Trends in Cognitive Sciences

CellPress

Feature Review

Neural Computations of Threat

Ifat Levy[1,*] and Daniela Schiller[2,*]

A host of learning, memory, and decision-making processes form the individual's response to threat and may be disrupted in anxiety and post-trauma psychopathology. Here we review the neural computations of threat, from the first encounter with a dangerous situation, through learning, storing, and updating cues that predict it, to making decisions about the optimal course of action. The overview highlights the interconnected nature of these processes and their reliance on shared neural and computational mechanisms. We propose an integrative approach to the study of threat-related processes, in which specific computations are studied across the various stages of threat experience rather than in isolation. This approach can generate new insights about the evolution, diagnosis, and treatment of threat-related psychopathology.

Highlights

The response to threat comprises multiple learning, memory, and decision-making processes.

These processes may be disrupted in anxiety and trauma-related disorders.

We describe five stages of processing: experience of imminent threat; formation of threat associations; post-association learning; storing and updating of these associations; and decision-making under threat.

These stages rely on overlapping computations and shared neural circuits.

We propose that, to reach a fundamental understanding of anxiety and trauma-related disorders, these processes should be studied together rather than in isolation.

Neural Computations of Threat: Learning, Memory, and Decision Making

How does the brain compute threat? When facing danger, an organism is tasked with learning precursors of threat, remembering those precursors for extended periods and across contexts, and making optimal decisions under stress. Anxiety and trauma can impact each of these cognitive processes. Here we define threat as an organism, an object, or a situation that is likely to inflict damage on an organism's physical or mental wellbeing. We review the neural computations underlying adaptive and maladaptive threat learning, memory, and decision making. Beginning with an encounter with a threatening situation, we follow the associations born out of this event, the elaboration of these associations, the formation and reformation of threat memories, and how threat experience shapes decision making (Figure 1). For each phase, we describe the neural computations performed on incoming, retrieved, or projected information, and their manifestation in post-traumatic stress disorder (PTSD) and anxiety disorders.

The juxtaposition of the various stages of threat experience highlights the interconnected nature of these processes, with common computations and overlapping neural regions (Figure 2 and Box 1). At the heart of this process are computations during ambiguous situations, where uncertainty could be reduced through information gathering, proactive anticipation of consequences, and the retrieval and updating of relevant memories, for the purpose of making predictions and choices more accurate. We use the term 'computation' in accordance with Marr's three levels [1], whereby the term refers to the goal of the computation and the logic by which it can be performed. We argue that the different stages of threat experience share computational goals, and therefore the algorithmic and implementation levels could also overlap (Box 2). This set of computations ought to comprise the values of predictive cues, actions and outcomes, **prediction error** (see Glossary) that drives learning, dynamically adjusted learning rates, and the uncertainty surrounding these estimations. These values could be learned through various policies such as trial and error or by learning about the structure of the environment.

Although these computations are deployed throughout the threat experience, the information they process differs depending on the phase, be it the initial encounter, subsequent learning, memory retrieval, or decision making. This approach generates interesting predictions about how clusters of symptoms may organize, and proposes considerations for diagnosis and

[1]Departments of Comparative Medicine, Neuroscience, and Psychology, Yale University, New Haven, CT, USA
[2]Department of Psychiatry, Department of Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

*Correspondence:
ifat.Levy@yale.edu (I. Levy) and
daniela.schiller@mssm.edu (D. Schiller).

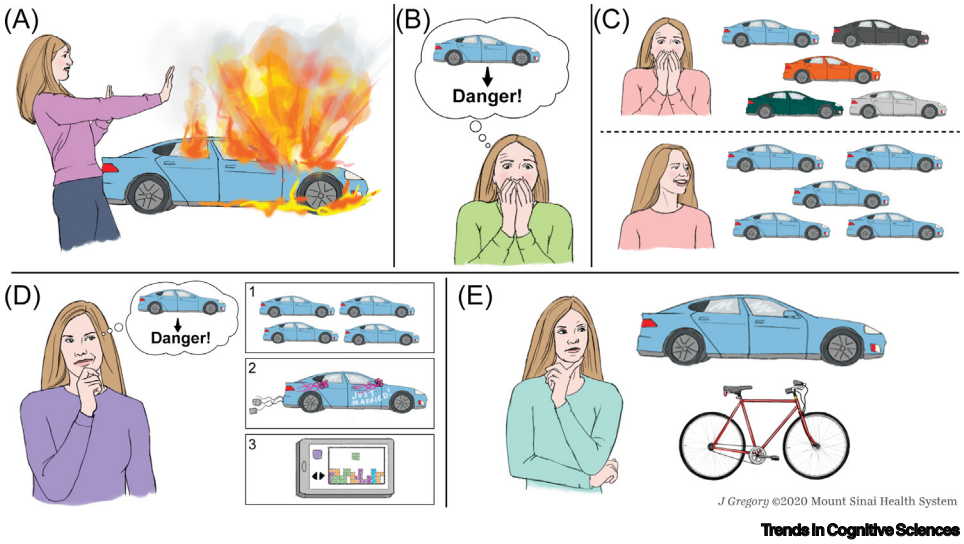

Figure 1. The Stages of Threat Experience. Experiencing a life-threatening event, in this case an aversive emotional memory of a car explosion (A), may result in associative learning (B) where a neutral stimulus (the blue car) becomes threatening as it predicts danger (explosion). The learned association then competes with or influences new associations (C). For example, generalization of the association to other stimuli [(C), top] or extinction learning [(C), bottom], where repeated exposure to blue cars diminishes the threat response, may occur. A more permanent way of diminishing the learned threat response is by modifying the original association through reconsolidation updating (D). A reminder cue may trigger the memory and destabilize it, requiring restabilization (reconsolidation) to return it to a stable state. In the course of destabilization, updates may occur in several ways, such as extinction (top), counterconditioning (middle; car associated with a positive outcome such as a wedding), or sensorimotor interference (depicted here as a Tetris game). The new information these processes provide is incorporated into the memory (extinction, counterconditioning) or depletes neural resources of reconsolidation (sensorimotor interference). Finally, threat learning interacts with processes of decision making and attitudes toward loss, risk, and ambiguity (E). For example, when facing a choice between riding in a car or on a bicycle, threat-related processes may bias the choice toward the less threatening option. The depiction of the stages of threat experience (A–E) does not mean to indicate any sequential order or independence. The stages are intertwined throughout the threat experience.

treatment, as the following sections elaborate. This overview suggests that rather than treating anxiety and PTSD as disorders of multiple distinct processes – heightened emotional reactivity, aberrant learning, impaired inhibition, overgeneralization, hyperavoidance, maladaptive memories, or biased decision making – a unifying approach may prove more efficient. From a computational standpoint, anxiety and PTSD are disorders of prediction – the estimation of future threats.

The Stages of Threat Experience

The Experience of Imminent Threat

Encountering a life-threatening situation engages neural computations that consider information about the environment and the source of threat. These computations design defensive policies and select adaptive responses for execution. The **threat imminence continuum** model [2], which maps defensive behaviors onto levels of threat imminence (how far away a predator is in time and space), provides a platform for prey–predator relations to assess the neural circuits and computations for survival [3,4]. In the first, 'safe' stage, there is no threat, but an encounter with a predator may occur in the distant future. Individuals may experience occasional anxiety and flash forward toward possible future threats. Cognitive control and emotion regulation could keep this process in check. Next is the 'pre-encounter threat' – the predator is not present, but may surface at any moment. Individuals may experience anticipatory anxiety and exhibit vigilance and preparatory behaviors. In the more dangerous 'post-encounter threat' the prey, not yet detected, observes the predator. This step generates encounter anxiety, involving close

Glossary

Ambiguity: a type of uncertainty in which likelihoods for potential outcomes are not precisely known; equivalent to estimation uncertainty.
Associability: the degree to which a cue has previously been accompanied by a surprising outcome. Can be used to gate the learning rate in models of reinforcement learning.
Avoidance learning: learning a response to evade aversive outcomes or the stimuli that predict them.
Consolidation: a period in which a newly acquired memory is stabilized.
Counterconditioning: pairing a threat-conditioned stimulus with a rewarding outcome (or a reward cue with an aversive outcome) to weaken a learned association.
Destabilization: the shift of a reactivated memory into an unstable state, making it susceptible to change.
Extinction learning: repeated presentations of a previously conditioned stimulus, in the absence of the associated outcome, result in a temporary decline of memory expression.
Instrumental conditioning: learning to associate actions with aversive or appetitive outcomes.
Loss aversion: the relative weight that the decision maker assigns to losses compared with gains.
Model-based computations: learning algorithms that rely on internal models of the environment, including action-state transition probabilities.
Model-free computations: learning algorithms that learn the expected value of actions or stimuli based on sampling and direct experience of outcomes.
Pavlovian conditioning: learning to associate stimuli with aversive or appetitive outcomes.
Prediction error: the difference between the obtained and expected outcomes, used to drive learning in reinforcement learning models.
Predictive value: in the context of models of aversive learning, the degree of threat predicted by a cue.
Reactivation: exposure to memory reminders, which may lead to destabilization of the neural representation of the memory.
Reconsolidation: the process of restabilization of a destabilized memory, allowing to update it with new information. Disruption of the

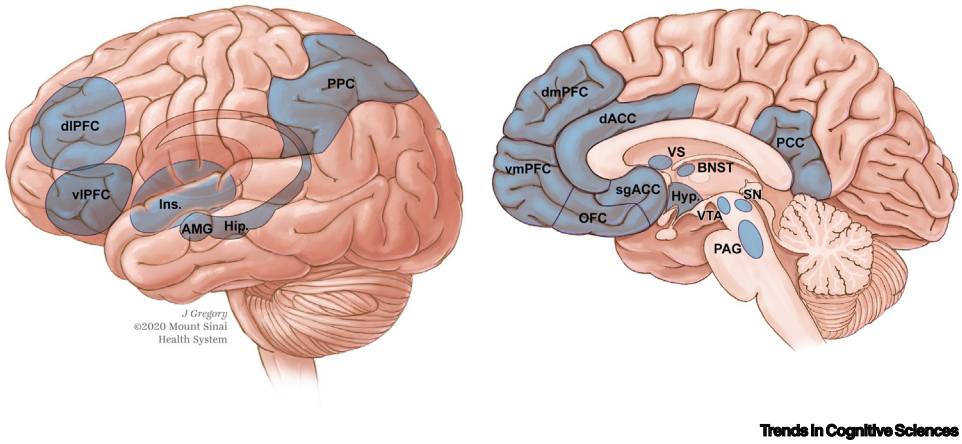

Figure 2. Neural Basis of Threat. Brain schema depicting working hypotheses, based on extant evidence, for brain regions and neural circuits involved in threat reactivity, learning, and decision making. Rather than different regions being uniquely engaged in separate processes, review of the evidence suggests that most regions are engaged in more than one function, and that the concepts of learning, memory, and decision making are difficult to isolate behaviorally and computationally. For example, regions typically involved in physiological reactivity, such as the periaqueductal gray (PAG), also contribute to the formation of Pavlovian associations by providing a teaching signal to the amygdala (AMG). Regions typically assigned to associative learning, such as the amygdala, hippocampus (Hip.), and ventral striatum (VS), are also involved in decision making. Regions often assumed to have a role in decision making (valuation and choice), such as the orbitofrontal cortex (OFC), ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC), the posterior parietal cortex (PPC), and the lateral PFC, are also involved in learning. The insula (Ins.) has been implicated in decision making, learning, and reactivity. Together, these regions converge into a global network that conducts a similar set of computations across various phases of experience. The separate investigation of specific types of learning, memory, and decision-making processes is thus not conducive to a comprehensive understanding of a unified global network. A refined approach would first define the computational problem – for example, how the brain predicts outcomes under uncertainty in a volatile environment given certain stimuli or actions – and then examine how each region in this global network contributes to these computations (e.g., informing value, tracking associability, computing prediction error). This process should then be iterated across the various stages of threat experience (e.g., initial encounter, memory retrieval, decision making), as well as other domains such as reward. BNST, bed nucleus of stria terminalis; Hyp., hypothalamus; dmPFC, dorsomedial PFC; dlPFC, dorsolateral PFC; vlPFC, ventrolateral PFC; PCC, posterior cingulate cortex; SN, substantia nigra; VTA, ventral tegmental area.

inspection and anticipation of the predator's moves, strategic freezing to avoid detection and gather information, and avoidance estimation. Finally, the prey is under most extreme danger during the 'circa-strike' phase, when the predator is attacking. In that attack mode, the predator could be distant enough to allow a feeling of fear and rapid thoughts examining the situation and assessing escape routes. Fight or flight ensues as the predator gets closer yet without contact. The final point of contact provokes hard-wired, fast, often poorly executed reactions of freezing and panic [5].

