we can infer the configuration of these underlying latent variables that best explain the output of the system. These latent variables are governed by parameters that dictate the nature of the observed data that we have access to, and it is the goal of the model to uncover how these latent variables are coupled and what parameter values best explain the observed data and predict the system's behavior at future time points. This is, generally speaking, achieved by minimizing the error between the observed data and the predictions made by the model.

There are two fundamental equation structures that one encounters in most computational models: differential equations for modeling continuous time data and difference equations for modeling discrete time data. A differential equation (which models how a variable of interest, x, evolves over continuous time) takes the following general form:

$$\frac{dx}{dt} = f(z(t), \theta, I(t))$$

where z(t) is the latent variable vector z(t) = (z_1(t), z_2(t), ..., z_n(t)) over continuous time t; θ refers to the set of latent variable parameters; and I(t) = (I_1(t), I_2(t), ..., I_n(t)), reflecting the input into the system over continuous time t (Fig. A.1). Altogether then, how some observable variable of interest changes over time is a function of the latent variables, their parameters, and the inputs the system receives. Example data sets that would require continuous time models are EEG, fMRI, and LFP time-series data. A model of this form, could, for example, capture how an observable variable of interest – such as an fMRI BOLD voxel – changes in intensity over time as a function of a given input (e.g., an experimenter-controlled stimulus such as an image or sound) and its estimated latent variable values (e.g., the number and type of cells in the voxel, their receptor profile, etc.).

Difference equations, on the other hand, are used for modeling data that evolve over discrete time, and differs from the above differential form in the following way:

$$x(t + 1) = f(z(t), \theta, I(t))$$

Notice how, unlike the differential model, time in the difference model progresses at segmented intervals (e.g., t+1, t+2, … , t+n). Reinforcement learning models typically utilize discrete time models to characterize how a system learns and behaves as it interacts with the environment and advances through discrete state transitions. Now that we have outlined what a model is and introduced some of the basic terms, we will now look at how to build a model and use it to estimate latent variable parameter values.

Forward models and the inverse problem

When modeling a system, we are trying to identify the latent variable settings that are causing the observations we measure. In the case of psychiatry, a clinician is trying to find a pathophysiological profile (i.e., the latent variables and their likely values) that best explains a given patient's symptomatology (i.e., the observable behavior of the system). Given this correspondence, we can cast the diagnostic process as a modeling problem wherein the physician is attempting to infer the state of the latent variables of the patient.

Models can run both forward and backward. A forward model is used to simulate data and use that simulated data to predict the behavior of the system. That is:

$$model\,parameters \;\rightarrow\; simulated\;data$$

A forward model is simply composed of an array of latent variables and their parameter values. A fundamental insight that computational modeling brings to the table is that we can invert the directionality and infer latent variable parameter values given some set of observations. That is:

$$actual\,data \rightarrow inferred\,model\,parameters$$

Solving this inverse mapping problem simply requires fitting the model's parameters to the data we observe. By taking a forward model and exploring various parameter values for the set of parameters θ, we can generate data and compare that simulated data to observed data. This allows us to then assign a value, known as the likelihood, to the probability that we would see the data we observe given a set of coupled latent variables and an estimate of their parameter values θ. By exploring various parameter values for a given model, typically using an optimization algorithm, one can identify the parameter value configuration that provides the highest probability of generating the observed data. This procedure is referred to as maximum likelihood estimation, and is denoted:

$$\widehat{\theta}_{ML} = argmax_{\theta}\{L(\theta, x)\}$$

where $\widehat{\theta}_{ML}$ indicates the specific configuration of the model's parameter values θ that maximize the likelihood L of seeing some observed data x. Thus, a forward model furnishes us with the ability to simulate consequences (e.g., symptoms) from causes (e.g., underlying pathophysiology).

