ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202601.00199

Safety Learning and Anxiety Disorders: A Three-Stage Neural Mechanism Framework

Authors: Dong Zhanpeng, Zhang Jie, Zhang Yarui, Lei Yi, Lei Yi

Date: 2026-01-23T14:27:09+00:00

Abstract

Safety learning refers to an adaptive mechanism by which individuals suppress fear through the identification of safety signals (cues indicating the absence of threat). Impairment of this function is closely associated with the onset of anxiety and related disorders. Building on existing research, the present study proposes a three-stage neurobiological framework of safety learning, composed of "perceptual evaluation of safety, acquisition of safety, and behavioral expression of safety." This framework emphasizes that individuals with anxiety disorders exhibit corresponding neural abnormalities across these three stages of safety learning: in the perceptual evaluation stage, abnormalities are observed in circuits centered on the insula and sensory cortex, which undermine effective identification of safety cues; in the acquisition stage, abnormalities in midbrain-striatal circuits hinder the formation of safety associations and the experience of positive emotions; and in the behavioral expression stage, dysfunction of inhibitory circuits centered on the hippocampus makes it difficult for individuals to retrieve safety memories to suppress fear responses. Future research should employ paradigms and measurement techniques with higher ecological validity to elucidate stage-specific abnormalities in safety learning among anxious individuals, thereby promoting the translational application of safety learning in clinical interventions.

Full Text

Safety Learning and Anxiety Disorders: A Three-Stage Neural Mechanism Framework

DONG Zhanpeng, ZHANG Jie, ZHANG Yarui, LEI Yi
Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu 610066, China

Abstract
Safety learning refers to an adaptive mechanism through which individuals suppress fear by identifying safety signals—cues that predict the absence of threat—and impairments in this process are closely associated with the emergence of anxiety and related disorders. Based on existing evidence, the present work seeks to propose a three-stage neurobiological framework of safety learning, comprising safety perception and appraisal, safety acquisition, and safety behavioral expression. This framework emphasizes that individuals with anxiety disorders exhibit stage-specific neural abnormalities across the three phases of safety learning. Specifically, during the perception and appraisal stage, abnormalities in circuits centered on the insula and perceptual cortices impair the effective identification of safety cues; during the acquisition stage, dysfunctions in midbrain–striatal circuits hinder the formation of safety associations and the acquisition of positive emotional experiences; and during the behavioral expression stage, functional abnormalities in hippocampus-centered inhibitory circuits make it difficult to retrieve safety memories to suppress fear responses. Future research should employ paradigms and measurement approaches with higher ecological validity to characterize stage-specific abnormalities in safety learning among anxious individuals, thereby facilitating the translational application of safety learning in clinical interventions.

Keywords: safety learning, anxiety, neural mechanisms

Safety learning refers to the adaptive mechanism by which individuals identify "no-threat" cues to inhibit fear responses (Christianson et al., 2012; Laing & Harrison, 2021; Laing et al., 2024; Tashjian et al., 2021). For example, encountering a tiger in the wild triggers intense fear and escape behaviors, whereas observing the same animal through a safety barrier at a zoo elicits no fear. The barrier serves as a safety signal, enabling individuals to assess the situation as non-threatening and thereby suppress fear responses. This capacity to learn safety is crucial for navigating complex and dynamic environments. When impaired, it often leads to maladaptive behaviors and can develop into anxiety disorders. Existing research indicates that various anxiety-related disorders—including social anxiety disorder (SAD), generalized anxiety disorder (GAD), post-traumatic stress disorder (PTSD), and obsessive-compulsive disorder (OCD)—are associated with abnormalities in safety learning (Jovanovic et al., 2012; Cooper et al., 2018; Cooper & Dunsmoor, 2021; Lewis et al., 2024), manifesting primarily as difficulty distinguishing between safety and threat stimuli, inefficient safety acquisition, and failure of fear inhibition.

Although anxiety is closely linked to deficits in safety learning, previous understanding of anxiety pathogenesis has focused predominantly on fear learning research. This line of work suggests that anxiety disorders arise from excessive generalization or persistence of fear responses, with neurobiological investigations primarily targeting regions such as the amygdala, dorsal anterior cingulate cortex (dACC), and insula that show hyperactivity during fear expression (Graham & Milad, 2011; Sevenster et al., 2018; 雷怡 et al., 2018; 申思怡 et al.,

2023). Additionally, fear extinction—the process of establishing a new association between a conditioned threat stimulus (CS) and "threat absence" (safety) to inhibit fear responses—is often used to study how people unlearn fear. This process provides a theoretical foundation for clinical interventions such as exposure therapy and can be considered a special form of safety learning. However, relying solely on this "fear-centric" perspective to understand safety acquisition may be limited, as it fails to explain common manifestations in anxiety disorders such as difficulty identifying safety and lack of positive affect. Therefore, it is essential to further examine how individuals learn safety signals and their role in fear regulation.

Recent research suggests that safety learning is an inhibitory learning mechanism distinct from fear learning. At the behavioral level, individuals learn that a CS predicts "no threat," thereby using learned "safety" to reduce fear responses (Christianson et al., 2012; Kong et al., 2014). Unlike fear learning, which produces defensive behaviors such as freezing and avoidance (LeDoux & Pine, 2016), safety learning manifests as fear reduction. For instance, Pollak et al. (2008) found in mice that freezing behavior (a conditioned fear response) significantly decreased during safety signal presentation, indicating that safety signals can attenuate conditioned fear expression. Using the elevated plus maze (EPM) paradigm, researchers further observed that safety-trained mice showed increased entries and time spent in open arms, reflecting reduced anxiety and enhanced exploratory behavior. Human studies provide convergent evidence, showing that safety signals learned in safety learning tasks significantly reduce subjective threat appraisal (Odriozola et al., 2024), offering important theoretical and methodological foundations for developing anxiety interventions (Laing et al., 2024).

At the neural circuit level, although both threat and safety learning depend on amygdala-prefrontal-hippocampal circuits, different subregions are activated. For example, in rodents, the prelimbic cortex (PL)—considered homologous to human dACC/dorsomedial prefrontal cortex (dmPFC)—is more involved in fear expression and discrimination, whereas the infralimbic cortex (IL)—homologous to human ventromedial prefrontal cortex (vmPFC)—is more involved in fear inhibition and safety cue processing (Sangha et al., 2020). Duvarci (2024) further noted that fear association formation is primarily supported by the periaqueductal gray/dorsal raphe (PAG/DR)-central amygdala (CeA) dopamine circuit, which is necessary for threat association formation. In contrast, safety association formation relies on dopaminergic activity in the ventral tegmental area (VTA)-anteromedial nucleus accumbens (amNAc) circuit, which becomes enhanced during safety learning to support safety association establishment.