A hierarchical neuroanatomical organization traces the threat imminence continuum. As the predator approaches, brain activity shifts from the prefrontal cortex (PFC) to the midbrain. Two parallel paths support defensive approach and avoidance, originating from PFC areas through cingulate cortex areas, to the hippocampus, amygdala, and hypothalamus, terminating on the midbrain periaqueductal gray (PAG) and dorsal raphe nucleus [3]. Converging neuroanatomical and functional evidence across species, including rodents, non-human primates and humans, supports this organizational scheme, attributing feelings of anxiety and fear and cognitive regulation to higher-order cortical areas, and freezing, escaping, and panic to the amygdala, hypothalamus, and PAG, respectively [3,6–8].

A proximal threat engages rapid, reflexive, and narrowly targeted actions, and therefore has limited computational resources. Decisions during this phase are likely to rely on **model-free**

reconsolidation process may result in memory impairment.
Risk: a type of uncertainty in which likelihoods for potential outcomes are fully known. In the context of learning, this is also known as expected or irreducible uncertainty.
Subjective value: the utility of an option to the decision maker, integrating over all of the option's properties, including potential outcomes and their likelihoods, as subjectively perceived by the decision maker.
Threat imminence continuum: mapping of defensive behaviors into stages of threat, ranging from a 'safe' stage (no threat) to the most extreme 'circa-strike' stage (an attack occurs).
Unexpected uncertainty: a surprising change in the probabilistic structure of the environment.
Volatility: frequency of changes in the probabilistic structure of the environment. The learning rate should be higher in volatile, compared with stable, environments.

computations, sustaining the repetition of previously reinforced actions. The more distal points of encounter allow time to assess the environment and consider alternative courses of action. It is plausible that, in addition to model-free responses, such threats initiate **model-based**

Box 1. The Neural Mechanisms of Threat Detection and Modification

Innate and learned threat processes are distinguishable in the brain. The neural circuits of innate threats can be described by three main functional units: a detection unit subserved by sensory systems that gather sensory information signaling the presence of threat; an integration unit where sensory information converges, directing the recruitment of downstream structures that produce the adaptive response, with the integration occurring at the level of the amygdala and hypothalamus; and an output unit comprising brainstem structures, including the PAG, directly producing adaptive physiological and behavioral responses to the threatening stimuli [7,168]. The experience of innate threat instructs a learning process that forms a memory of the threatening event.

The neural mechanisms of threat acquisition, extinction, and other forms of threat modulation (Figure I), are centered around the routing of information to and from the amygdala and within amygdala nuclei [169,170]. During threat conditioning, sensory inputs arrive at the amygdala through either a thalamo-cortico-amygdala pathway or a direct thalamo-amygdala pathway. Those sensory inputs, signaling the neutral (to be conditioned) stimulus and the aversive outcome, converge onto neurons in the lateral amygdala (LA). The stimulus–outcome convergence induces long-term potentiation of stimulus input synapses, such that when the stimulus later appears alone, its input will sufficiently drive LA outputs, triggering the threat response. Within the amygdala, the LA relays information directly to the central nucleus (CE) or via the basal nucleus. There is also evidence that the basal/lateral nuclei (BLA) and CE process information in parallel and not only serially [171]. The CE is the major output structure of the amygdala. CE projections to the hypothalamus, PAG, and other regions mediate the behavioral and physiological threat response (freezing, change in heart rate and blood pressure, and release of stress hormones) [23,170,172–174].

LA neurons' aversive responses correspond to prediction error encoding, because these cells respond strongly to unexpected aversive outcomes but reduce their firing when the outcomes are predicted by conditioned stimuli. The prediction error signal is created by an amygdala–PAG feedback circuit. The conditioned stimulus recruits the CE to activate a specific population of PAG neurons, which in turn inhibit aversive signaling before it reaches the LA, thereby resetting threat learning levels and controlling conditioned threat behaviors [29,175–177].

Along the borders of the BLA and CeA lie islands of inhibitory (GABAergic) neurons. These are the intercalated cell masses (ITCs) that exert inhibitory control over the amygdala as part of an 'off switch' system. The major amygdala nuclei further divide into internal partitions. For example, the CE divides into lateral and medial parcels. The lateral part of the CE continuously inhibits the medial part, keeping the amygdala's output under control. The BLA overturns this effect by projecting to the lateral CE via ITCs. The medial CE, consequently, now free of inhibition, enables the threat response [178,179]. In addition to the threat responsive population of 'on' cells in the amygdala, there are 'off' cells that are responsive to stimuli that signal extinction, as well as BLA to nucleus accumbens projecting neurons that signal reward [127,180,181].

Amygdala threat responses are short-lived: they last only a few hundred milliseconds and therefore cannot be responsible for the stimulus-evoked sustained threat response, which typically lasts at least a few seconds. In rodents, the dorsal part of the medial PFC, the prelimbic PFC, is the region that maintains and prolongs threat responses. The adjacent infralimbic PFC mediates the diminution of the threat response seen following extinction [182–185]. The putative human homologs of these regions are the dACC and the vmPFC, respectively [186]. Retrieval of extinction memory involves potentiated inhibitory circuits in the BLA and increased medial PFC output to the amygdala [178,180,187]. Inputs from the hippocampus, insula, and thalamus, among other regions, further modulate the amygdala's threat response [182,183,188].

Regions of the medial PFC arbitrate between freezing and avoidance responses. The prelimbic PFC gates the impact of BLA inputs to the ventral striatum during avoidance [49]. BLA projections to the ventral striatum and prelimbic PFC projections to the BLA both facilitate avoidance, whereas prelimbic PFC projections to the ventral striatum diminish avoidance [189]. The infralimbic PFC suppresses freezing mediated by the CeA [190]. The retrieval of avoidance memory relies on prelimbic PFC projections to the BLA, and avoidance extinction relies on projections from the infralimbic PFC to the BLA and ventral striatum [191].

Consistent with these findings, research in humans showed that the degree to which the amygdala and striatum are synchronized with regions in the medial PFC during avoidance learning predicted avoidance success [51]. Striatal activation differentiated between participants that exerted control over conditioned stimuli versus those that did not, and corresponded to diminished return of threat responses [50,52,192]. Theoretical formulations suggest that individuals' estimates of agency based on past experience with controllable and uncontrollable outcomes adaptively calibrate their proactive or reactive behavioral strategies [193].

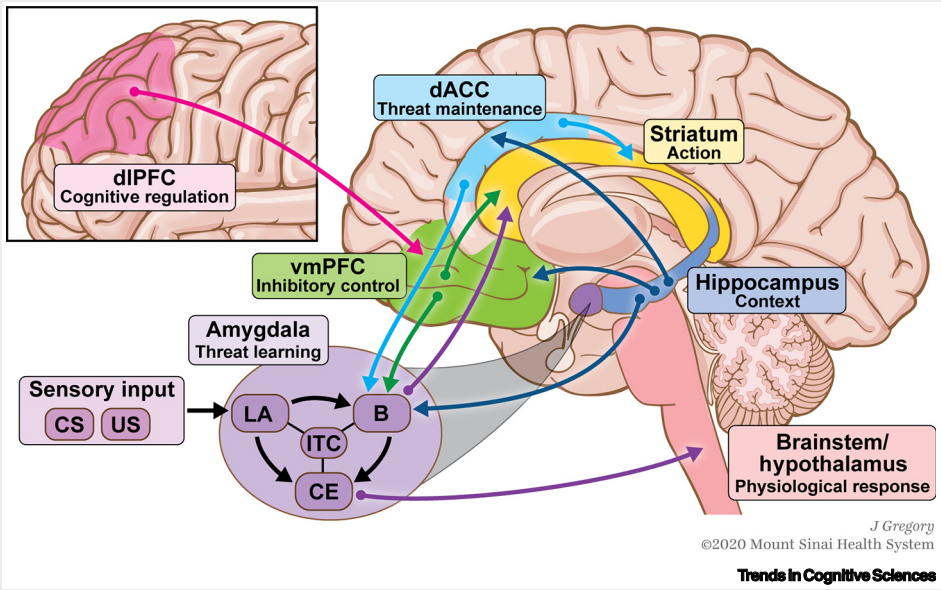

Figure I. The Neurocircuitry of Threat Learning and Modification. A schematic representation of the inputs into the amygdala, information transfer within the amygdala, and amygdala outputs, and their supposed functional role in threat learning and regulation. The dACC and the vmPFC represent the putative human homologs of the rodent PL and IL, respectively. B, basal amygdala; LA, lateral amygdala; CE, central amygdala; dACC, dorsal anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; PL, prelimbic prefrontal cortex; IL, infralimbic prefrontal cortex; CS, conditioned stimulus; US, unconditioned stimulus.

computations – the prospective anticipation of action consequences using a constructed map or a model. This computational policy favors accuracy and strategic planning [9,10] and allows 'offline' testing of potential courses of action using mental simulations. Computational methods such as Dyna [11] aim to identify the optimal policy for a particular situation by simulating actions and their consequences within an internal model of the environment. The mental simulations provide data to train the model and improve predictions in the absence of actual threat experiences.

People who experienced a traumatic event and went on to develop post-traumatic stress, and people with anxiety disorders (Box 2), often exhibit some of the same behaviors along the threat imminence continuum, albeit in an excessive manner and inappropriate contexts. For example, according to the DSM-5 [12], among the criteria for a PTSD diagnosis are hypervigilance and avoidance as in the pre-encounter anticipatory anxiety phase. Anxiety disorders are accompanied by frequent intermittent anxiety involving rumination, worry, and over-strategizing. This suggests that PTSD and anxiety disorders may involve deficient computations of spatiotemporal threat, particularly influencing the reliance on model-based planning during anticipatory and encounter anxiety. Consistent with this idea, high-trait anxiety affects escape decisions from a virtual predator, but only when the threat is distal rather than imminent [13]. Other studies, in the domain of reward learning, demonstrate the vulnerability of model-based computations to stress [14] particularly in depression [15], and to lifetime stress [16], as well as to self-reported intrusive thought [17], a symptom of generalized anxiety and PTSD. Additional research is required to examine how greater reliance on one system versus another relates to vulnerability or resilience to traumatic stress, and which clusters of symptoms correspond to these underlying behavioral policies.

Beyond direct experience, individuals can also be impacted by secondhand, vicarious experience, by witnessing the experience of others or through verbal instructions (Box 2). This form of social

Box 2. PTSD and Anxiety

PTSD was first introduced as a diagnosis in 1980, in the third edition of the American Psychiatric Association (APA)'s *Diagnostic and Statistical Manual of Mental Disorders* (DSM). The initial definition attracted controversy and was revised over the years [194]. The most substantial conceptual change occurred in the latest DSM-5, with the removal of PTSD from the category of anxiety disorders [195]. Instead, the disorder was placed in a new diagnostic category named 'Trauma and Stressor-related Disorders'. This new categorization is unique among psychiatric disorders: while all other DSM diagnostic categories are conceptually grouped by symptom characteristics, this is the only category that requires an exposure to a stressful event as a precondition.

Accordingly, most fundamental to the nosology of PTSD is criterion A: exposure to a traumatic event. The definition of trauma includes actual or threatened death, serious injury, or sexual violence [12]. This specific definition indicates that not all stressful events (e.g., psychological stressors like losing a job or a divorce) qualify as trauma. Exposure to trauma, according to criterion A, comprises not only direct personal exposure, but also witnessing trauma to others or indirectly experiencing trauma through the traumatic experience of a close individual. The assessment of PTSD symptoms is valid only if criterion A is met. The symptoms must begin or worsen following the traumatic event, without assuming any causal or etiological inference. The symptom groups are: intrusions, avoidance, negative alterations in cognition and mood, and alterations in arousal and reactivity. This new organization emphasizes avoidance, now making it a requirement in order to meet the diagnostic criteria for PTSD.