Almost every model will have some degree of explanatory power, however, and there is considerable pathophysiological heterogeneity between patients which will be best captured with different models. For these reasons, it is important to use more than one model, compare them, and select the best one. With a forward model in hand, we can now perform such model comparisons by using the maximum likelihood estimation (MLE) metric to select among them. It turns out that this is not a very effective method for model selection, however, and while MLE is valuable for providing an intuition for how model parameters can be estimated, model comparison is rarely done using MLE owing to a variety of shortcomings associated with this approach (e.g., it is prone to overfit data). Fortunately, more robust parameter inference methods exist, namely Bayesian models (see Appendix B).

Having outlined the basics of computational modeling, we can now examine how these models can be applied to better our understanding of psychiatric disease.

Appendix B

Current challenges in psychiatry

The goal of computational psychiatry is, of course, to improve our understanding of psychiatric disorders so that we may develop new effective treatments and improve the quality of life of patients. The growing body of evidence that I outlined in Chapter 2 strongly suggests that: **(1)** psychiatric dysfunction is due to a maladaptive interaction between underlying brain information processing vulnerabilities and the environment; **(2)** we should guide treatment development to address the underlying information processing dysfunction(s) in the brain that are relevant to a given patient; and **(3)** appropriate tests can likely be developed that will allow us to identify information processing vulnerabilities in an individual, gauge risks or future maladaptive behavior, and provide the possibility of prevention.

The standard model in psychiatric nosology has held that categorical descriptions furnished by the Diagnostic and Statistical Manual of Mental Disorders (e.g., agoraphobia, trichotillomania, depersonalization, bulimia nervosa, etc.) map onto a set of hidden physiological causes generating the psychiatric condition under consideration; however, this does not appear to be the case, since different patients diagnosed with the same psychiatric disorder often exhibit a wide range of varying cognitive and physiologic measures. Likewise, patients from different diagnostic categories can exhibit very similar cognitive and physiologic symptoms. This phenomenon is described by the principles of equifinality and multifinality – the notion that, in a complex open system, many unique pathways (sets of dysfunctions) result in the same outcome (the same symptoms), and any given dysfunction can give rise to multiple divergent observations (symptoms), respectively.

A new approach to psychiatric nosology: the Bayesian Integrative Framework

To capture the full complexity of psychiatric nosology, we need to recognize tiers of causal influence in the origin, instantiation, and symptomatology of psychiatric disease (Flagel et al., 2016). In this novel framework, putative causes lead to hidden physiological states, physiological states relate to a range of continuously distributed latent variables, and latent variables give rise to symptoms which form the basis of categorical and dimensional assignments made by physicians (Fig. B.1). Latent variables are akin to the dimensional constructs provided by the Research Domain Criteria approach (NIMH, 2019b) (reward responsiveness, cognitive control, perception of self and others, habit learning, threat reactivity, etc.), which are grounded in a complex milieu of putative causes (genetics, prenatal and perinatal factors, trauma, developmental experiences, etc.) and difficult-to-observe physiological states (aberrant neurotransmission, synaptic dysregulation, glial dysfunction, functional hypo- or hyper-connectivity across networks, etc.).

The Bayesian Integrative Framework builds on the clinical observations obtained from a patient. These include putative observable causes (e.g., risk genes,environmental insults, exposure to trauma, etc.), symptoms (e.g., hallucinations and their characteristics, depressed mood and its persistence, etc.), and how responsive symptoms have been to specific treatments. A generative model can then factor in these data and make probabilistic inferences (i.e., the model can infer the posterior distributions over the parameters governing the latent variables) regarding the patient's location in latent variable space – which is analogous to the concept of diagnoses – and their most likely trajectory through that space – their prognosis. This can then inform the prescription of treatment (see Fig. B.1). Furthermore, Bayesian models provide a method by which to compare models and determine which one offers the best fit to the data (is the most accurate) but also requires the lowest dimensional parameter space (is the least complex), a procedure which is critical given the fact that any model will have some explanatory power (see Appendix A). Of course, one would not expect the clinician to do these calculations explicitly, but they can be factored into computerized decision-support systems (such as apps) derived from these generative models.