At the more microscopic synaptic and molecular level, both safety and fear association formation and maintenance involve long-term potentiation (LTP) mediated by NMDA and AMPA receptors, but LTP occurs in different neuronal types. Fear association LTP primarily occurs in excitatory neurons of the lateral amygdala (LA) to support fear memory formation and consolidation (Palchaud-

huri et al., 2024). Safety association LTP, however, occurs more frequently in GABAergic intercalated cells between the basolateral amygdala (BLA) and CeA, which inhibit CeA to reduce fear output (Asede et al., 2022). Additionally, Harb et al. (2021) demonstrated that brain-derived neurotrophic factor (BDNF) deficiency does not significantly affect fear association formation but impairs safety association establishment, suggesting that safety learning may rely more heavily on BDNF to support synaptic plasticity and learning updates in relevant circuits compared to fear learning. These findings indicate that safety learning is not simply the opposite of fear learning but rather an adaptive learning mechanism supported by relatively independent neural systems with distinct functional orientations (see Table 1).

Table 1 Differences Between Safety and Fear Learning

Safety Learning	Fear Learning
Establishes "CS→no threat/safety" association (inhibitory learning) VTA→amNAc dopamine circuit encodes prediction error signals for "threat absence" ; vmPFC is critical for safety cue processing and fear inhibition	Establishes "CS→threat" association (excitatory learning) PAG/DR→CeA dopamine circuit encodes prediction error signals for "threat presence" ; dmPFC/dACC is more important for fear expression
Molecular and Plasticity Mechanisms LTP for safety learning occurs more in BLA-CeA GABAergic intercalated cells (ITCs); BDNF deficiency impairs safety association formation	**Molecular and Plasticity Mechanisms** LTP for fear learning occurs primarily in excitatory neurons of LA, dependent on NMDA/AMPA receptors; less affected by BDNF deficiency
Behavioral Manifestations Fear/anxiety reduction and increased approach behavior	**Behavioral Manifestations** Enhanced fear, avoidance, and vigilance

To more systematically reveal the neural mechanisms of safety learning, Laing et al. (2022a) proposed a framework dividing safety learning into "acquisition" and "expression"stages based on functional and mechanistic differences in safety association formation and subsequent behavioral responses. While this dichotomous framework offers a new perspective for understanding safety learning mechanisms, it may be overly simplistic. It fails to fully capture the complete process of information reception and integration in real-world contexts and provides insufficient detail about the specific neural circuits involved in the expression stage. Therefore, expanding and deepening the existing theoretical framework is necessary to more comprehensively describe the neural mechanisms underlying safety learning.

In summary, safety learning is not merely the simple counterpart of fear learning but an adaptive behavioral process supported by relatively independent neural systems. Investigating its neural mechanisms is crucial for understanding anxiety and related disorders. However, a key question remains: At which neurobiological stages do deficits in safety learning manifest in individuals with anxiety and related disorders?

Addressing this question, the present article proposes a "perception-acquisition-expression" three-stage neural mechanism framework for safety learning based on a review of existing research. This framework aims to more comprehensively reveal the processes of safety learning and the stage-specific abnormalities in anxious individuals, thereby providing new insights for precision interventions and mechanistic research on anxiety disorders.

2.1 Definition of Safety Learning

From a fear learning perspective, "safety" lacks explicit learning value and is typically defined as the absence of a "threat" outcome. From a safety learning perspective, however, safety establishment is viewed as an active learning process in which individuals gradually form stable associations between safety cues and no-threat outcomes. Tashjian et al. (2021) conceptualize safety as an individual' s perception of current and future survival possibilities, existing on a continuum from "zero safety" to "complete safety." "Zero safety" refers to imminent and unavoidable danger where threat is no longer potential or mitigable; in such contexts, fear emerges as an emotional response to danger and may intensify with increasing threat level. Complete safety refers to the absence of actual or anticipated future threats, where no fear response occurs.

Safety can be further categorized into contextual safety (e.g., shelter), extinction-based safety (previously dangerous stimuli become safe), safety signals (threat is mitigated by external factors), and prospective safety (future safety is anticipated through behavior). In other words, the core of safety learning lies in associating cues with safety outcomes (e.g., absence of aversive stimuli or positive outcomes). In existing research, this cue-safety outcome association learning is manifested not only in the inhibition of original threat memories during fear extinction but also in active discrimination between threat and safety cues in safety discrimination tasks, as well as in conditioned inhibition and avoidance tasks where individuals use safety signals to suppress or avoid potential threat outcomes (Grasser & Jovanovic, 2021; Sangha et al., 2020). Based on these definitions, we next review common experimental paradigms of safety learning to reveal its unique role in adaptive behavior.

2.2 Research Paradigms of Safety Learning

Safety learning is defined in research as a Pavlovian conditioning process that trains individuals to associate a stimulus with a safe outcome, thereby inhibiting fear responses to threat (Laing et al., 2024). Prediction error (PE) theory pro-

vides an important explanatory framework for this field. PE occurs when actual outcomes differ from expectations, reflecting discrepancies between reality and prediction and serving as a core mechanism driving "new learning" (Rescorla & Wagner, 1972). Based on this framework, researchers commonly use discriminative conditioning and conditioned inhibition paradigms to investigate safety learning.

The discriminative conditioning paradigm is a classic Pavlovian approach that shapes the safety properties of specific stimuli. In this paradigm, two or more CSs are typically presented, where at least one stimulus (CS+) is paired with an aversive unconditioned stimulus (US, e.g., electric shock) to become a threat signal, while another stimulus (CS-) is always presented alone and never paired with the US. This design is applied in two types of studies: threat-safety discrimination studies (Harrison et al., 2017; Mueller et al., 2024), where individuals learn to discriminate between CS+ and CS-, with CS- serving as a contrast condition that acquires relative safety meaning through its consistent no-threat property; and fear extinction studies, where individuals first establish a CS+-US association during acquisition and then undergo extinction training where CS+ is repeatedly presented alone without the US. This "expected threat did not occur" experience generates PE, prompting individuals to update the meaning of CS+ from threat to safety (曹杨婧文 et al., 2019; Deng et al., 2021; Rashidi et al., 2025). In such tasks, "safety" emerges through re-evaluation of previously conditioned fear cues. Therefore, the discriminative conditioning paradigm can reveal individuals' stimulus discrimination abilities and mechanisms of extinction memory formation, with high ecological validity and clinical significance (Zhao et al., 2021; Trent et al., 2025).