One reason for the separation of PTSD from anxiety is the ample evidence that PTSD involves emotions outside the range of fear and anxiety (e.g., anger, guilt, shame). To be diagnosed with an anxiety disorder, a person must experience fear or anxiety that are out of proportion and impair normal function. Fear refers to an emotional response to an immediate threat, often triggering a fight or flight reaction. Anxiety is more diffused, referring to the anticipation of a future threat, typically manifested in muscle tension and avoidance behaviors.

The neural correlates of PTSD and anxiety largely involve structural and functional aberrations in the amygdala, PFC, and hippocampus. PTSD patients exhibit exaggerated amygdala reactions to negative and trauma-related stimuli, hypoactivation of the vmPFC, and impaired hippocampus-dependent context learning, as well as neuroendocrine dysregulation [196]. Animal models differentiate neural circuits underlying the response to immediate present threat versus uncertain threats (i.e., anxiety). Uncertain threats (e.g., unpredictable shocks) engage the bed nucleus of stria terminalis (BNST), which mediates the transfer of information between the amygdala and the ventral striatum and modulates defensive reactions [197].

The first line of defense against PTSD is prolonged exposure therapy [198,199] – the repeated exposure to trauma-related cues – leading to desensitization (akin to laboratory procedures of extinction). The treatment is ineffective in about 20–30% of patients, and approximately 20% fail to complete the full course of treatment. There is little empirical evidence to support pharmacological treatments for PTSD [200,201], which are typically selective serotonin reuptake inhibitors. The development of novel effective treatments is desperately needed. Possible directions could consider drug-assisted behavioral therapy [202], reconsolidation-based pharmacological and behavioral treatments [203–205], a synergistic approach combining multiple behavioral and neural processes [75,86] and the temporal progression of treatment [83].

behavior capitalizes on existing neural mechanisms of direct learning, in addition to processing of social information, and is evident across species [18–20]. Socially formed associations can also shape subsequent decision making the same way that direct learning does [21].

Experiencing imminent threat may result in a cascade of: (i) the formation of threat associations; (ii) post-association learning; (iii) storing and updating of these associations; and (iv) decision making under threat (Figure 1). It is possible that the particular processes occurring in the first stage would impact the four following stages. For example, a less-imminent threat encounter might result in a weaker memory that is easier to extinguish; a highly stressful encounter with more imminent threat could produce a memory that is **reconsolidation** resistant; and the degree of uncertainty during initial encounters might shape the flexible update of initial learning, etc. The next sections examine the cascade of processes following the initial threat learning.

Backpropagation of Threat

The computational processes during imminent threat focus on the immediate needs and current environment of prey. However, they also rely on prior learning, such as previously acquired

associations, reinforced actions, and learned models, called upon in the service of the moment. The merging of past and present experience yields new associations and updated models. In the simplest form, predation becomes associated with neutral stimuli in the environment. The threat of the predator backpropagates to the stimuli that predict it, a process known as **Pavlovian conditioning** [22,23].

Various learning models are used in this research, aiming to capture various types of information. Theories of associative learning, such as the Rescorla–Wagner (RW) and reinforcement learning models, envision that learning is driven by surprise, formalized as prediction error – the difference between the outcome expected and the outcome received. The **predictive value** of the stimulus – the level of threat that it predicts – changes proportionally with the magnitude of the error at a rate that is not in itself influenced by the learning [9,24]. Overgeneralization of the learning to cues that were not associated with threat [25] may be a hallmark of anxiety disorders [26].

The temporal difference model extends the RW model, allowing predictions of accumulated discounted future outcomes rather than the immediate only [9]. Other theories, such as the Pearce–Hall model, focus on the predictive efficiency of the cue. Here, to learn cue–reinforcer associations, individuals track a quantity termed **associability**, which reflects the degree to which a cue has previously been accompanied by surprise (positive or negative prediction error). The associability of a cue gates the amount of future learning about that cue, depending on whether it has previously been a poor or a reliable predictor of an outcome. In this way, associability accelerates learning to cues whose predictions are poor and decelerates it when predictions become reliable [27].

Several brain regions play a role in associative learning (Figure 2 and Box 1). There is evidence for the encoding of aversive prediction errors in activation patterns in the amygdala [28–31] and the striatum [32–34]. The striatal activation may result from dopaminergic inputs from the ventral tegmental area (VTA) [35,36], although the evidence for a striatal role in aversive prediction error may not be as strong as the evidence for its role in reward prediction error [37,38]. Both amygdala and striatum have been implicated in tracking of associability [33,34,39,40].

An augmented 'hybrid' RW model controls learning rates dynamically, based on the Pearce–Hall learning rule. In humans undergoing threat conditioning, measured by skin conductance response (a measure of autonomic nervous system activity), the hybrid model better captured cue-specific associabilities, over and above value expectations [40–42]. A study in combat veterans used the hybrid model to estimate the influence of prediction errors on cue-specific associabilities in each learning trial [34]. This subject-specific parameter positively corresponded to PTSD symptoms [measured by the Clinician Administered PTSD Scale (CAPS)]. Thus, by assigning more weight to prediction errors, the more trauma-affected individuals exaggerated their adjustment to cues that did not predict what they had expected [34,43]. Amygdala and striatal tracking of cue value throughout learning negatively corresponded to CAPS (less-faithful neural representation of value related to worse diagnosis). Better striatum tracking of associability corresponded with lower symptom severity, and partially mediated the positive relationship between prediction error weight and CAPS [34].

In parallel with the passive formation of cue–outcome Pavlovian associations, individuals can readily associate their actions with outcomes, a process termed instrumental (or operant) conditioning [44–46]. A classic example of **instrumental conditioning** in the context of threat is **avoidance learning** (Box 1), where an animal learns to prevent or minimize contact with an aversive outcome (e.g., electric shocks) or the stimuli that predict it [47]. Animal studies identify

at least two opposing pathways subserving active avoidance learning: a lateral amygdala – basal amygdala – nucleus accumbens pathway required for active avoidance; and a competing lateral amygdala – central amygdala – PAG pathway mediating freezing to conditioned stimuli. The infralimbic and prelimbic PFC subregions serve as the arbitrators, mediating the transition from reaction to action by suppressing freezing and facilitating avoidance [48,49] (Box 1). Human studies are consistent with these findings, demonstrating the involvement of amygdala and striatum and their interactions with the medial PFC [50–52].

Avoidance by itself is an adaptive response to danger, but unwarranted and excessive avoidance is a hallmark of PTSD and anxiety disorders [48,53]. Avoidance symptoms can be subdivided into passive and active. Passive avoidance is the lack of action (e.g., strategic freezing) whereas active avoidance involves emitting an action that circumvents the aversive outcome. The fact that active avoidance is a form of learning driven by the absence of the aversive reinforcer poses a challenge for learning theories. The psychologists O.H. Mowrer and Neal Miller provided a conceptual framework for active avoidance as a two-factor learning process, where threat is first acquired through Pavlovian conditioning, and then actions that reduce the conditioned threat are reinforced through instrumental conditioning [47,54,55]. Reinforcers of avoidance behavior include negative reinforcement by the removal of threat-associated cues and positive reinforcement by cues associated with safety [48,49].

In the context of both Pavlovian and instrumental conditioning, the models we have surveyed explain predictions of expected threat, but ignore uncertainty about these predictions (Box 3). Instead of point estimates, Bayesian learning models include estimates of prediction uncertainty that rule a dynamic learning rate [42,56–58]. In stable environments, experiences from the distant past are informative in predicting the future, and transient changes should be largely ignored. Natural environments, however, are seldom stable; rather, **unexpected uncertainty** (Box 3) may arise, when the probabilistic structure of the environment changes abruptly [59–61]. When action–outcome contingencies change (e.g., the outcome probability substantially drops), only

Box 3. Types of Uncertainty

Learning and decisions about threats involve multiple forms of uncertainty. The learning literature distinguishes between expected and unexpected uncertainty [206].

Expected uncertainty – also known as irreducible uncertainty [60,207] – arises from the probabilistic nature of outcomes in a familiar environment. In the laboratory, expected uncertainty occurs when outcome probabilities are fixed and well learned (e.g., when each cue presentation is associated with an 80% chance of an electric shock). While the outcomes themselves are uncertain (shock or no shock), their variance is expected, and, assuming a stable environment, these outcomes do not provide useful information and should not drive learning.

Before learning is well established, the lack of experience gives rise to estimation uncertainty [60,208] – uncertainty about the probabilistic structure of the environment. This type of uncertainty can be reduced with additional evidence, and signals how much learning is required. In the economics literature, irreducible uncertainty and estimation uncertainty are termed 'risk' and 'ambiguity', respectively (Box 4).

Unexpected uncertainty arises when expectations about the statistical structure of the environment are violated (e.g., if the shock probability abruptly drops to 20%). Determining that the uncertainty is unexpected, rather than part of the stochastic nature of the stable environment, is challenging [59]. Expected uncertainty can provide a baseline for the level of uncertainty against which surprising events should be compared [61], considering the learner's belief about the world [209]. Changes in the probabilistic structure of the environment – for example, in a volatile environment – should lead to an increased learning rate, such that predictions rely more on recent events. Normative statistical models, such as Bayesian models, and their approximations, provide accurate predictions for behavior [56,60,62,65] but include complex computations, and it is unclear how these computations are implemented by the brain [61]. One model, synaptic metaplasticity (the ability to change synaptic states without measurable changes in synaptic efficacy), was proposed as a biologically plausible mechanism for adjustment of learning rates based on unexpected uncertainty [210].

recent experiences should inform learning, to allow quick adaptation to the changing conditions. Human participants are able to incorporate estimates of unexpected uncertainty in their learning [62–65]. Individuals adjust their learning rate in response to variations in **volatility** – the frequency of changes in action–outcome contingencies (or the mean level of unexpected uncertainty [65]) – separately for potential rewards and punishments [66]. Pupil dilation and BOLD signals in the locus coeruleus, which are indicators of arousal, reflect subjective estimates of unexpected uncertainty [63,67]. Activity in the anterior cingulate cortex tracks subjective estimates of volatility and reflects individual differences in learning rate [65]. Using a hierarchical Bayesian learning model, estimates of uncertainty during a probabilistic choice task predicted subjective stress and arousal, exemplifying a tight link between stress responses and environmental uncertainty [68]. These studies suggest that conditioned threat responses do not correspond simply to the outcome prediction, but rather to the degree of uncertainty surrounding that prediction.

Individuals with trait anxiety learn more from recent punishments than healthy controls [69] but are slow to adapt their learning rates in response to changes in threat volatility, and show reduced pupil response to volatility [70]. Social contexts may exacerbate such reduced adaptability [71]. Inappropriate adjustment to changes in probability structure may also lead to poor decision making, and contribute to increased symptoms if aversive outcomes are perceived as less predictable and less avoidable [70].

Overall, extant evidence suggests that aberrant neural computations of value, prediction error, associability, and estimations of uncertainty are related to anxiety and PTSD. The models described above capture two important facets of associative learning: reinforcement learning formalizes predictions of long-term accumulated outcomes; and Bayesian models track uncertainty around learned associations. Models that merge the two computations have also been proposed [57].

Flexible Threat Associations

Following the encounter with danger, an individual will emit defensive responses triggered by the conditioned stimuli, but not for long. Eventually, for adaptive energy maintenance, those defensive responses will dissipate and new learning will take their place. The inappropriate lingering of learned defensive responses is part of a major PTSD symptom cluster in the DSM-5 [12], defined as alterations in arousal and reactivity (e.g., hypervigilance) that began or worsened after the trauma. A prime mechanism that counteracts threat conditioning is **extinction learning** – the decline in responding to a stimulus that previously signaled danger, following repeated nonconsequential exposures [72]. Extinction can also be learned vicariously [73] or through imagination of the conditioned cues [74] by capitalizing on neural mechanisms of direct learning.