How to estimate the posterior and compare Bayesian models

Using the above Bayesian framework, how would a clinician (or, more accurately, the software that the clinician is using) go about inferring the posterior distributions over a model's latent variables in light of new observations (e.g., new test results, new symptoms, response to treatment, etc.)? How exactly would one compare two or more of these models in order to identify the one best supported by the available evidence? Given some set of parameter values θ (e.g., the estimated latent variable values in a patient's generative model) and some data set D (i.e., putative causes, symptoms, diagnoses), Bayes' theorem allows us to reason about the probability that a model with the parameter values we have defined would generate the data we are observing. The evidence in favor of a given model, called the model evidence or marginal likelihood, is calculated by simply marginalizing (i.e., summing out) all the parameters to arrive at the probability of observing the data p(D), that is:

$$model\,evidence \;=\; \int p(D, \theta)\; d\theta$$

Bayesian models often incorporate the model evidence into the same general form that we saw in Chapter 2. The model evidence, which increases as the model becomes more accurate and is penalized as the model becomes more complex, is a metric which can be used to compare models and identify the most predicative and parsimonious among the models being considered. Putting it all together then, we have the following general form for describing a Bayesian model:

$$p(\theta|D) \;= \frac{p(D|\theta)\,p(\theta)}{\int p(D,\theta)\; d\theta}$$

If, as is often the case, the model being used has many latent variable parameters, the model evidence will become a high-dimensional integral, thus making the posterior hard (or impossible) to solve analytically. However, methods exist for approximating the posterior, and these methods form the foundation of most Bayesian models. Two such posterior approximation methods – Markov Chain Monte Carlo and variational inference – are so ubiquitous in the fields of theoretical neuroscience and computational psychiatry that they warrant further discussion. In brief, Monte Carlo methods approximate the posterior via random sampling, while variational methods approximate the posterior by varying the parameters of a simpler, known distribution until it closely matches the posterior.

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) refers to a family of algorithms that allow you to sample from the posterior distribution via an intelligent random search process. Here, we will discuss a common MCMC implementation known as the Metropolis-Hastings algorithm. Under the Metropolis-Hastings algorithm, a value in high-dimensional parameter space is selected at random from a known distribution (i.e., a normal distribution referred to as the proposal distribution). An acceptance rule based on the ratio of the current sample from the posterior over the previous sample from the posterior is applied to this random draw from the proposal distribution to determine whether it is accepted (i.e., a ratio ≥ 1) or rejected (i.e., a ratio < 1). If accepted, the previous value becomes the new mean of the proposal distribution for the next random draw, giving rise to a random walk.

This Monte Carlo (i.e., random sampling) process repeats arbitrarily many times, forming a chain of samples which are used to estimate the true posterior according to this algorithm which essentially biases the samples to occur in areas with a higher posterior probability (Fig. B.2). After running multiple chains that converge on the same distribution, the end result of the Metropolis-Hastings algorithm is a distribution that can be quantified and used to infer the shape of the true posterior distribution.

Variational inference

The goal of variational inference is to vary the parameters of the approximate posterior q(θ), a distribution of the model's parameter estimates, with the aim of making it as close as possible to the true posterior, p(θ|D). This is captured by the following equation:

$$model\,evidence\ =\ KL(q\,||\,p)\ +\ lower\ bound$$

where the model evidence is the marginal likelihood of observing the data (for a given model); KL is the Kullback-Leibler divergence, a non-negative, non-symmetric measure of how similar two probability distributions are (e.g., the approximate posterior and the true posterior) and which is equal to zero when the two distributions are identical; and the lower bound is a constraint on how small the model evidence can be, such that the model evidence ≥ the lower bound. Intuitively, we want to simply minimize the KL divergence between the approximate posterior and the true posterior, but we cannot do this owing to the implicit dependence of the KL divergence on directly computing the model evidence. We can, however, set the approximate posterior equal to the expectation (i.e., the probability-weighted average) of the likelihood (Fig. B.3), a term that is independent from the model evidence. By adjusting the parameters of the approximate posterior, we can then maximize the likelihood of observing the data given our parameter estimates $\widehat{\theta}$. Maximizing this expectation of the likelihood term, which is the lower bound on the model evidence, is equivalent to minimizing the KL divergence but does not require that we analytically solve the model evidence. As a result, this method of estimating the true posterior is known as the expectation-maximization algorithm.