The conditioned inhibition paradigm, also known as feature-negative discrimination, is another classic experimental approach to studying safety learning (Kribakaran et al., 2024; Krueger et al., 2024). Its core focus is learning how specific cues can disarm fear responses to existing threat signals by investigating the inhibitory relationship between safety and threat stimuli (Holland, 1989; Holland & Lamarre, 1984; Lovibond & Lee, 2021). In safety learning tasks using this paradigm, individuals learn that a conditioned stimulus (A, i.e., CS+) is associated with an aversive US in some trials. In other trials, CS+ is presented simultaneously with another stimulus (called the conditioned inhibitor, X), forming an AX- compound, and the US is omitted. Here, X in the AX- compound generates PE through repeated "threat absence that violates expectation" and is learned as a safety signal. Individuals realize that X' s presence predicts that danger will not occur even when the threat cue A+ is present. Thus, safety learning tasks using the conditioned inhibition paradigm emphasize "threat absence" as a reinforcing condition. As a "safety signal," X' s appearance represents "threat absence," and through this reinforcement mechanism of "threat absence," individuals actively learn to associate X with "safety."

Additionally, other experimental paradigms involve associating cues with safety outcomes. For example, in active avoidance tasks (Kleine et al., 2023; Fisher

& Urcelay, 2024), individuals can learn specific behaviors to avoid threats and experience a sense of "relief" upon successful avoidance. In backward fear conditioning tasks (Andreatta et al., 2010; Gründahl et al., 2022), the US precedes the CS, which then signals the safety period following US termination. Safety assessment studies investigate how individuals dynamically predict future safety by integrating external threat cues with self-protection information in threatening contexts (Tashjian et al., 2025). These paradigms further enrich the research perspective on safety learning, demonstrating that safety learning not only inhibits fear but also generates positive emotional and motivational value to guide subsequent behavior. Thus, learning "safety" is not a single fear inhibition process but a complex adaptive mechanism encompassing cognitive appraisal, motivational drive, and emotional regulation.

In summary, safety learning is a PE-driven conditioning process whose core lies in forming "stimulus-safety"associations to inhibit fear responses. Discriminative conditioning emphasizes learning a stimulus itself as a safety signal, while conditioned inhibition emphasizes that a stimulus inhibits fear when co-present with threat cues. Active avoidance, backward conditioning, and safety assessment paradigms provide complementary approaches, collectively offering important empirical pathways for understanding safety learning deficits in anxiety and related disorders.

2.3 The Acquisition-Expression Framework of Safety Learning Neural Mechanisms

Given the differences observed in safety acquisition and expression processes in human and animal studies, Laing et al. (2022a) proposed an "acquisition-expression" framework for safety learning based on neurobiological studies of Pavlovian conditioned inhibition and fear extinction.

In this framework, "safety acquisition" is defined as the process of establishing stimulus-safety associations through repeated trial-and-error, driven by PE mechanisms. This process involves neural circuits comprising the basal ganglia, thalamus, insula, and association cortex. Specifically, during early acquisition, the midbrain-striatum circuit plays a key role, with the dorsal striatum (DS)-substantia nigra pars compacta (SNc) circuit primarily encoding inhibitory prediction error signals (CS→no US) and the ventral striatum (VS)-VTA circuit primarily encoding excitatory prediction error signals (CS→US). As safety learning progresses and safety outcomes become predictable, association memories are ultimately consolidated through basal ganglia-thalamus-cortex circuits.

The framework further posits that "safety expression" refers to behavioral, physiological, and cognitive operations based on acquired safety information, comprising three core components: threat response inhibition, positive affect and hedonic value, and memory maintenance and recall. First, threat response inhibition is the most direct functional manifestation of safety expression, involving brain regions such as vmPFC, hippocampus, and orbitofrontal cortex (OFC),

which directly or indirectly inhibit amygdala activity to reduce fear responses to safety stimuli (CS-). Second, safety stimuli carry certain positive affective and reward values associated with activation in reward-related regions like vmPFC. Additionally, safety expression depends on memory maintenance and contextual regulation of safety associations. The hippocampus and vmPFC jointly participate in retrieving safety memories and contextual judgments, determining when and under what circumstances to initiate inhibitory responses. Overall, this framework emphasizes that safety expression is not a passive reflection of non-threat states but a structured cognitive-neural process encompassing affective evaluation, memory, and regulated behavioral responses.

2.4 Limitations of Existing Frameworks

The acquisition-expression framework of safety learning was the first to systematically delineate the neural processes of human safety learning, offering a new perspective for understanding safety processing deficits in emotional disorders. However, this dichotomous framework has several limitations. First, it cannot capture the entire process of individual safety learning. Before the learning stage, individuals engage in preliminary perception of stimuli, which importantly influences subsequent learning—a factor evident in previous safety learning research (Kong et al., 2014) and anxiety circuit studies (Calhoon & Tye, 2015). Abnormalities in these processes may also affect safety learning. Second, the framework' s definition of striatal circuits during the learning stage has certain limitations; DS and VS circuits may reflect more complex functional divisions rather than simple excitatory-inhibitory oppositions. The dopamine PE signal generated after threat avoidance, which resembles reward and is subjectively accompanied by a "relief" experience (Badarnee et al., 2025), may be essential for establishing safety associations rather than merely a post-acquisition manifestation. Finally, the acquisition-expression framework does not elaborate on differences in fear inhibition neural circuits across safety learning types; for example, conditioned inhibition-based safety learning may depend on hippocampal-cingulate interactions, which differ from the vmPFC-amygdala circuit underlying fear extinction (Meyer et al., 2019). These factors collectively indicate the need to further deepen and expand existing theoretical frameworks to more clearly explain the neural mechanisms of individual safety learning.

Therefore, to address these limitations, this study constructs a three-stage "perception-acquisition-expression"neural mechanism framework for safety learning. This framework aims to systematically reveal the dynamic neural processes of safety learning and provide a new theoretical perspective for understanding the abnormal mechanisms in anxiety and related disorders. The following sections elaborate on the neural bases of these three stages and discuss how anxiety and related disorders manifest abnormalities at each stage.

3 A Three-Stage Neural Mechanism Framework for Safety Learning

Neurobiological research indicates that safety learning is a complex process involving multiple stages and coordinated neural systems. Although Laing et al. (2022a) distinguished acquisition and expression stages based on neurobiological differences in safety association formation and subsequent behavioral responses (see Section 2.3), thereby partially explaining the stage-specific characteristics of safety learning neural mechanisms, existing research shows that dopamine circuits related to positive emotion processing (Grill et al., 2024) and hippocampus-vmPFC circuits related to memory storage (Meyer et al., 2019) also participate in safety learning. In other words, the safety acquisition process involves not only threat prediction modification but also positive emotion and safety memory processing, and should therefore be viewed as a process that simultaneously integrates safety association formation and positive affective processing.