Formalizing an associative learning model to describe extinction has been challenging. The RW model, for example, views extinction as the unlearning of associative contingencies due to the omission of the outcome, but this fails to account for the return of extinguished responses under various circumstances (e.g., spontaneous recovery, renewal, reinstatement). The Pearce–Hall model, alternatively, classifies extinction as new learning where omission of the outcome in the presence of the conditioned stimulus creates a second association, such that threat and extinction associations compete for expression. However, it is possible that a mixed model, assuming cooperation between unlearning and new learning, best describes an individual's internal representation [75]. The latent cause model [76] captures that cooperation: individuals update the associative weight between the stimulus and the outcome (i.e., unlearning) given small prediction errors, but infer that a second rule is likely to be in effect (i.e., new learning) given large prediction errors. In this way, an individual does not simply learn a single stimulus–outcome

association, but rather parses experiences into latent causes, each with its own associative weight, thus constructing a structure of the environment. During extinction, unlearning would dominate over new learning by minimizing deviations from the individual's expectations, such as transitioning from conditioning to extinction by gradually changing association strengths over time [77]. The model predicts that individuals who assume a single latent cause, and update only the initial threat memory, will show weaker recovery of the extinguished response. Individuals who form a new extinction memory, segmenting their experience into two latent causes, are more likely to show greater recovery from extinction [78].

How the brain implements structure learning is unclear. A possible circuit mechanism for structure learning postulates that latent cause representations, possibly originating from the orbitofrontal cortex (OFC), activate hippocampal dentate gyrus cells, reflecting the likelihood of the active cause. The VTA computes the discrepancy (i.e., the prediction error) between the associative weights that the inferred cause predicts and the observed contingencies. The prediction error signal is then transmitted via dopaminergic projections to the amygdala, hippocampus, striatum, and PFC, thereby triggering memory updating or new learning; small prediction errors alter the associative weight of the inferred cause, whereas large prediction errors trigger neurogenesis of dentate gyrus granule cells to generate a new latent cause [79,80].

Beyond extinction, the power of the latent cause and other unifying models lies in their ability to explain a multitude of post-conditioning phenomena, such as generalization of the response to other stimuli, blocking learning to other stimuli, second-order conditioning, **counterconditioning**, and more [57,79,81,82], thereby generalizing the RW model and addressing its explanatory limitations. Further model development is required to capture the effect of multiple extinction sessions, which may diverge into a separate circuit mechanism [83], and capture clinical therapeutic processes more faithfully [84]. Another consideration is the lingering impact of PTSD, which can last years and even decades. The longevity of the disorder is difficult to explain using current learning models, each capturing a sliver of the symptoms. A synergistic model describing how multiple learning phenomena work together may explain the persistence of defensive motivational states [85,86].

Life Cycle of Threat Memory

Aversive events serve as teaching signals instructing synaptic plasticity in the amygdala, resulting in threat memory storage (Box 1). Interventions such as amnesic agents (e.g., protein synthesis inhibitors) or electroconvulsive shock stimulation interrupt **consolidation** – the stabilization period following the acquisition of a new memory. Once a memory has formed, however, it goes into a life cycle of oscillations between periods of neural stability and instability (Figure 3). Memory **reactivation** and **destabilization** trigger those periods of instability, providing opportunities for modification. The same interventions that disrupt consolidation also interrupt – and therefore support the existence of – a destabilization period following the reactivation of old memories. This phenomenological similarity inspired the term reconsolidation, referring to the active process necessary to restabilize a memory after it has been reactivated and destabilized [87,88].

The instability period, when old memories resurface, presumably serves the purpose of incorporating new relevant information into the memory, instead of forming a new separate memory of present events [89,90]. Experimental evidence supporting this theory introduced extinction learning following the reactivation of a conditioned threat stimulus, leading to long-term reduction in conditioned threat responses [91–95]. Another form of intervention used counterconditioning, which was demonstrated in the context of appetitive conditioning. Here, a drug-related conditioned stimulus was repeatedly paired with an aversive outcome (e.g., a disgusting image) following reactivation

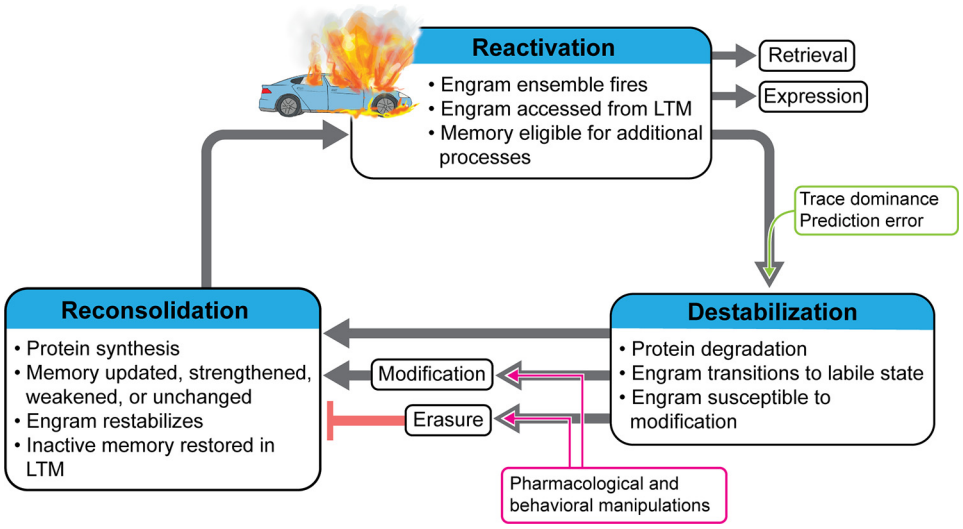

Figure 3. Life Cycle of Threat Memory. Memory encoding is the strengthening of neural connections through long-term synaptic restructuring, a process occurring when different regions coactivate. A memory trace, or engram, is therefore not a physical entity like a stored object, but rather the disposition of neural circuits to fire upon triggering by a certain reminder. During reactivation, a memory becomes active when the engram ensemble fires. Reactivated memory is eligible for retrieval, behavioral expression, and/or destabilization. This means that memories that are retrieved and expressed are not necessarily destabilized, and that destabilization can even occur covertly without behavioral expression. Destabilization involves a cascade of cellular and molecular processes (e.g., protein degradation) that instigate the transition of the engram from a stable to an unstable state. At this point, the memory becomes susceptible to modification. Among the computational principles that govern destabilization is trace dominance: a memory will be more malleable to amnestic manipulations to the extent that it has control over behavior at the time of treatment. Prediction error is one of the parameters that influence trace dominance and facilitate destabilization. Memory restabilization (reconsolidation) requires protein synthesis, among other cellular/molecular events. Successful reconsolidation restabilizes the ensemble and restores the memory into its inactive state. Manipulations that interrupt reconsolidation (pharmacological agents, behavioral interference, memory enhancers) will reroute the memory toward erasure, strengthening, weakening, or updating. A memory that has not been erased may cycle into another sequence of reactivation, destabilization, and reconsolidation. LTM, long-term memory.

[96,97]. Other types of noninvasive interference could be effective, including those that deplete the neural resources needed for reconsolidation, such as extensive sensory motor tasks [98] (Figure 1).

How does new learning following reactivation differ from a standard associative learning session? Why would reactivation–extinction, for example, lead to a more permanent reduction of threat responses while standard extinction allows their return? The latent cause model, described in the preceding text, offers a theoretical solution by formalizing the dynamic interplay between learning and memory [79]. Applying the latent cause theory to the case of post-retrieval memory modification stipulates that when a memory is retrieved, the brain assumes that the previously inferred cause (originally assigned to the remembered event) is once again active. This inference makes the memory eligible for updating because new information, now attributed to the original cause, merges and thus changes the original memory. For example, the associative weights of a cue–outcome association will permanently decrease due to the merging of extinction learning; in other words, the memory has been updated. If the brain otherwise infers a new cause for the surprising event, a new memory will be formed.

One of the parameters that nudges latent cause assignment, as featured in the model, is the duration of the reminder cue. Reminder duration can be conceptualized as the length of exposure or multiple repeated exposures with short gaps [79]. A brief exposure to the reminder cue favors

the assignment of the reminder to the initial threat learning cause. With longer reminder durations, prediction errors accrue, facilitating the inference of a new latent cause [99–104]. The model explains many of the boundary conditions of reconsolidation updating that have been observed in laboratory experiments, and makes testable predictions about when a memory will or will not maintain its original form [79].

One of the most important goals of clinicians treating anxiety and PTSD is to facilitate a change that is enduring in patients. Achieving enduring change that is not easily prone to relapse conceivably requires three essential components: reactivating the problematic memories along with the emotions they elicit; altering those memories by having a corrective emotional experience during reconsolidation; and building enduring semantic structures onto the updated memories, by implementing new behaviors and ways of engagement with the world [105,106]. Some forms of therapy, such as coherence therapy, are built on the principles of memory reconsolidation and are designed to maximally optimize this process [107–109].

Decision Making under Threat

When a threat is detected or remembered, a rapid decision-making process ensues, resulting in an approach (e.g., attack) or avoid reaction. Long after the imminent threat has dissipated, decisions about cues that predict threat are likely to share some of the mechanisms with those initial processes. These decisions may be adaptive (e.g., approaching a threat that could be overcome, avoiding a real threat or a cue that predicts it) or maladaptive (e.g., avoiding a threat that could be overcome or a benign cue).

The decisions that individuals make depend on their available courses of action, the potential outcomes of each action, and the likelihood of each outcome. Common models of choice posit that decision makers integrate the various properties of each option to compute its idiosyncratic **subjective value**, and then choose the most valuable option [110]. Although this may seem straightforward, we do not fully understand how these computations are implemented in the brain. We do know that activation patterns in a network of multiple brain areas reflect subjective values that are inferred from behavior (Figure 2). Activity in two areas in particular, the ventromedial PFC (vmPFC) and the ventral striatum, is consistent with the encoding of subjective values across different categories and under varying conditions [111,112]. Other putative value regions include more dorsal regions of the medial PFC [111,113], the OFC [114,115], the posterior cingulate cortex (PCC) [116], and the posterior parietal cortex (PPC) [117].

While it is likely that activity in these areas encodes the value of potential rewards, the evidence for value encoding of punishments, or threats, is less robust. Some neuroimaging studies in humans report overlapping representations of positive and negative values [117–119]. Other studies describe distinct representations of rewards and punishments in different brain areas, with more medial representations of reward value and more lateral representations of the value of punishments [120–122]. Single-unit studies in animals also report both distinct [123] and overlapping [124] representations of rewards and punishments in the dopaminergic midbrain, habenula, and medial PFC. In addition, some brain areas, including the vmPFC, OFC, and PPC, encode aversive values in a monotonic manner (with decreasing activation for more-aversive values), whereas other areas encode value in a u-shaped manner, with high activation for both high rewards and high punishments, consistent with salience representation. The salience network includes the ventral striatum, dorsal anterior cingulate cortex (dACC), anterior insula, temporoparietal junction [111,114,117,125], and amygdala [126,127]. Thus, the ventral striatum is the only region exhibiting both monotonic and u-shaped representations of value,

indicating a dual role for this structure in encoding value and salience [36,111,114,125]. Whether the brain first signals the salience of the stimulus (how important it is) to orient attention and then determines its valence (positive or negative), or whether value is computed first, with salience information (its absolute value) extracted later or in parallel, is an open question.

The likelihoods of potential outcomes are an important factor affecting the subjective value of available options. For example, a 10% chance of sustaining an injury is not as bad as a 50% chance of sustaining the same injury. Only seldom, however, are these likelihoods completely known – a type of uncertainty known as **risk** or 'irreducible uncertainty'. In most cases, likelihoods cannot be precisely estimated; rather, there is some **ambiguity** or 'estimation uncertainty' around those likelihoods (Box 3). Repeated sampling of the environment and experiencing various outcomes can reduce ambiguity. Subjective values are influenced not just by the objective levels of risk and ambiguity around potential outcomes, but also by individuals' subjective perception of risk and ambiguity and their attitudes toward these sources of uncertainty. Since risk and ambiguity attitudes are only weakly correlated across individuals, they are likely to involve somewhat separate cognitive mechanisms (Box 4).

Box 4. Attitudes toward Risk and Ambiguity

Since most decisions are made under partly ambiguous conditions – where likelihood estimates exist but are not precise – individuals' behavior under uncertainty is modulated both by their risk attitude and by their ambiguity attitude [211]. To estimate these attitudes in the laboratory, researchers use simple tasks in which participants are required to make a series of choices between various uncertain and certain options. To elicit risk preferences, the choices are between options whose outcomes and outcome probabilities are fully known (e.g., 50% chance to win $10), where one option is 'riskier' and the other is 'safer' [212]. A risky option is one that offers a small probability for a high reward; a safe option would offer a smaller reward, but at a higher probability. For example, a 50% chance to win $10 offers $5 on average, but at a higher risk than a sure bet of $5. A risk-averse individual would prefer options that offer smaller amounts at higher probabilities over ones that offer higher amounts at lower probabilities, even if the expected value (the product of the probability and amount) of the latter is higher. A risk-seeking individual would show an opposite preference.