Taken together then, as the lower bound is iteratively maximized it approaches the model evidence (which is equivalent to the KL divergence between the approximate posterior and the true posterior approaching zero), the result being an increasingly accurate approximation to the true posterior distribution without ever having to directly compute the model evidence (Fig. B.3).

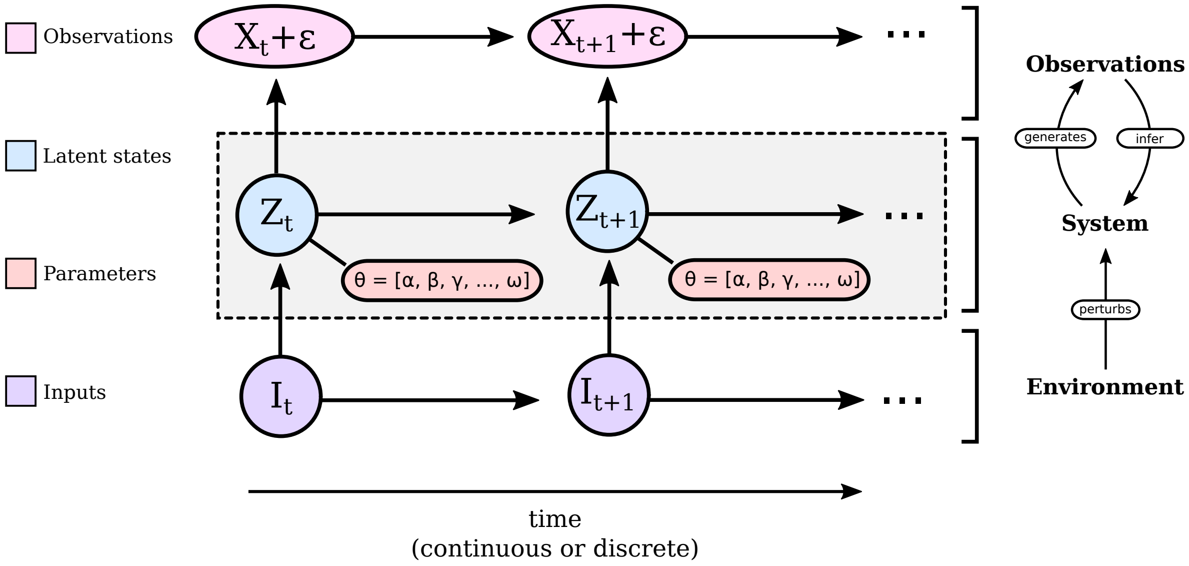

Observations
Latent states
Parameters
Inputs
$X_t+\varepsilon$
$X_{t+1}+\varepsilon$
...
Z_t
Z_{t+1}
...
θ = [α, β, γ, ..., ω]
θ = [α, β, γ, ..., ω]
I_t
I_{t+1}
...
Observations
generates
infer
System
perturbs
Environment
time
(continuous or discrete)

Figure A.1 A schematic of a computational model. The system under investigation (the gray box) receives inputs (I) and transforms those inputs into some measurable observation (X, plus some noise ε). The aim of a given model is to identify and characterize the parameters (θ) underlying the latent states (Z) of the system such that the model is able to explain and/or predict the observations that the system generates.