Furthermore, research demonstrates that individuals exhibit different neural response patterns to safety and threat cues during early stimulus processing stages. Event-related potential (ERP) studies show that N1 (approximately 100 ms post-stimulus) and P2 amplitude (approximately 200 ms post-stimulus) are primary components reflecting early perceptual processing (Favero et al., 2023): N1 primarily reflects primary perceptual processing, while P2 is associated with initial stimulus significance evaluation. Safety and threat stimuli show distinct amplitude patterns in these early ERP components. For example, under backward masking conditions (where a mask immediately follows stimulus presentation, preventing conscious awareness), threat cues elicit larger N1 amplitudes than safety cues (Mei et al., 2024). Dou et al. (2022) further demonstrated that stimulus social attributes modulate early safety perception processing, finding that partner' s voices as safety cues significantly reduced subjective threat perception and showed lower P2 amplitudes compared to stranger voices. These findings indicate that stimuli undergo differential processing at early stages, which may subsequently influence safety acquisition and expression.

Additionally, Grasser and Jovanovic (2021) noted, based on a review of neuroimaging studies, that individuals' ability to perceive and discriminate threat and safety cues is closely related to nervous system maturation, particularly the development of subcortical structures (e.g., amygdala, thalamus, basal ganglia) and their connections. During childhood, maturation of these subcortical structures and their connections supports more effective perception and discrimination of environmental threat and safety cues. Moreover, Calhoon and Tye (2015) proposed in their anxiety circuit model that threat and safety cue processing can be divided into early "detection" and "interpretation" stages, primarily relying on the thalamus, primary sensory cortex, insula, and amygdala to initially analyze and integrate stimulus salience and emotional meaning. Anxiety may cause abnormal activity in the neural circuits involved in these two stages.

Based on these insights, we further conceptualize the neural processing of safety learning into three stages: (1) The safety perception and appraisal stage involves detection and early processing of external stimuli, providing support for safety association formation and subsequent response regulation. Its neural basis relies on functional connectivity networks among the thalamus, cortex, insula, and amygdala to evaluate stimulus safety value; (2) The safety acquisition stage can be further subdivided into two sub-processes: safety association formation and maintenance, where individuals form stable associations between stimuli and no-threat outcomes through PE-driven learning, and positive affect acquisition, where repeated safety outcomes generate positive emotional value attribution to safety stimuli. At the neural level, the midbrain-striatum circuit plays a central role in supporting safety association formation and positive emotional experience generation, while long-term maintenance and consolidation of safety associations depend on hippocampal-prefrontal cortex circuits to lay the foundation for subsequent safety memory retrieval and behavioral expression; (3) The safety behavioral expression stage manifests as individuals' ability to effectively inhibit fear responses using safety signals after acquisition. At the neural level, this involves interactions among limbic, prefrontal, and temporal cortices that jointly support retrieval of safety memories across contexts to achieve effective fear inhibition. The following sections elaborate on the key brain regions and their relatively independent functional mechanisms in each of the three safety learning stages.

3.1 Safety Perception and Appraisal: The Cortex-Thalamus-Insula-Amygdala Circuit

Safety perception and appraisal is the process by which individuals judge whether an encountered stimulus is non-threatening. The thalamus-amygdala pathway plays an important role in this process. Early rodent studies found that fear acquisition primarily involves the medial geniculate body (MGB)-amygdala circuit (Romanski & LeDoux, 1992). Subsequently, Heldt and Falls (2006) examined whether this pathway affects safety signal processing in mice using a conditioned inhibition task. They found that lesions of the MGB or auditory cortex alone, which are important for fear conditioning, did not affect the inhibitory effect of safety signals on fear, but lesions of a larger auditory thalamic area (including MGB plus surrounding nuclei) disrupted safety signal effects. These results suggest that safety and threat information perception may rely on similar sensory pathways but differ in specific neural structures. Increasing evidence indicates that emotional processing does not depend on two fixed sensory pathways but involves widespread brain regions. The thalamus serves as a "relay station" for brain information, where almost all sensory information entering the cerebral cortex is processed and forwarded (Marcuse et al., 2025). Meanwhile, the thalamus may also participate in emotional processing through extensive interactions with the prefrontal cortex, insula, and amygdala (Pessoa, 2017). Therefore, safety-related information may be transmitted to the amygdala via the thalamus to form initial emotional

responses, but forming cognitive representations of safety may require refined processing and cognitive evaluation through the prefrontal cortex.

During safety learning, vmPFC is considered critical for safety signal perception and appraisal. vmPFC shows valence sensitivity when processing different types of emotional stimuli (positive or negative) and is regarded as a key region for integrating external stimuli with subjective emotional experiences to form emotional value representations (Hiser & Koenigs, 2018). A recent human multivoxel pattern analysis (MVPA) study (Tashjian et al., 2025) further distinguished vmPFC's role in threat versus safety assessment. Using a safety estimation task that simultaneously presented external threat cues (animals) and protective cues (weapons) that could inhibit threats, participants were asked to predict their safety status. Results showed significant vmPFC activation during assessment, with posterior vmPFC being more sensitive to threat information and anterior vmPFC more sensitive to protective information that could inhibit threats. Thus, safety judgments involve cognitive processing of stimulus meaning in prefrontal cortex, with anterior vmPFC potentially being a key region for subjective safety representation.

The insula may also participate in safety cue perception and appraisal. Foilb et al. (2021) trained rats to discriminate between danger and safety signals and found that Fos protein levels (reflecting neuronal activation) in the middle insular cortex (mIC) were significantly activated during safety-threat discrimination tasks, but did not significantly correlate with behavioral discrimination indices (ability to discriminate CS+ from CS-). This may indicate that mIC is not directly involved in regulating fear response intensity but participates in processing sensory input salience and contextual changes. Previous research suggests that the insula receives exteroceptive signals (pain, aversive stimuli), interoceptive signals (heartbeat, respiration), and cognitive-emotional inputs from prefrontal regions (e.g., vmPFC) and amygdala, integrating these different sources to generate bodily and environmentally relevant emotional experiences (Zhang et al., 2024). Therefore, the insula may serve as an information integration node in safety learning, combining external sensory cues with internal emotional state evaluations to support formation of safety-related perceptual experiences.

In summary, safety stimulus perception and appraisal likely depends on functional circuits among vmPFC, thalamus, insula, and amygdala (Figure 1 [Figure 1: see original paper]A), which may participate in integrating multimodal sensory information with cognitive evaluation to support safety judgments about external cues.