Most people tend to be risk averse when making choices between moderate potential monetary gains [213,214]. There is high variability across individuals, however, in the degree of risk aversion, and a minority of people exhibit risk neutrality (i.e., they choose based on the expected value alone) or even risk seeking [215]. This is important, because it means that individual risk attitudes may be associated with particular personality traits or psychopathological symptoms. The picture is a bit more complex, however, because preferences change when the choice is between losses rather than gains. Here, too, there is wide variability in individual preferences, but most people tend to exhibit risk seeking rather than risk aversion. For example, in a choice between losing $8 for sure and taking a 50% chance of losing $20, many are likely to take the risk [213]. Importantly, individuals' attitude toward risk in the gain domain does not predict their attitude to risk in the domain of losses (Figure I) [214,215], suggesting that these are two separate characteristics that may be differentially associated with personality traits and clinical symptoms.

To elicit ambiguity attitudes, participants are also asked to make choices when some (or all) of the information about outcome probabilities is withheld. Ambiguity aversion is commonly observed when the choice is between potential gains [216–218]. In the domain of losses, there is some evidence for reduced ambiguity aversion, or even no effect of ambiguity [215,219]. Importantly, risk and ambiguity attitudes are only weakly, if at all, correlated across individuals (Figure I) [128,137,145,215,220,221], suggesting that they rely on cognitive mechanisms that are at least partly separable. The individual's behavior when making choices between potential gains also does not predict their choices between potential losses. This means that there is likely to be no single unified trait of 'uncertainty attitude'. Rather, it seems that how the individual copes with uncertainty is a complex process, affected by several attitudes, which are largely independent.

A question of ongoing research is to what extent risk and ambiguity attitudes elicited in the laboratory reflect behavior outside of the laboratory [222]. While there is some evidence for consistent risk attitudes across domains [129,223], task specifics may affect estimates of these attitudes [224,225] (although this may arise from changes in the perceived risk rather than the attitude toward risk [226]). It is also unclear to what extent risk and ambiguity attitudes are stable across time (i.e., reflect a personality trait) and to what extent they reflect state variables, although it is possible that risk attitudes are more stable and ambiguity attitudes more transient [221].

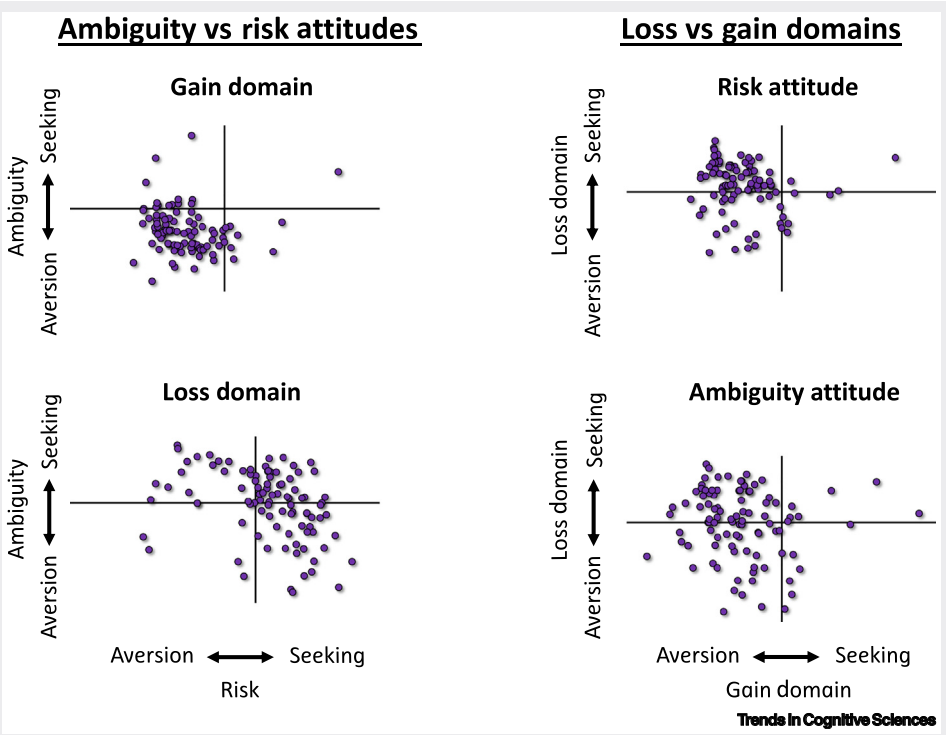

Figure I. Risk and Ambiguity Attitudes in the Gain and Loss Domains Comprise Four Largely Independent Decision Characteristics. Data reanalyzed from [215]. *n* = 99 participants from the general population between the ages of 12 and 77 years.

How potential outcomes and their likelihood estimates are integrated in the brain is not completely clear. The integrated subjective value, reflecting an individual's uncertainty attitudes, is encoded, as expected, in the valuation network [113,128,129]. There is also ample evidence that uncertainty is reflected by neural activity in several brain areas, including the ventral striatum [130–135], PPC [136–138], the anterior insula [131,135,136,139,140], the lateral OFC and ventrolateral PFC (vlPFC) [131,133,137,140,141], and the ACC [142]. Some of these studies highlight potential differences between neural activation patterns encoding risk and ambiguity [134,137,141]. Activity in the PPC [137,138,141], as well as the PPC's structure [143,144], reflects individual risk attitudes. In some studies, activity in the vlPFC reflects ambiguity attitudes and not risk attitudes [134,137,141], but other studies report correlation with risk attitudes [133,142], and a study in lesion patients also implicates the vlPFC in processing both risk and ambiguity [145]. Finally, the structural and functional connectivity of the amygdala is associated with individual risk attitudes [146]. Taking these findings together, while there is abundant evidence for multiple representations of uncertainty in the brain, the underlying neural substrates of those representations, and whether different types of uncertainty are represented separately, remains unclear.

In addition to the integration of outcomes with their likelihoods, the decision maker also needs to integrate the values of potential rewards and punishments. For example, using the car (Figure 1) will be fast and convenient, but will also bring back memories of the aversive event. The weight given to potential gains compared with potential losses (akin to threats) – or the degree of **loss aversion** [147] – varies substantially across individuals, and serves as another source for individual differences in decision making. In humans, subjective value representations in the vmPFC and

ventral striatum integrate over potential appetitive and aversive outcomes of available options [118,148] and reflect the individual's degree of loss aversion, as estimated from their choice behavior [118]. Research in animals also implicates the amygdala in value integration. While the activity of amygdala neurons reflects the value of both potential rewards and potential punishments [149], lesion studies in rats suggest that the basolateral amygdala (BLA) has a specific role in integrating rewards with punishments. BLA lesions led to increased choices of large rewards accompanied by potential punishments, but did not impair sensitivity to the potential punishments [150]. Optogenetic inhibition of the BLA during the simultaneous receipt of reward and punishment also increased risk taking but inhibition during the deliberation phase had an opposite effect [151], suggesting a heterogeneous role for the amygdala in value integration, which may rely on the orchestrated activity of separate neuronal populations. Interestingly, unpredictable outcomes – even in the absence of interaction with motivational value – lead to sustained activity in the amygdala, in both humans and mice [152].

Thus, decision making as a whole integrates outcomes and their likelihoods, expected rewards and punishments and their weights (loss aversion), various sources of uncertainty (irreducible or risk, and reducible or ambiguity), and the individual's attitudes toward uncertainty (risk and ambiguity aversion). The decision maker then needs to compare the integrated values of the different options to reach a choice; how this is done is a subject of ongoing research. It is possible that integrated values are compared downstream of value computations [153] or that the choice process is an inherent part of iterative value computations [154,155].

Following a traumatic event, changes in any of the computations of valuation and decision making may be observed [132]. In rodents, there is some evidence for malleability of valence encoding. During acute stress, the presentation of rewards induces punishment-like responses in the lateral habenula (i.e., a switch from the typical decreased firing rate to an increased firing rate), consistent with a shift from value to salience encoding [156]. A different type of response shift occurs in the nucleus accumbens, where neurons switch their preference from rewards to punishments in a stressful environment [157].

In combat veterans with PTSD, the ventral striatum appears to shift from value to salience encoding [158]. These individuals also exhibit enhanced aversion to ambiguity around potential losses in a simple monetary choice task compared with combat veterans who did not develop PTSD [159]. In a trauma-exposed community sample, self-reported intolerance of uncertainty correlated with the severity of PTSD symptoms [160]. The evidence from animal research suggests causality, i.e., that at least some of the changes in computations of valuation and decision making in humans may result from the trauma. However, it is also possible that some of the observed differences reflect predisposition for the development of PTSD, which preceded the traumatic events. Abnormally enhanced intolerance of uncertainty may constitute a transdiagnostic factor across anxiety disorders [161], given that anxiety-related pathologies, including obsessive compulsive disorder (OCD) [162], generalized anxiety disorder (GAD) [163], and social anxiety [164] also evince increased intolerance of uncertainty.

The intolerance of uncertainty observed when anxious individuals and those with trauma-related psychopathology make choices is likely to affect how they learn. An intriguing question is to what extent variations in uncertainty attitudes shape learning about threats. For example, in Pavlovian and instrumental conditioning, outcomes are used to reduce ambiguity about the probabilistic structure of the environment. Ambiguity aversion may hinder learning, because the individual will make an effort to avoid the ambiguous situation; alternatively, it may strengthen acquisition, if the individual will be increasingly motivated to reduce the level of ambiguity. Similarly, extinction,

generalization, relearning, and other post-association processes all require the individual to cope with varying levels of uncertainty. Whether the individual's approach to uncertainty in the context of decision making also underlies their handling of uncertainty in the course of learning is an open question.

Concluding Remarks

A common method of investigation segments threat experience into separate stages and processes, studying and diagnosing them independently. In reality, however, learning, memory, and decision making never occur in isolation and are constantly intertwined. Formalizing first principles may capture the basic computational processes that are a common thread, while considering the stages of threat experience as the settings for these computations.

As described in the previous sections, a fine-grained look at the initial encounter with imminent threat reveals a series of computations involving predictions about the outcomes of environmental cues and actions in an uncertain environment. New cue–outcome and action–outcome associations – arising via accumulated trial and error or models of the environment – are stored in memory, guiding future behaviors. Newly formed associations will go on to compete with, or facilitate, other associations, which together will adjust the individual's behavior to a changing environment. The post-association phase further involves additional computations that segment the stream of experience into distinct clusters, or hidden causes. This type of computation plays a major role in the retrieval of learned associations, as it determines whether a certain memory will be updated or remain unchanged, arbitrating between new learning and unlearning. Finally, choices based on learned associations incorporate estimations of uncertainty to predict and optimize future outcomes.

PTSD and anxiety disorders may be linked to aberrant computations at any stage of threat processing [165]. Improper computations may be specific to threat or extend to other domains, such as reward processing or social interactions. Improper computations may also be limited to a particular processing stage (e.g., learning) or shared by several stages (e.g., learning and decision making). For example, the increased adjustment to surprising cues in PTSD may be specific to threat learning or reflect a general learning deficiency, which is also at play when learning about rewards. The intertwined nature of the various cognitive processes, the overlap in their neural circuits, and the likely contribution of interactions between these processes to mental disorders also reinforce the significance of studying whole-brain computations and connectivity patterns [166].

To decipher the neural mechanisms of disorders such as PTSD, and to translate insights into the clinic (Box 5), comprehensive approaches across domains and processes should be used. Throughout the stages of threat experience, an individual confronts two core computational problems: making predictions about long-term accumulated outcomes and tracking uncertainty in the environment. These computations are first used in the initial assessment of the threat, next in the use of cues and actions to confront it and to appropriately update memories in the aftermath, and finally in the decisions shaped by this experience. To mimic this in the laboratory, we could identify and track a specific set of computations (e.g., computation of estimation uncertainty, or ambiguity) and examine it during tasks of learning, memory, and decision making, under threat as well as reward experience. This approach could identify whether the dysfunction lies in the computation itself and is domain general or in a specific domain, possibly due to disruptions in the incoming and processing of a specific type of information. Such approaches, in conjunction with longitudinal designs [167], are likely to generate new insights about threat computations in the healthy brain, and about disruptions in these computations in disease (see Outstanding Questions).