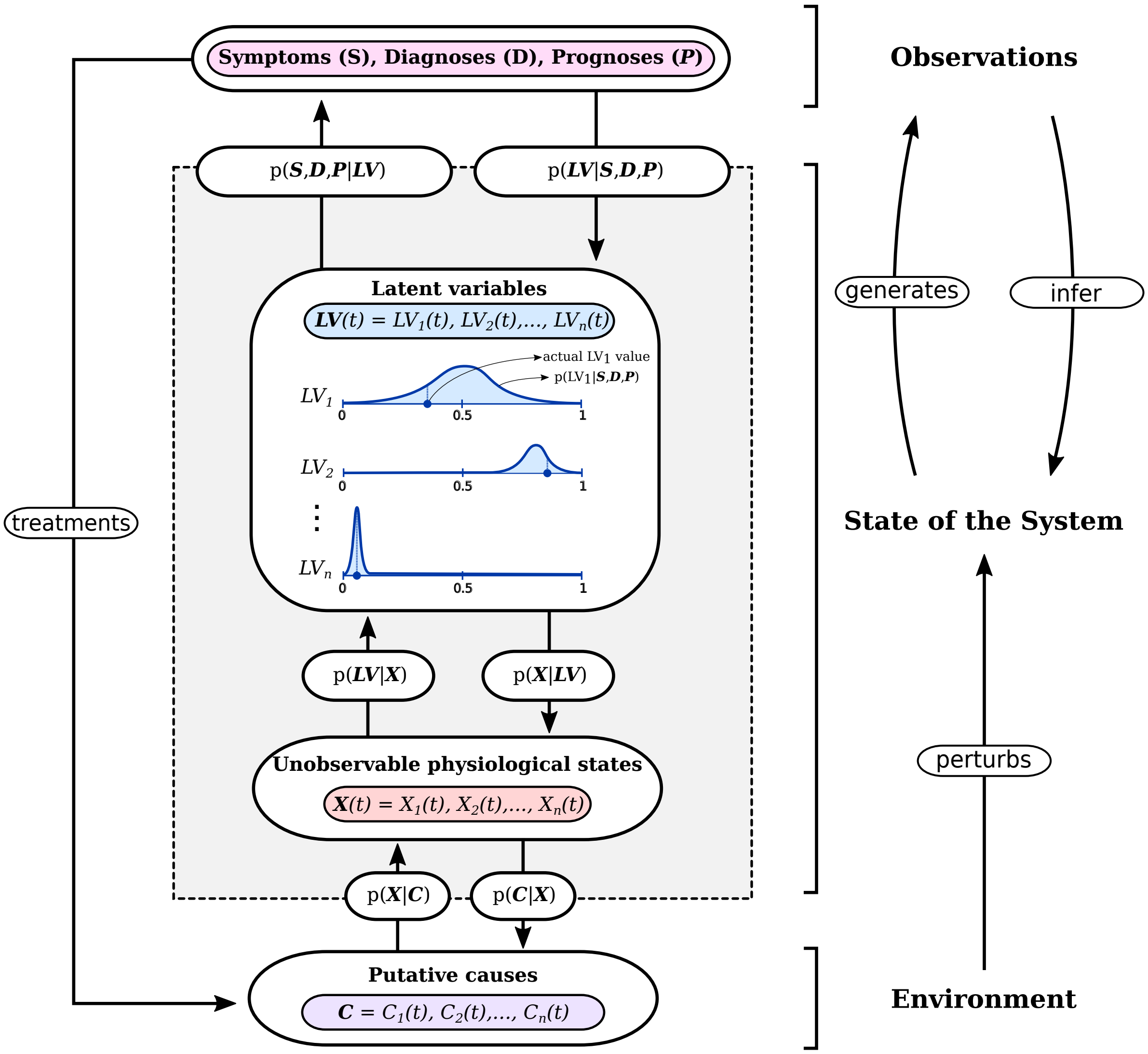

Symptoms (S), Diagnoses (D), Prognoses (*P*)
Observations
p(***S***,***D***,***P***|***LV***)
p(***LV***|***S***,***D***,***P***)
generates
infer
Latent variables
LV(*t*) = *LV*$_1$(*t*), *LV*$_2$(*t*),..., *LV*$_n$(*t*)
actual LV$_1$ value
p(LV$_1$|***S***,***D***,***P***)
LV$_1$
0
0.5
1
LV$_2$
0
0.5
1
⋮
LV$_n$
0
0.5
1
treatments
State of the System
p(***LV***|***X***)
p(***X***|***LV***)
perturbs
Unobservable physiological states
X(*t*) = *X*$_1$(*t*), *X*$_2$(*t*),..., *X*$_n$(*t*)
p(***X***|***C***)
p(***C***|***X***)
Putative causes
C = *C*$_1$(*t*), *C*$_2$(*t*),..., *C*$_n$(*t*)
Environment

Figure B.1 Putative causes engender unobservable (or difficult to observe) physiological changes which in turn affect a range of latent variables (where the blue dots indicate the patient's actual position along a given latent variable and the clinical estimate of that position is depicted as a probability distribution over that variable). The