Figure 1. Three-stage neural mechanism circuits of safety learning. Note: In Figure 1(A), sensory cortex includes somatosensory cortex (postcentral gyrus in parietal lobe and precentral gyrus in frontal lobe), visual cortex (in occipital lobe), and auditory cortex (in temporal lobe). In Figure 1(B), the thalamus is medial to the striatum; midbrain-striatum involves two circuits: dorsal striatum-substantia nigra and ventral striatum-ventral tegmental area (both VTA and

substantia nigra are located in midbrain; striatum includes ventral and dorsal parts).

3.2.1 Safety Association Formation and Maintenance: Midbrain-Striatum Circuit and Hippocampus-Ventromedial Prefrontal Circuit

During the learning stage, individuals need to associate stimuli with safety and integrate them into memory networks. The acquisition-expression framework (Laing et al., 2022a) suggests that safety acquisition depends on two subcortical circuits related to associative learning: the DS-SNc circuit plays an important role in inhibitory prediction (CS→no US), while the VS-VTA circuit is more related to excitatory prediction (CS→US). However, this excitatory-inhibitory distinction may have limitations. For example, Salinas-Hernández et al. (2023) found that optogenetic inhibition of VTA projections to amNAc (located in VS) impaired mice' s ability to learn that "CS is now safe," resulting in impaired extinction learning. This suggests that VS is not solely involved in excitatory prediction but may also participate in encoding threat absence. A recent study provides new perspective on the division of labor between these two circuits in safety learning. Grill et al. (2024) used positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) in a human reversal learning task and found that the associative striatum (in DS) released dopamine at task rule reversal points. Stronger dopamine release was associated with higher sensitivity to "outcome deviating from expectation" and faster behavioral adjustment. In contrast, VS dopamine activity responded more to reward presence/absence without participating in rule reversal. Therefore, in safety learning, VS and DS likely reflect functional division of labor rather than simple inhibitory-excitatory opposition: DS is primarily involved in rule switching and behavioral strategy adjustment, while VS plays a key role in safety value formation and updating. Conditioned inhibition safety learning typically involves safety signals co-occurring with threat signals, thereby changing the outcome of threat cues, which is closer to rule/context switching and thus more DS-involved. Extinction and reversal primarily reflect re-evaluation and updating of threat cue value itself, making them more VS circuit-related.

After acquisition, safety associations are relayed through the thalamus to emotional memory storage circuits, namely the hippocampus-vmPFC circuit, integrating safety associations into individuals' cognitive structures. Rodent studies indicate that safety memory formation depends on prefrontal-hippocampal circuits (Kreutzmann, 2020; Lingawi et al., 2019). In this circuit, the hippocampus may participate in initial encoding of safety information, while the prefrontal cortex may be responsible for further consolidation and expression of safety memories. Tashjian et al.' s (2025) human MVPA study further suggests that anterior vmPFC may be a key region supporting long-term maintenance of safety memories. The study found that anterior vmPFC showed not only specific activation to safety stimuli during safety judgments but also maintained specific activation to safety cues after task completion, a phenomenon not ob-

served for fear cues. Additionally, Wiemer et al. (2024) found in human fMRI research that the hippocampus contributed to both threat and safety memories, while vmPFC specifically predicted safety memory formation and supported long-term maintenance through connections with hippocampus and amygdala. Moreover, a recent intracranial electroencephalography (iEEG) study (Pacheco-Estefan et al., 2025) using an ABC paradigm combined with representational similarity analysis (RSA) found that during extinction learning, amygdala theta oscillations (low-frequency rhythms generated by synchronized local neuronal activity) enhanced for "safety" stimuli in a specific time window (1.18-1.75 seconds before US presentation). This suggests the amygdala may participate in safety information encoding during extinction learning. Further analysis revealed that the stability of stimulus representation in the amygdala during extinction significantly correlated with that in the hippocampus, indicating that extinction memory formation may depend on coordinated activity between hippocampus and amygdala.

In summary, safety association formation primarily involves processing in midbrain-striatum circuits and hippocampus-amygdala-vmPFC circuits (Figure 1Ba). The former is activated during safety signal acquisition, providing important input for subsequent information processing, while the latter is more involved in safety memory consolidation and long-term storage.

3.2.2 Acquisition of Positive Affect: Ventral Tegmental Area-Ventral Striatum-Prefrontal Cortex Circuit

During safety association formation, individuals typically simultaneously establish positive affective representations of safety signals. Neurobiological research indicates that the VTA-VS circuit is commonly considered involved in reward behavior (Jeong et al., 2022; Morales & Margolis, 2017). However, recent studies show that the VTA-VS circuit is also important for threat and safety-related learning (Cho et al., 2021; Ko et al., 2023; Luo et al., 2018). Laing et al. (2022b) collected subjective emotional ratings of safety signals in human fMRI research, finding that safety signals were subjectively rated as more positive and that early safety signal processing was accompanied by specific VTA activation. Thiele et al. (2021) found that during human fear extinction, learning that "US did not occur (threat absence)" generated reward-like prediction error signals in VS. Therefore, VTA and VS activation during fear extinction and safety signal processing may be related to safety value updating and positive subjective evaluation of safety signals.

VTA-VS circuit dopamine activity is typically considered regulated by prefrontal cortex, primarily involving OFC and vmPFC. Atlas et al. (2016) used a reversal fear conditioning experiment where participants learned correspondences between two faces and electric shocks with multiple reversals. Results showed that OFC activity updated immediately upon receiving reversal instructions, with OFC being relatively more active under safety expectations. Anatomical evidence indicates that OFC projects to VS, which partially influences mid-

brain dopamine neuron activity via the habenula, thereby participating in regulation of reward prediction error signals important for emotion, motivation, and decision-making (Rolls et al., 2020; Rolls, 2023). Therefore, OFC may primarily participate in safety value representation and coordinate with the VTA-VS circuit to help individuals update safety-related value information.

Numerous studies also show that vmPFC is highly sensitive to stimulus emotional valence during emotional processing, with activation levels increasing significantly with positive stimulus valence and subjective pleasantness (Bezmaternykh et al., 2025; Suzuki & Tanaka, 2021; Winecoff et al., 2013), possibly reflecting vmPFC' s role in encoding emotional experiences into subjective value signals. In safety learning, vmPFC and VS activation positively correlates with participants' subjective positive ratings of safety signals (Harrison et al., 2017; Savage et al., 2021; Tashjian et al., 2025). Therefore, vmPFC may participate in safety signal value encoding through interaction with the dopamine system, supporting formation and maintenance of safety representations.

Overall, acquisition of positive affective experiences in safety learning may be related to dopamine projection activity in the VTA-VS-prefrontal cortex (OFC/vmPFC) circuit (Figure 1Bb). This circuit' s activity encourages individuals to treat safety signals as reward-like reinforcers, generating approach motivation and positive feedback toward safety stimuli.