Outstanding Questions

What explains the lingering effects of PTSD? Current learning models cannot explain the longevity of the disorder. A comprehensive model encompassing multiple learning, memory, and decision-making features may be needed.

To what extent are model-based, versus model-free, computations involved in the development and maintenance of post-trauma and anxiety symptomatology?

How are value computations in the brain used to generate choice? Subjective values of threats, as well as rewards, are encoded in a network of brain areas, but whether and how these values are compared to produce choice is unclear.

What is the relationship between neural representations of value and salience? Are they computed in parallel or is one estimated first and then used to calculate the other?

Are the neural mechanisms that encode uncertainty shared across learning, memory, and decision-making processes as well as across both threats and rewards?

To what extent do aberrant computations confer variability in traumatic stress and anxiety and to what extent do they result from stressful experiences?

Box 5. From Algorithms to Feelings

In this review, we have approached threat experience as a computational process involving sensory and internal inputs computed into a behavioral output. Where should we place the subjective feeling of fear within this framework? The felt quality of emotion, the feeling of threat, is a conscious mental state. This creates a gap that is difficult to fill between what we can measure and the conscious experience [227]. A class of theories approaches consciousness by differentiating first-order representations – mental representations of our situation or states in the world (e.g., visual perception, threat) – which could be either cortical or subcortical representations, versus higher-order representations, which are representations of other representations, often attributed to the function of cortical regions [228]. From this perceptive, our review of the evidence suggests that threat computations belong to the category of first-order representations, involving not only a subcortical defensive circuitry but also computations of value, state, uncertainty, volatility, and more, occurring non-consciously and engaging multiple frontal and parietal cortical regions. By contrast, the feeling of fear belongs to higher-order representations, engaging other circuits and integrating several processes, including non-conscious memories, pre-existing schemas, and mental models [229,230]. Considering fear as a representation or output of a more basic computational operation inspired calls to revisit our terminology to reflect the subject of investigation: using 'threat' to describe the inner working of defensive survival circuits and 'fear' as a category of conscious experience [231].

Conscious, verbal descriptions of emotions in terms of their valence and arousal resonate with non-conscious computations of utility and vigor [196,232,233], but there are no specific computations that map onto the conscious experience of fear. Nevertheless, conscious experience may influence decisions and shape the output of neural computations. For example, informing individuals about aversive contingencies influenced activation in the striatum and OFC to feedback-driven learning. Amygdala responses resisted verbal warnings and changed only when individuals had direct experience of the relationships between cues and outcomes [234]. Remembered feelings could influence the arbitration of choices by creating an anticipation about how one might feel upon incurring an outcome [235,236]. Moods can bias the perceptions of outcomes and sway decisions. For example, a negative outcome would be perceived as worse when one is in a bad mood. Moods could be formalized as the cumulative impact of differences between actual and expected outcomes (i.e., a running average of recent prediction errors). By reflecting the momentum of an outcome in a particular environment, mood helps an individual to account for the statistics of the environment. This could be beneficial for steering away from an environment that induces bad mood, for example, since the magnitude and frequency of negative surprise increasingly grow [237].

Understanding the distinction between the feeling of fear and the neural computations of threat can inform our approach to PTSD and anxiety disorders. Identifying from which level symptoms arise can advise pharmacological and behavioral treatment. The interactions between feelings and emotions could refine the assessment of aberrant behaviors as well as facilitate treatment [197].

Acknowledgments

This work was supported by the US National Institute of Mental Health grants R01MH105535 (to D.S.) and R01MH118215 (to I.L.) and NSF grant BCS-1829439 (to I.L.). The authors thank Temidayo Orederu and Hyojung Seo for helpful comments on previous versions of the manuscript.

References

1. Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman
2. 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 *Evolution and Learning* (Bolles, R.C. and Beecher, M.D., eds), pp. 185–212, Lawrence Erlbaum Associates
3. Mobbs, D. *et al.* (2020) Space, time, and fear: survival computations along defensive circuits. *Trends Cogn. Sci.* 24, 228–241
4. Mobbs, D. *et al.* (2015) The ecology of human fear: survival optimization and the nervous system. *Front. Neurosci.* 9, 55
5. Qi, S. *et al.* (2018) How cognitive and reactive fear circuits optimize escape decisions in humans. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3186–3191
6. Fanselow, M.S. (1994) Neural organization of the defensive behavior system responsible for fear. *Psychon. Bull. Rev.* 1, 429–438
7. Gross, C.T. and Canteras, N.S. (2012) The many paths to fear. *Nat. Rev. Neurosci.* 13, 651–658
8. McNaughton, N. and Corr, P.J. (2004) A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. *Neurosci. Biobehav. Rev.* 28, 285–305
9. Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, MIT Press
10. Daw, N.D. (2018) Are we of two minds? *Nat. Neurosci.* 21, 1497–1499
11. Sutton, R.S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Mach. Learn. Proc.* 1990, 216–224
12. American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders* (5th edn), American Psychiatric Association
13. Fung, B.J. *et al.* (2019) Slow escape decisions are swayed by trait anxiety. *Nat. Hum. Behav.* 3, 702–708
14. Otto, A.R. *et al.* (2013) Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci. U. S. A.* 110, 20941–20946
15. Heller, A.S. *et al.* (2018) Model-based learning and individual differences in depression: the moderating role of stress. *Behav. Res. Ther.* 111, 19–26
16. Radenbach, C. *et al.* (2015) The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology* 53, 268–280
17. Gillan, C.M. *et al.* (2016) Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife* 5, e11305

18. Debiec, J. and Olsson, A. (2017) Social fear learning: from animal models to human function. *Trends Cogn. Sci.* 21, 546–555
19. Lindstrom, B. *et al.* (2018) A common neural network differentially mediates direct and social fear learning. *Neuroimage* 167, 121–129
20. Olsson, A. and Phelps, E.A. (2007) Social learning of fear. *Nat. Neurosci.* 10, 1095–1102
21. Lindstrom, B. *et al.* (2019) Social threat learning transfers to decision making in humans. *Proc. Natl. Acad. Sci. U. S. A.* 116, 4732–4737
22. Pavlov, I.P. (1927) *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*, Oxford University Press
23. LeDoux, J.E. (2000) Emotion circuits in the brain. *Annu. Rev. Neurosci.* 23, 155–184
24. Rescorla, R.A. and Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II* (Black, A.H. and Prokasy, W.F., eds), pp. 64–99, Appleton-Century-Crofts
25. Resnik, J. *et al.* (2011) Auditory aversive learning increases discrimination thresholds. *Nat. Neurosci.* 14, 791–796
26. Laufer, O. *et al.* (2016) Behavioral and neural mechanisms of overgeneralization in anxiety. *Curr. Biol.* 26, 713–722
27. Pearce, J.M. and Hall, G. (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87, 532–552
28. Belova, M.A. *et al.* (2007) Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron* 55, 970–984
29. McNally, G.P. *et al.* (2011) Placing prediction into the fear circuit. *Trends Neurosci.* 34, 283–292
30. Klavir, O. *et al.* (2013) Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning. *Neuron* 80, 1290–1300
31. McHugh, S.B. *et al.* (2014) Aversive prediction error signals in the amygdala. *J. Neurosci.* 34, 9024–9033
32. Delgado, M.R. *et al.* (2005) An fMRI study of reward-related probability learning. *Neuroimage* 24, 862–873
33. Schiller, D. *et al.* (2008) From fear to safety and back: reversal of fear in the human brain. *J. Neurosci.* 28, 11517–11525
34. Homan, P. *et al.* (2019) Neural computations of threat in the aftermath of combat trauma. *Nat. Neurosci.* 22, 470–476
35. Bromberg-Martin, E.S. *et al.* (2010) Dopamine in motivational control: rewarding, aversive and alerting. *Neuron* 68, 815–834
36. Brooks, A.M. and Berns, G.S. (2013) Aversive stimuli and loss in the mesocorticolimbic dopamine system. *Trends Cogn. Sci.* 17, 281–286
37. Garrison, J. *et al.* (2013) Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* 37, 1297–1310
38. Delgado, M.R. *et al.* (2008) The role of striatum in aversive learning and aversive prediction errors. *Philos. Trans. R. Soc. B* 1511, 3787–3800
39. Roesch, M.R. *et al.* (2010) Neural correlates of variations in event processing during learning in basolateral amygdala. *J. Neurosci.* 30, 2464–2471
40. Li, J. *et al.* (2011) Differential roles of human striatum and amygdala in associative learning. *Nat. Neurosci.* 14, 1250–1252
41. Zhang, S. *et al.* (2016) Dissociable learning processes underlie human pain conditioning. *Curr. Biol.* 26, 52–58
42. Tzovara, A. *et al.* (2018) Human Pavlovian fear conditioning conforms to probabilistic learning. *PLoS Comput. Biol.* 14, e1006243
43. Brown, V.M. *et al.* (2018) Associability-modulated loss learning is increased in posttraumatic stress disorder. *Elife* 7, e30150
44. Skinner, B.F. (1938) *The Behavior of Organisms: An Experimental Analysis*, D. Appleton-Century
45. Tolman, E.C. (1932) *Purposive Behavior in Animals and Men*, The Century Company
46. Hull, C.L. (1943) *Principles of Behavior: An Introduction to Behavior Theory*, D. Appleton-Century
47. Mowrer, O.H. (1960) *Learning Theory and Behavior*, Wiley
48. LeDoux, J.E. *et al.* (2017) The birth, death and resurrection of avoidance: a reconceptualization of a troubled paradigm. *Mol. Psychiatry* 22, 24–36
49. Diehl, M.M. *et al.* (2019) The study of active avoidance: a platform for discussion. *Neurosci. Biobehav. Rev.* 107, 229–237
50. Delgado, M.R. *et al.* (2009) Avoiding negative outcomes: tracking the mechanisms of avoidance learning in humans during fear conditioning. *Front. Behav. Neurosci.* 3, 33
51. Collins, K.A. *et al.* (2014) Taking action in the face of threat: neural synchronization predicts adaptive coping. *J. Neurosci.* 34, 14733–14738
52. Boeke, E.A. *et al.* (2017) Active avoidance: neural mechanisms and attenuation of Pavlovian conditioned responding. *J. Neurosci.* 37, 4808–4818
53. Hofmann, S.G. and Hay, A.C. (2018) Rethinking avoidance: toward a balanced approach to avoidance in treating anxiety disorders. *J. Anxiety Disord.* 55, 14–21
54. Mowrer, O.H. (1940) Anxiety-reduction and learning. *J. Exp. Psychol.* 27, 497–516
55. Miller, N.E. (1948) Studies of fear as an acquirable drive .1. Fear as Motivation and Fear-Reduction as Reinforcement in the Learning of New responses. *J. Exp. Psychol.* 38, 89–101
56. Mathys, C. *et al.* (2011) A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* 5, 39
57. Gershman, S.J. (2015) A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* 11, e1004567
58. Dayan, P. *et al.* (2000) Learning and selective attention. *Nat. Neurosci.* 3, 1218–1223
59. Bland, A. and Schaefer, A. (2012) Unexpected uncertainty, volatility and decision-making. *Front. Neurosci.* 6, 00085
60. Payzan-LeNestour, E. and Bossaerts, P. (2011) Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput. Biol.* 7, e1001048
61. Soltani, A. and Izquierdo, A. (2019) Adaptive learning under expected and unexpected uncertainty. *Nat. Rev. Neurosci.* 20, 635–644
62. Nassar, M.R. *et al.* (2010) An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* 30, 12366–12378
63. Payzan-LeNestour, E. *et al.* (2013) The neural representation of unexpected uncertainty during value-based decision making. *Neuron* 79, 191–201
64. McGuire, J.T. *et al.* (2014) Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* 84, 870–881
65. Behrens, T.E. *et al.* (2007) Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221
66. Pulcu, E. and Browning, M. (2017) Affective bias as a rational response to the statistics of rewards and punishments. *Elife* 6, e27879
67. Nassar, M.R. *et al.* (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* 15, 1040
68. De Berker, A.O. *et al.* (2016) Computations of uncertainty mediate acute stress responses in humans. *Nat. Commun.* 7, 10996
69. Aylward, J. *et al.* (2019) Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nat. Hum. Behav.* 3, 1116–1123
70. Browning, M. *et al.* (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. *Biol. Psychiatry* 77, 47s–48s
71. Lamba, A. *et al.* (2020) Anxiety impedes adaptive social learning under uncertainty. *Psychol. Sci.* 31, 592–603
72. Dunsmoor, J.E. *et al.* (2015) Rethinking extinction. *Neuron* 88, 47–63
73. Golkar, A. *et al.* (2016) Neural signals of vicarious extinction learning. *Soc. Cogn. Affect. Neurosci.* 11, 1541–1549
74. Reddan, M.C. *et al.* (2018) Attenuating neural threat expression with imagination. *Neuron* 100, 994–1005.e4
75. Clem, R.L. and Schiller, D. (2016) New learning and unlearning: strangers or accomplices in threat memory attenuation? *Trends Neurosci.* 39, 340–351
76. Gershman, S.J. *et al.* (2015) Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* 5, 43–50