patient's position in latent variable space influences their symptoms and subsequent diagnoses and prognoses, and treatments themselves feed back into the list of putative causes.

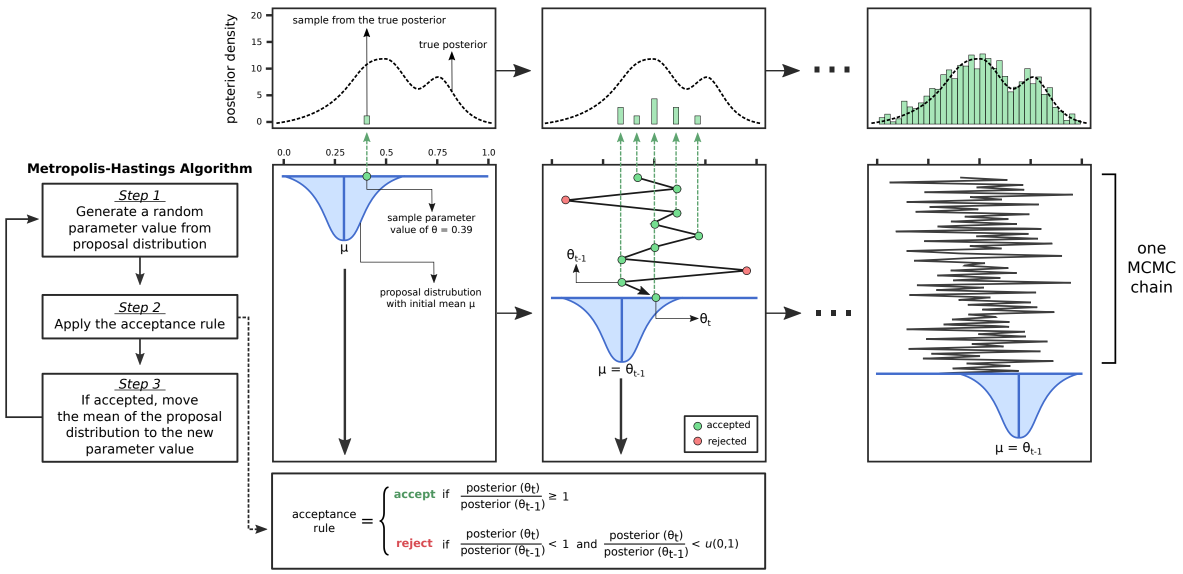

Metropolis-Hastings Algorithm
Step 1
Generate a random parameter value from proposal distribution
Step 2
Apply the acceptance rule
Step 3
If accepted, move the mean of the proposal distribution to the new parameter value
posterior density
20
15
10
5
0
sample from the true posterior
true posterior
0.0
0.25
0.5
0.75
1.0
sample parameter value of θ = 0.39
μ
proposal distrubution with initial mean μ
θ_{t-1}
θ_t
$\mu = \theta_{t-1}$
● accepted
● rejected
one MCMC chain
$\mu = \theta_{t-1}$
acceptance rule = accept if $\frac{\text{posterior }(\theta_t)}{\text{posterior }(\theta_{t-1})} \geq 1$
reject if $\frac{\text{posterior }(\theta_t)}{\text{posterior }(\theta_{t-1})} < 1$ and $\frac{\text{posterior }(\theta_t)}{\text{posterior }(\theta_{t-1})} < u(0,1)$