3.3 Safety Behavioral Expression: Hippocampus-Centered Fear Inhibition Circuit

Acquired safety memories can be retrieved when individuals encounter contexts where safety stimuli are present, helping them inhibit fear responses. This process may involve anterior hippocampus (aHC)-vmPFC-amygdala (Figure 1Ca) and aHC-dACC (Figure 1Cb) circuits.

Current research indicates that functional connectivity between hippocampus and prefrontal-amygdala circuits forms the neural basis of extinction memory (Anderson & Floresco, 2022; Kredlow et al., 2022). Milad and Quirk (2012) noted that in this circuit, the amygdala generates and inhibits fear responses, the hippocampus provides contextual information determining whether extinction memories can be retrieved, and vmPFC activates during extinction recall to inhibit fear through top-down regulation of the amygdala. Their coordinated action ensures flexible use of safety information to inhibit fear across contexts. Compared to traditional univariate methods, MVPA can more sensitively identify representational changes and memory reinstatement in key regions like limbic system and vmPFC after extinction learning, providing finer-grained evidence for understanding human extinction recall mechanisms. For example, Hennings et al. (2022) used encoding-retrieval similarity (ERS) analysis to tag neural representations formed during fear acquisition and extinction and tested their reactivation 24 hours later in memory tests. Results showed that both vmPFC and aHC displayed neural reinstatement of extinction memories 24

hours after fear extinction, suggesting these regions may be key areas supporting safety memory recall after extinction learning. Bauer et al. (2025) further used ERS to compare neural reinstatement after standard extinction versus counterconditioning extinction (where CS+ is paired with positive outcomes during extinction). Twenty-four hours after extinction learning, individuals showed significant reinstatement of neural representations for counterconditioned safety cues in vmPFC, an effect not observed for standard extinction or threat stimuli. Together, these findings suggest vmPFC may be a key region supporting extinction recall, with safety memories carrying stronger positive value being more easily reactivated in vmPFC.

Furthermore, safety learning based on the conditioned inhibition paradigm shows distinct neural circuit activation patterns compared to extinction learning. A series of safety signal learning studies (Meyer et al., 2019; Odriozola & Gee, 2021; Odriozola et al., 2024) found that when safety compound stimuli were presented (safety signal and threat stimulus co-occurring), individuals' fear responses—including skin conductance responses (SCR) or US expectancy ratings—were significantly reduced compared to threat stimulus alone. Brain imaging results further showed that aHC-dACC circuit functional connectivity was significantly enhanced during safety compound stimulus presentation, while the traditional extinction pathway (vmPFC-amygdala) showed no significant differences. Given that dACC activity is typically considered to evaluate current environmental threat levels (Holtz et al., 2012; Hur et al., 2020), this suggests that during safety signal learning, coordinated activity between anterior hippocampus and dACC may support safety memory retrieval to modulate threat representations and reduce fear responses.

Notably, these two fear inhibition mechanisms show different developmental trajectories. A recent study comparing fear extinction retention across children, adolescents, and adults found that adults showed better behavioral fear inhibition accompanied by stronger vmPFC activation and more stable vmPFC-amygdala and vmPFC-hippocampus functional connectivity. In contrast, children and adolescents lacked such mature circuit support and were more prone to fear response recovery (Widegren et al., 2025). This suggests that extinction-based safety learning ability depends on circuit maturation, with children and adolescents lacking mature vmPFC regulation support and being more vulnerable to fear recovery. In contrast, conditioned inhibition-based safety signal learning may become the primary fear inhibition pathway during adolescence. Pattwell et al. (2016) found that "context + safety cue" combinations effectively attenuated fear memories in mice, with inhibitory effects persisting from adolescence (postnatal day 30) to adulthood (postnatal day 60). This result was associated with rapid dendritic spine proliferation in mouse adolescent PL and transient enhancement of projections from ventral hippocampus CA1 (homologous to human aHC) to PL. Kribakaran et al. (2022, 2024) found in human safety signal learning tasks that both adults and children could effectively attenuate fear responses through safety signal learning at the behavioral level. Further mediation analysis showed that in younger adolescents (<11.45 years),

trauma exposure was associated with stronger aHC-subgenual anterior cingulate cortex (sgACC) connectivity, whereas older adolescents (>16.89 years) showed the opposite pattern—trauma exposure was associated with weaker aHC-sgACC functional connectivity. Moreover, weakened connectivity in this circuit further predicted PTSD symptom severity (Kribakaran et al., 2024). These results suggest that adolescents with trauma history may show transient compensatory enhancement of aHC-sgACC functional connectivity early on, but this compensation gradually diminishes with developmental progression.

In summary, safety behavioral expression—manifested as successful retrieval of acquired safety memories to inhibit fear responses—likely involves coordinated activity among prefrontal cortex, sensory cortex, and limbic system. Specifically, extinction-based safety learning depends on maturation of the aHC-vmPFC-amygdala circuit and is more stable in adults, while inhibitory safety learning relies more on the aHC-dACC circuit and plays an important role during childhood and adolescence. These two circuits jointly support effective retrieval of safety information to inhibit fear in appropriate contexts.

4.1 Threat Perception Amplification and Cognitive Regulation Failure

Individuals with anxiety often show inability to effectively identify safety signals during safety learning (Haaker et al., 2015; Wong & Lovibond, 2021). This deficit may stem from structural and functional connectivity abnormalities in perception and appraisal circuits (Figure 2 [Figure 2: see original paper]A), leading to amplified threat perception and cognitive regulation failure.

Existing research finds that anxious individuals show significantly enhanced insula-amygdala connectivity (Brühl et al., 2014; Nicolas et al., 2023; Sehlmeyer et al., 2011), leading to stronger emotional experiences. A recent neuroimaging study (Langhammer et al., 2025) showed that across anxiety disorders—including panic disorder/agoraphobia, SAD, and specific phobia—patients exhibited enhanced insula-thalamus functional connectivity compared to healthy individuals, reflecting increased sensitivity to internal bodily sensations. Dong et al. (2025) further demonstrated through functional gradient analysis that GAD patients showed significantly higher functional gradient values between lateral thalamus and sensory cortex compared to healthy individuals. This gradient disruption reflects diminished thalamic filtering of sensory input, causing even weak external stimuli to be amplified by sensory cortex. These neural circuit abnormalities may be related to excessive threat perception and amplified fear responses in anxious individuals, making them more likely to interpret ambiguous or neutral stimuli as potential threats. Additionally, anxious individuals show functional abnormalities in vmPFC-insula circuits. For example, GAD patients exhibit reduced functional connectivity between vmPFC and posterior-mid insula compared to healthy individuals, with connectivity strength negatively correlating with anxiety sensitivity (Steinhäuser et al., 2023). This suggests GAD patients may have deficits in top-down cognitive control, making it difficult to effectively regulate bottom-up emotional signals.