77. Gershman, S.J. *et al.* (2013) Gradual extinction prevents the return of fear: implications for the discovery of state. *Front. Behav. Neurosci.* 7, 164
78. Gershman, S.J. and Hartley, C.A. (2015) Individual differences in learning predict the return of fear. *Learn. Behav.* 43, 243–250
79. Gershman, S.J. *et al.* (2017) The computational nature of memory modification. *Elife* 6, e23763
80. Schuck, N.W. and Niv, Y. (2019) Sequential replay of nonspatial task states in the human hippocampus. *Science* 364, eaaw5181
81. Pearce, J.M. (1987) A model for stimulus generalization in Pavlovian conditioning. *Psychol. Rev.* 94, 61
82. Lissek, S. and van Meurs, B. (2015) Learning models of PTSD: theoretical accounts and psychobiological evidence. *Int. J. Psychophysiol.* 98, 594–605
83. Orederu, T. and Schiller, D. (2018) Fast and slow extinction pathways in defensive survival circuits. *Curr. Opin. Behav. Sci.* 24, 96–103
84. Vervliet, B. *et al.* (2013) Fear extinction and relapse: state of the art. *Annu. Rev. Clin. Psychol.* 9, 215–248
85. Feder, A. *et al.* (2019) The biology of human resilience: opportunities for enhancing resilience across the lifespan. *Biol. Psychiatry* 86, 443–453
86. Corchs, F. and Schiller, D. (2019) Threat-related disorders as persistent motivational states of defense. *Curr. Opin. Behav. Sci.* 26, 62–68
87. Agren, T. (2014) Human reconsolidation: a reactivation and update. *Brain Res. Bull.* 105, 70–82
88. Haubrich, J. *et al.* (2020) Impairments to consolidation, reconsolidation, and long-term memory maintenance lead to memory erasure. *Annu. Rev. Neurosci.* 43, 297–314
89. Lee, J.L. (2009) Reconsolidation: maintaining memory relevance. *Trends Neurosci.* 32, 413–420
90. Nader, K. and Hardt, O. (2009) A single standard for memory: the case for reconsolidation. *Nat. Rev. Neurosci.* 10, 224–234
91. Monfils, M.H. *et al.* (2009) Extinction–reconsolidation boundaries: key to persistent attenuation of fear memories. *Science* 324, 951–955
92. Schiller, D. *et al.* (2010) Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* 463, 49–53
93. Agren, T. *et al.* (2012) Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science* 337, 1550–1552
94. Lee, J.L. *et al.* (2017) An update on memory reconsolidation updating. *Trends Cogn. Sci.* 21, 531–545
95. Cahill, E.N. and Milton, A.L. (2019) Neurochemical and molecular mechanisms underlying the retrieval–extinction effect. *Psychopharmacology* 236, 111–132
96. Paulus, D.J. *et al.* (2019) Prospects for reconsolidation-focused treatments of substance use and anxiety-related disorders. *Curr. Opin. Psychol.* 30, 80–86
97. Kuijer, E.J. *et al.* (2020) Retrieval–extinction and relapse prevention: rewriting maladaptive drug memories? *Front. Behav. Neurosci.* 14, 23
98. James, E.L. *et al.* (2015) Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychol. Sci.* 26, 1201–1215
99. Hu, J. *et al.* (2018) Reminder duration determines threat memory modification in humans. *Sci. Rep.* 8, 8848
100. Eisenberg, M. *et al.* (2003) Stability of retrieved memory: inverse correlation with trace dominance. *Science* 301, 1102–1104. https://doi.org/10.1126/science.1086881
101. Pedreira, M.E. and Maldonado, H. (2003) Protein synthesis subserves reconsolidation or extinction depending on reminder duration. *Neuron* 38, 863–869
102. Suzuki, A. *et al.* (2004) Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *J. Neurosci. Off. J. Soc. Neurosci.* 24, 4787–4795. https://doi.org/10.1523/jneurosci.5491-03.2004
103. Lee, J.L. *et al.* (2006) Reconsolidation and extinction of conditioned fear: inhibition and potentiation. *J. Neurosci. Off. J. Soc. Neurosci.* 26, 10051–10056. https://doi.org/10.1523/JNEUROSCI.2466-06.2006
104. Cassini, L.F. *et al.* (2017) On the transition from reconsolidation to extinction of contextual fear memories. *Learn. Mem.* 24, 392–399
105. Lane, R.D. *et al.* (2015) Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: new insights from brain science. *Behav. Brain Sci.* 38, e1
106. Lane, R.D., Nadel, L., eds (2020) *Neuroscience of Enduring Change: Implications for Psychotherapy*, Oxford University Press
107. Ecker, B. *et al.* (2015) Minding the findings: let's not miss the message of memory reconsolidation research for psychotherapy. *Behav. Brain Sci.* 38, e7
108. Ecker, B. (2020) Erasing problematic emotional learnings. In *Neuroscience of Enduring Change: Implications for Psychotherapy* (Lane, R.D. and Nadel, L., eds), pp. 273–299, Oxford University Press
109. Gray, R.M. *et al.* (2017) The reconsolidation of traumatic memories (RTM) protocol for PTSD: a case study. *J. Exp. Psychother.* 20, 47–61
110. Kable, J.W. and Glimcher, P.W. (2009) The neurobiology of decision: consensus and controversy. *Neuron* 63, 733–745
111. Bartra, O. *et al.* (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* 76, 412–427
112. Levy, D.J. and Glimcher, P.W. (2012) The root of all value: a neural common currency for choice. *Curr. Opin. Neurobiol.* 22, 1027–1038
113. Piva, M. *et al.* (2019) The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *Elife* 8, e44939
114. Zhang, Z. *et al.* (2017) Distributed neural representation of saliency controlled value and category during anticipation of rewards and punishments. *Nat. Commun.* 8, 1907
115. Padoa-Schioppa, C. and Assad, J.A. (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226
116. Kable, J.W. and Glimcher, P.W. (2007) The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* 10, 1625–1633
117. Kahnt, T. *et al.* (2014) Disentangling neural representations of value and salience in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 111, 5000–5005
118. Tom, S.M. *et al.* (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315, 515–518
119. Fujiwara, J. *et al.* (2009) Segregated and integrated coding of reward and punishment in the cingulate cortex. *J. Neurophysiol.* 101, 3284–3293
120. O'Doherty, J. *et al.* (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* 4, 95–102
121. Kim, S.H. *et al.* (2015) Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning. *Soc. Cogn. Affect. Neurosci.* 10, 1219–1227
122. Liu, X. *et al.* (2011) Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* 35, 1219–1236
123. Monosov, I.E. and Hikosaka, O. (2012) Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. *J. Neurosci.* 32, 10318–10330
124. Del Arco, A. *et al.* (2020) Unanticipated stressful and rewarding experiences engage the same prefrontal cortex and ventral tegmental area neuronal populations. *eNeuro* 7 ENEURO.0029-20.2020
125. Litt, A. *et al.* (2011) Dissociating valuation and saliency signals during decision-making. *Cereb. Cortex* 21, 95–102
126. O'Neill, P.-K. *et al.* (2018) Basolateral amygdala circuitry in positive and negative valence. *Curr. Opin. Neurobiol.* 49, 175–183
127. Namburi, P. *et al.* (2015) A circuit mechanism for differentiating positive and negative associations. *Nature* 520, 675–678
128. Levy, I. *et al.* (2010) Neural representation of subjective value under risk and ambiguity. *J. Neurophysiol.* 103, 1036–1047
129. Levy, D.J. and Glimcher, P.W. (2011) Comparing apples and oranges: using reward-specific and reward-general