Figure B.2 The Metropolis-Hastings algorithm is a common method for sampling from the posterior distribution when it is analytically intractable. Both the Metropolis-Hastings algorithm and the expectation-maximization algorithm (see Supp. Fig. 4) are useful tools when dealing with high-dimensional posteriors (see Supp. Fig 2) that are difficult or impossible to sample from directly.

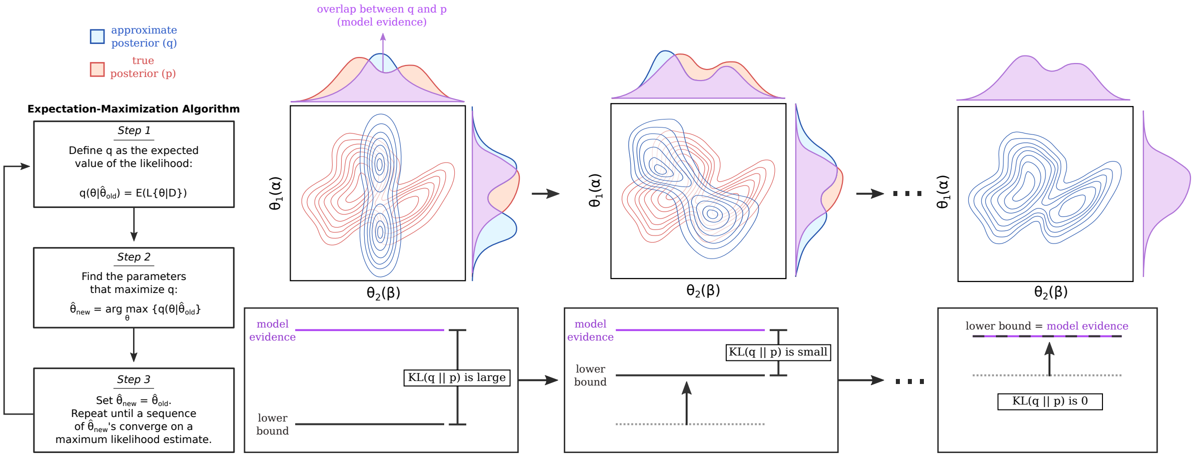

approximate posterior (q)
true posterior (p)
overlap between q and p (model evidence)
Expectation-Maximization Algorithm
Step 1
Define q as the expected value of the likelihood:
$q(\theta|\hat{\theta}_{old}) = E(L\{\theta|D\})$
Step 2
Find the parameters that maximize q:
$\hat{\theta}_{new} = \arg\max_{\theta} \{q(\theta|\hat{\theta}_{old})\}$
Step 3
Set $\hat{\theta}_{new} = \hat{\theta}_{old}$.
Repeat until a sequence of $\hat{\theta}_{new}$'s converge on a maximum likelihood estimate.
$\theta_1(\alpha)$
$\theta_2(\beta)$
$\theta_1(\alpha)$
$\theta_2(\beta)$
$\theta_1(\alpha)$
$\theta_2(\beta)$
model evidence
KL(q || p) is large
lower bound
model evidence
KL(q || p) is small
lower bound
lower bound = model evidence
KL(q || p) is 0

Figure B.3 The expectation-maximization algorithm is another common approach for approximating from the posterior distribution. Unlike the Metropolis-Hastings algorithm, the expectation-maximization algorithm does not sample from the posterior, but rather tweaks the parameters of a simpler, known distribution until it approximates the more complex, unknown posterior distribution.