In summary, existing research suggests that anxious individuals have functional abnormalities in neural circuits related to perception and appraisal, which may be associated with heightened sensitivity to internal and external signals and weaker top-down regulation, thereby affecting their ability to discriminate between safety and threat signals.

4.2 Sluggish Safety Learning and Weakened Positive Feedback

Inefficient safety learning is a characteristic feature of anxiety disorders, with neurobiological manifestations primarily involving dopamine system and striatal dysfunction (Figure 2B), leading to difficulty in effectively using PE mechanisms to establish safety associations that inhibit fear across learning types.

In fear extinction learning, dopamine activity in the VTA-nucleus accumbens (NAc) pathway is abnormal in anxious patients. Normally, when expected threats do not occur (e.g., CS+ without shock), the brain generates negative PE (contrary to expectation) to update the cognition that "this stimulus is safe." However, anxious individuals show significantly reduced or absent neural responses to this PE signal, leading to extinction difficulties (Papalini et al., 2020). Additionally, reduced functional connectivity between VS and prefrontal regions (e.g., vmPFC) in anxious patients may hinder safety value updating and subsequent safety memory formation (Wen et al., 2022). In conditioned inhibition learning, low safety learning efficiency stems from abnormal PE encoding of safety signals in DS, preventing effective use of safety signals to inhibit threat. Laing et al. (2021) found that high trait-anxious individuals showed higher threat expectancy for conditioned inhibition safety signals (X-) during the learning stage. Subsequent neuroimaging research revealed that learning safety signals was associated with DS-SN circuit activation (Laing et al., 2022b). However, this study used healthy adults, and whether anxious individuals show abnormalities in this circuit requires direct verification in clinical or high-anxiety samples. Thus, deficits in safety acquisition mechanisms in anxious individuals may stem from abnormal striatal functional connectivity, affecting PE encoding processes. However, human conditioned inhibition safety learning studies currently lack direct evidence of circuit abnormalities in anxious individuals, representing an important direction for future research.

Moreover, anxious individuals often struggle to obtain positive feedback from safety signals. Their reward system is suppressed, manifesting as abnormal dopamine signaling in vmPFC and weakened functional coupling between vmPFC and reward systems (e.g., VS), making it difficult to extract positive feedback signals from safe contexts (Plas et al., 2024; Sartori et al., 2024; Whittaker et al., 2018). This may cause anxious patients in exposure therapy to fail to obtain sufficient safety feedback even after multiple exposures due to blunted reward systems, thereby preventing fear extinction. For example, insufficient dopamine signaling in SAD may reduce pleasure from social interactions (e.g., positive feedback from being accepted), exacerbating sensitivity to negative evaluation and avoidance tendencies in social contexts (Nusslock &

Alloy, 2017; Carlton et al., 2020). Additionally, Sequeira et al. (2021) showed that activation levels in striatal (particularly NAc) dopaminergic reward circuits are closely related to safety learning effectiveness in anxious individuals. Higher reward sensitivity may enhance motivation and positive emotional experiences during exposure exercises and increase responsiveness to social rewards such as therapist feedback, thereby strengthening safety learning and threat re-evaluation. Conversely, individuals with lower reward responsiveness may struggle to obtain positive signals from successful exposure experiences, leading to inadequate safety learning and ultimately affecting treatment efficacy. Overall, anxiety is closely associated with functional abnormalities in dopamine circuits centered on the striatum. These circuits play key roles in affective valence and reward value evaluation, and their dysfunction may impair encoding of positive subjective value of safety signals, reducing the ability to obtain positive feedback from safety signals and thereby affecting safety association formation.

4.3 Abnormal Safety Behavioral Expression

Another characteristic feature of anxious individuals is the inability to effectively update safety memories after learning, which is closely related to structural and functional abnormalities in hippocampus-centered fear inhibition circuits (Figure 2C).

The prefrontal-amygdala circuit is widely considered to participate in extinction memory integration and fear inhibition (Gunther et al., 2022; Kenwood et al., 2022). A recent meta-analysis (Kausche et al., 2025) showed that individuals with anxiety disorders have impaired long-term discrimination and memory regulation of threat and safety information. Compared to healthy individuals, they subjectively perceive safety stimuli as more threatening and show physiological fear responses (e.g., startle reflex, skin conductance) to safety signals comparable to threat stimuli. This abnormality is closely related to neurobiological mechanisms of safety memory. Under normal conditions, the hippocampus indirectly regulates prefrontal inhibition of amygdala hyperactivation by retrieving safety memories. However, hippocampal dysfunction caused by anxiety may hinder execution of prefrontal (especially vmPFC) inhibitory functions, preventing effective fear suppression and affecting safety memory expression (Lebois et al., 2019).

Additionally, according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (American Psychiatric Association, 2022), OCD and PTSD have high clinical comorbidity with anxiety disorders. Related neuroimaging studies show these mental disorders are also accompanied by functional abnormalities in hippocampus-centered neural circuits and brain regions. For example, adolescents with PTSD show obvious developmental deficits in safety memory neural circuits, including decreased hippocampal volume and weakened amygdala-prefrontal coupling with age (Herringa, 2017). Similarly, OCD patients show significantly reduced hippocampal volume and weakened

functional connectivity between hippocampus and prefrontal cortex/amygdala compared to healthy controls (Boedhoe et al., 2017; Cooper & Dunsmoor, 2021). This may cause OCD patients to reinforce automatic execution of stereotyped behaviors (e.g., repeatedly checking doors and windows despite clear memory evidence of having closed them) because hippocampal abnormalities prevent integration of "confirmed safety" signals into current contexts. Wen et al. (2022) further confirmed in an fMRI study of 338 participants (including healthy controls, anxiety-related disorder patients, and PTSD patients) that abnormalities in hippocampus-centered circuits are closely related to anxiety-related disorders. Healthy individuals showed significantly enhanced functional connectivity between vmPFC and key nodes including hippocampus and amygdala during late fear extinction learning, supporting safety memory formation. Patient groups lacked this dynamic enhancement, with some showing decreased connectivity during late extinction that correlated with symptom severity.