subjective value representation in the brain. *J. Neurosci.* 31, 14693–14707
130. Dreher, J.-C. *et al.* (2006) Neural coding of distinct statistical properties of reward information in humans. *Cereb. Cortex* 16, 561–573
131. Preuschoff, K. *et al.* (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51, 381–390
132. Symmonds, M. *et al.* (2010) A behavioral and neural evaluation of prospective decision-making under risk. *J. Neurosci.* 30, 14380–14389
133. Tobler, P.N. *et al.* (2007) Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J. Neurophysiol.* 97, 1621–1632
134. Hsu, M. *et al.* (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683
135. Kuhnen, C.M. and Knutson, B. (2005) The neural basis of financial risk taking. *Neuron* 47, 763–770
136. Huettel, S.A. *et al.* (2005) Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. *J. Neurosci.* 25, 3304–3311
137. Huettel, S.A. *et al.* (2006) Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49, 765–775
138. Symmonds, M. *et al.* (2011) Deconstructing risk: separable encoding of variance and skewness in the brain. *Neuroimage* 58, 1139–1149
139. Preuschoff, K. *et al.* (2008) Human insula activation reflects risk prediction errors as well as risk. *J. Neurosci.* 28, 2745–2752
140. Mohr, P.N. *et al.* (2010) Neural foundations of risk–return trade-off in investment decisions. *Neuroimage* 49, 2556–2563
141. Bach, D.R. *et al.* (2009) Neural activity associated with the passive prediction of ambiguity and risk for aversive events. *J. Neurosci.* 29, 1648–1656
142. Christopoulos, G.I. *et al.* (2009) Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *J. Neurosci.* 29, 12574–12583
143. Gilaie-Dotan, S. *et al.* (2014) Neuroanatomy predicts individual risk attitudes. *J. Neurosci.* 34, 12394–12401
144. Grubb, M.A. *et al.* (2016) Neuroanatomy accounts for age-related changes in risk preferences. *Nat. Commun.* 7, 13822
145. FeldmanHall, O. *et al.* (2019) The functional roles of the amygdala and prefrontal cortex in processing uncertainty. *J. Cogn. Neurosci.* 31, 1742–1754
146. Jung, W.H. *et al.* (2018) Amygdala functional and structural connectivity predicts individual risk tolerance. *Neuron* 98, 394–404.e4
147. Tversky, A. and Kahneman, D. (1991) Loss aversion in riskless choice – a reference-dependent model. *Q. J. Econ.* 106, 1039–1061
148. Park, S.Q. *et al.* (2011) Neurobiology of value integration: when value impacts valuation. *J. Neurosci.* 31, 9307–9314
149. Paton, J.J. *et al.* (2006) The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* 439, 865–870
150. Orsini, C.A. *et al.* (2015) Dissociable roles for the basolateral amygdala and orbitofrontal cortex in decision-making under risk of punishment. *J. Neurosci.* 35, 1368–1379
151. Orsini, C.A. *et al.* (2017) Optogenetic inhibition reveals distinct roles for basolateral amygdala activity at discrete time points during risky decision making. *J. Neurosci.* 37, 11537–11548
152. Herry, C. *et al.* (2007) Processing of temporal unpredictability in human and animal amygdala. *J. Neurosci.* 27, 5958–5966
153. Koscik, T.R. *et al.* (2020) Decomposing the neural pathways in a simple, value-based choice. *Neuroimage* 214, 116764
154. Hunt, L.T. *et al.* (2012) Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.* 15, S1–S3
155. Strait, C.E. *et al.* (2014) Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. *Neuron* 82, 1357–1366
156. Shabel, S.J. *et al.* (2019) Stress transforms lateral habenula reward responses into punishment signals. *Proc. Natl. Acad. Sci. U. S. A.* 116, 12488–12493
157. Reynolds, S.M. and Berridge, K.C. (2008) Emotional environments retune the valence of appetitive versus fearful functions in nucleus accumbens. *Nat. Neurosci.* 11, 423–425
158. Jia, R. *et al.* (2020) From value to saliency: neural computations of subjective value under uncertainty in combat veterans. *bioRxiv* Published online June 24, 2020. https://doi.org/10.1101/2020.04.14.041467
159. Ruderman, L. *et al.* (2016) Posttraumatic stress symptoms and aversion to ambiguous losses in combat veterans. *Depress. Anxiety* 33, 606–613
160. Oglesby, M.E. *et al.* (2017) Intolerance of uncertainty and post-traumatic stress symptoms: an investigation within a treatment seeking trauma-exposed sample. *Compr. Psychiatry* 72, 34–40
161. Rosser, B.A. (2019) Intolerance of uncertainty as a transdiagnostic mechanism of psychological difficulties: a systematic review of evidence pertaining to causality and temporal precedence. *Cogn. Ther. Res.* 43, 438–463
162. Pushkarskaya, H. *et al.* (2015) Decision-making under uncertainty in obsessive–compulsive disorder. *J. Psychiatr. Res.* 69, 166–173
163. Charpentier, C.J. *et al.* (2017) Enhanced risk aversion, but not loss aversion, in unmedicated pathological anxiety. *Biol. Psychiatry* 81, 1014–1022
164. Boelen, P.A. and Reijntjes, A. (2009) Intolerance of uncertainty and social anxiety. *J. Anxiety Disord.* 23, 130–135
165. Grupe, D.W. and Nitschke, J.B. (2013) Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* 14, 488–501
166. Kube, T. *et al.* (2020) Rethinking post-traumatic stress disorder – a predictive processing perspective. *Neurosci. Biobehav. Rev.* 113, 448–460
167. Admon, R. *et al.* (2013) A causal model of post-traumatic stress disorder: disentangling predisposed from acquired neural abnormalities. *Trends Cogn. Sci.* 17, 337–347
168. Silva, B.A. *et al.* (2016) The neural circuits of innate fear: detection, integration, action, and memorization. *Learn. Mem.* 23, 544–555
169. Schiller, D. and Delgado, M.R. (2010) Overlapping neural systems mediating extinction, reversal and regulation of fear. *Trends Cogn. Sci.* 14, 268–276
170. LeDoux, J.E. and Schiller, D. (2009) The human amygdala: insights from other animals. In *The Human Amygdala*, pp. 43–60, Guilford Press
171. Balleine, B.W. and Killcross, S. (2006) Parallel incentive processing: an integrated view of amygdala function. *Trends Neurosci.* 29, 272–279
172. Davis, M. and Shi, C. (1999) The extended amygdala: are the central nucleus of the amygdala and the bed nucleus of the stria terminalis differentially involved in fear versus anxiety? *Ann. N. Y. Acad. Sci.* 877, 281–291
173. Maren, S. (2001) Neurobiology of Pavlovian fear conditioning. *Annu. Rev. Neurosci.* 24, 897–931
174. Fanselow, M.S. and Poulos, A.M. (2005) The neuroscience of mammalian associative learning. *Annu. Rev. Psychol.* 56, 207–234
175. Fanselow, M.S. (1998) Pavlovian conditioning, negative feedback, and blocking: mechanisms that regulate association formation. *Neuron* 20, 625–627
176. Herry, C. and Johansen, J.P. (2014) Encoding of fear learning and memory in distributed neuronal circuits. *Nat. Neurosci.* 17, 1644
177. Ozawa, T. *et al.* (2017) A feedback neural circuit for calibrating aversive memory strength. *Nat. Neurosci.* 20, 90
178. Ehrlich, I. *et al.* (2009) Amygdala inhibitory circuits and the control of fear memory. *Neuron* 62, 757–771
179. Tye, K.M. *et al.* (2011) Amygdala circuitry mediating reversible and bidirectional control of anxiety. *Nature* 471, 358–362
180. Herry, C. *et al.* (2008) Switching on and off fear by distinct neuronal circuits. *Nature* 454, 600–606
181. Beyeler, A. *et al.* (2016) Divergent routing of positive and negative information from the amygdala during memory retrieval. *Neuron* 90, 348–361
182. Sotres-Bayon, F. *et al.* (2006) Brain mechanisms of fear extinction: historical perspectives on the contribution of prefrontal cortex. *Biol. Psychiatry* 60, 329–336

183. Quirk, G.J. and Mueller, D. (2008) Neural mechanisms of extinction learning and retrieval. *Neuropsychopharmacology* 33, 56–72
184. Burgos-Robles, A. *et al.* (2009) Sustained conditioned responses in prelimbic prefrontal neurons are correlated with fear expression and extinction failure. *J. Neurosci.* 29, 8474–8482
185. Burgos-Robles, A. *et al.* (2017) Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment. *Nat. Neurosci.* 20, 824–835
186. Milad, M.R. and Quirk, G.J. (2012) Fear extinction as a model for translational neuroscience: ten years of progress. *Annu. Rev. Psychol.* 63, 129–151
187. Quirk, G.J. *et al.* (2003) Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. *J. Neurosci.* 23, 8800–8807
188. Janak, P.H. and Tye, K.M. (2015) From circuits to behaviour in the amygdala. *Nature* 517, 284–292
189. Diehl, M.M. *et al.* (2020) Divergent projections of the prelimbic cortex bidirectionally regulate active avoidance. *Elife* 9, e59281
190. Moscarello, J.M. and LeDoux, J.E. (2013) Active avoidance learning requires prefrontal suppression of amygdala-mediated defensive reactions. *J. Neurosci.* 33, 3815–3823
191. Martinez-Rivera, F.J. *et al.* (2019) Prefrontal circuits signaling active avoidance retrieval and extinction. *Psychopharmacology* 236, 399–406
192. Wanke, N. and Schwabe, L. (2020) Dissociable neural signatures of passive extinction and instrumental control over threatening events. *Soc. Cogn. Affect. Neurosci.* 15, 625–634
193. Moscarello, J.M. and Hartley, C.A. (2017) Agency and the calibration of motivated behavior. *Trends Cogn. Sci.* 21, 725–735
194. North, C.S. *et al.* (2016) The evolution of PTSD criteria across editions of DSM. *Ann. Clin. Psychiatry* 28, 197–208
195. Pai, A. *et al.* (2017) Posttraumatic stress disorder in the DSM-5: controversy, change, and conceptual considerations. *Behav. Sci. (Basel)* 7, 7
196. Raber, J. *et al.* (2019) Current understanding of fear learning and memory in humans and animal models and the value of a linguistic approach for analyzing fear learning and memory in humans. *Neurosci. Biobehav. Rev.* 105, 136–177
197. LeDoux, J.E. and Pine, D.S. (2016) Using neuroscience to help understand fear and anxiety: a two-system framework. *Am. J. Psychiatry* 173, 1083–1093
198. Foa, E.B. (2011) Prolonged exposure therapy: past, present, and future. *Depress. Anxiety* 28, 1043–1047
199. Foa, E.B. and McLean, C.P. (2016) The efficacy of exposure therapy for anxiety-related disorders and its underlying mechanisms: the case of OCD and PTSD. *Annu. Rev. Clin. Psychol.* 12, 1–28
200. Stein, D.J. *et al.* (2009) Pharmacotherapy of posttraumatic stress disorder: a review of meta-analyses and treatment guidelines. *CNS Spectr.* 14, 25–31
201. Institute of Medicine (2008) *Treatment of Posttraumatic Stress Disorder: An Assessment of the Evidence*, National Academies Press
202. Lebois, L.A.M. *et al.* (2019) Augmentation of extinction and inhibitory learning in anxiety and trauma-related disorders. *Annu. Rev. Clin. Psychol.* 15, 257–284
203. Elsey, J.W.B. and Kindt, M. (2017) Tackling maladaptive memories through reconsolidation: from neural to clinical science. *Neurobiol. Learn. Mem.* 142, 108–117
204. Kida, S. (2019) Reconsolidation/destabilization, extinction and forgetting of fear memory as therapeutic targets for PTSD. *Psychopharmacology* 236, 49–57
205. Monfils, M.H. and Holmes, E.A. (2018) Memory boundaries: opening a window inspired by reconsolidation to treat anxiety, trauma-related, and addiction disorders. *Lancet Psychiatry* 5, 1032–1042
206. Angela, J.Y. and Dayan, P. (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46, 681–692
207. Kobayashi, K. and Hsu, M. (2017) Neural mechanisms of updating under reducible and irreducible uncertainty. *J. Neurosci.* 37, 6972–6982
208. Pulcu, E. and Browning, M. (2019) The misestimation of uncertainty in affective disorders. *Trends Cogn. Sci.* 23, 865–875
209. Faraji, M. *et al.* (2018) Balancing new against old information: the role of puzzlement surprise in learning. *Neural Comput.* 30, 34–83
210. Farashahi, S. *et al.* (2017) Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* 94, 401–414.e6
211. Levy, I. (2017) Neuroanatomical substrates for risk behavior. *Neuroscientist* 23, 275–286
212. Holt, C.A. and Laury, S.K. (2002) Risk aversion and incentive effects. *Am. Econ. Rev.* 92, 1644–1655
213. Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291
214. Abdellaoui, M. *et al.* (2007) Loss aversion under prospect theory: a parameter-free measurement. *Manag. Sci.* 53, 1659–1674
215. Tymula, A. *et al.* (2013) Like cognitive function, decision making across the life span shows profound age-related changes. *Proc. Natl. Acad. Sci. U. S. A.* 110, 17143–17148
216. Ellsberg, D. (1961) Risk, ambiguity, and the savage axioms. *Q. J. Econ.* 75, 643–669
217. Camerer, C. and Weber, M. (1992) Recent developments in modeling preferences – uncertainty and ambiguity. *J. Risk Uncertain.* 5, 325–370
218. Wakker, P.P. (2010) *Prospect Theory for Risk and Ambiguity*, Cambridge University Press
219. Baillon, A. and Bleichrodt, H. (2015) Testing ambiguity models through the measurement of probabilities for gains and losses. *Am. Econ. J. Microecon.* 7, 77–100
220. FeldmanHall, O. *et al.* (2016) Emotion and decision-making under uncertainty: physiological arousal predicts increased gambling during ambiguity but not risk. *J. Exp. Psychol. Gen.* 145, 1255–1262
221. Konova, A.B. *et al.* (2020) Computational markers of risky decision-making for identification of temporal windows of vulnerability to opioid use in a real-world clinical setting. *JAMA Psychiatry* 77, 368–377
222. Charness, G. *et al.* (2020) Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *J. Risk Uncertain.* 60, 99–123
223. Seaman, K.L. *et al.* (2016) Adult age differences in decision making across domains: increased discounting of social and health-related rewards. *Psychol. Aging* 31, 737–746
224. Frey, R. *et al.* (2017) Risk preference shares the psychometric structure of major psychological traits. *Behav. Genet.* 47, 686–687
225. Pedroni, A. *et al.* (2017) The risk elicitation puzzle. *Nat. Hum. Behav.* 1, 803–809
226. Holzmeister, F. and Stefan, M. (2020) The risk elicitation puzzle revisited: across-methods (in)consistency? *Exp. Econ.* Published online September 12, 2020. https://doi.org/10.1007/s10683-020-09674-8
227. Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press
228. Brown, R. *et al.* (2019) Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768
229. LeDoux, J. (2012) Rethinking the emotional brain. *Neuron* 73, 653–676
230. LeDoux, J. and Daw, N.D. (2018) Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nat. Rev. Neurosci.* 19, 269
231. LeDoux, J.E. (2014) Coming to terms with fear. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2871–2878
232. Lindquist, K.A. and Barrett, L.F. (2012) A functional architecture of the human brain: emerging insights from the science of emotion. *Trends Cogn. Sci.* 16, 533–540
233. Bach, D.R. and Dayan, P. (2017) Algorithms for survival: a comparative perspective on emotions. *Nat. Rev. Neurosci.* 18, 311
234. Atlas, L.Y. *et al.* (2016) Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *Elife* 5, e15192
235. DeWall, C.N. *et al.* (2016) How often does currently felt emotion predict social behavior and judgment? A meta-analytic test of two theories. *Emot. Rev.* 8, 136–143
236. Stefanova, E. *et al.* (2020) Anticipatory feelings: neural correlates and linguistic markers. *Neurosci. Biobehav. Rev.* 113, 308–324
237. Eldar, E. *et al.* (2016) Mood as representation of momentum. *Trends Cogn. Sci.* 20, 15–24