Furthermore, anxiety may also impair individuals'ability to use contextual safety signals to inhibit fear by damaging the hippocampus-dACC circuit. Odriozola et al. (2024) found that high trait-anxious individuals required stronger hippocampal activity to achieve equivalent inhibitory effects from safety signals, and showed significantly decreased hippocampus-dACC functional connectivity during late safety signal learning stages. This suggests high-anxious individuals have inefficient safety memory retrieval and difficulty stably maintaining safety associations formed through safety signals, thereby weakening fear inhibition capacity. However, current research on safety signal learning remains relatively scarce in clinical anxiety disorder populations. Future studies should further verify these findings in clinical anxiety disorder populations to clarify whether these neural features can serve as potential biomarkers for identifying and evaluating anxiety and related disorders.

Overall, integrating findings from different anxiety disorders and closely related psychiatric conditions such as PTSD and OCD reveals that persistent fear generalization and weakened behavioral inhibition may be related to functional dysregulation of neural circuits among hippocampus, prefrontal cortex, amygdala, and anterior cingulate cortex. This may affect individuals' ability to retrieve safety memories across different contexts, making it difficult to use learned "safety" to inhibit fear.

5 Summary and Outlook

Safety learning serves as a critical adaptive mechanism for individuals to identify safety signals and suppress fear responses, playing an important role in environmental adaptation and healthy development. By integrating existing evidence, this study proposes a three-stage neural mechanism framework for safety learning encompassing perception and appraisal, safety acquisition, and behavioral expression. This framework not only deepens understanding of the dynamic neural processes underlying safety learning but, more importantly, provides a more refined and explanatory pathological model for characterizing safety learning fea-

tures in anxiety and related disorders. The framework posits that mechanistic abnormalities in anxious patients may originate from amplified threat perception during the perception stage, abnormal PE coding and weakened positive feedback during the acquisition stage, and impaired safety memory retrieval during the expression stage. This stage-specific perspective provides new theoretical reference and potential intervention targets for advancing clinical translation of safety learning research.

Future research can be advanced in the following directions:

5.1 Strengthening Application of Conditioned Inhibition Paradigms in Safety Learning Research on Anxious Individuals

In recent years, the conditioned inhibition paradigm has been recognized for its unique advantages in representing safety learning, as it avoids competition between threat and safety memories and facilitates stable safety association formation (Laing & Harrison, 2021; Laing et al., 2024). However, conditioned inhibition safety learning research in clinical populations remains lacking. Additionally, most studies use low-semantic stimuli such as geometric shapes, which poorly reflect complex reactions to safety cues in real-world social contexts (Dunsmoor & Murphy, 2015), limiting ecological validity and generalizability. Future research should introduce natural stimuli with semantic or social attributes combined with MVPA methods to reveal neural representation differences of inhibitory safety signals between clinical anxiety patients and healthy individuals.

On the other hand, anxious individuals typically show abnormalities during safety memory recall stages. However, current "safety memory" recall research remains focused primarily on fear extinction tasks. In contrast, it remains unclear whether anxious individuals also show functional impairments during recall of conditioned inhibition, a more active and stable form of safety learning, or whether this mechanism might compensate for deficits in extinction recall—questions requiring future investigation. Moreover, limited research on conditioned inhibition recall may be related to paradigm complexity: since inhibitory safety signals require co-occurrence with threat stimuli to exert inhibitory effects, independently measuring their memory representations during recall presents greater challenges. A recent neurobiological study (Tashjian et al., 2025) provides insights for exploring recall mechanisms of inhibitory safety learning. The study presented protective and threat cues separately, had individuals pair them and imagine "combat scenarios" to assess threat levels, and then presented stimuli alone again after the task to measure neural representations of different cue types. Future research could improve upon this paradigm and use RSA methods to compare neural representation differences between anxious and healthy individuals during conditioned inhibition safety learning recall stages, thereby further revealing neurobiological mechanisms underlying recall stage abnormalities in anxious individuals.

5.2 Enhancing Positive Affective Experience in Safety Learning for Anxious Individuals

During safety learning, establishment of "positive affect" is typically considered an important indicator of successful safety learning. However, anxious individuals may struggle to obtain sufficient positive emotional experiences from safety learning, which may weaken safety learning effectiveness. Future anxiety intervention strategies based on safety learning should place greater emphasis on enhancing positive emotional experiences. For example, introducing rewards or other positive emotional experiences as safety learning outcomes, rather than simple threat absence, could strengthen positive value encoding of safety cues.

Additionally, current measurement of positive affect acquisition in safety learning remains overly dependent on subjective reports, which suffer from strong subjectivity and limited sensitivity (Laing et al., 2024; Odriozola & Gee, 2021). Future research should construct multimodal, objective safety assessment systems with ecological validity. For example, simultaneously collecting fMRI, EEG, and skin conductance neurophysiological signals combined with subjective assessments to identify and predict neuralphysiological features of individuals' "safety" experiences could support precise evaluation of intervention effects and further promote development of safety learning-based anxiety intervention strategies.

5.3 Developing Neurobiologically-Guided Stage-Specific and Developmentally-Informed Clinical Intervention Strategies

Based on this framework, future clinical intervention strategies should follow two principles: stage-specificity and developmental sensitivity. First, from a neurobiological perspective, safety learning does not rely on a single fear inhibition circuit but is a dynamic multi-stage process involving perception and appraisal, associative feedback, and memory expression. Anxiety and related mental disorders may originate from neuroregulatory abnormalities at different stages, necessitating attention to multi-regional coordinated mechanisms including insula, OFC, and striatum. Therefore, future research must further validate the three-stage neural processes of safety learning in clinical populations to provide research support for developing stage-specific, targeted neuromodulation strategies and differential treatment plans.

Second, intervention strategies must fully consider individual neurodevelopmental trajectories. Research shows that the fear inhibition circuit underlying safety signal learning—the hippocampus-anterior cingulate cortex circuit—is well-developed during adolescence, whereas the traditional fear extinction-dependent vmPFC-amygdala circuit remains immature during adolescence, potentially making adolescents more responsive to safety signal learning (Kribakaran et al., 2024; Odriozola & Gee, 2021). This developmental difference suggests that future intervention designs should adopt stage-specific, hierarchical treatment strategies, matching appropriate neurobiological intervention foci

according to developmental stages to optimize treatment effects and enhance the translational value of safety learning mechanisms in clinical practice.

References

曹杨婧文, 李俊娇, 陈伟, 杨勇, 胡琰健, 郑希付. (2019). 条件性恐惧记忆消退的提取干预范式及其作用的神经机制. 心理科学进展, 27(2), 268-277.

雷怡, 梅颖, 张文海, 李红. (2018). 基于知觉的恐惧泛化的认知神经机制. 心理科学进展, 26(8),

申思怡, 梅颖, 王金霞, 戴雨芊, 吴奇, 雷怡. (2023). 条件学习视角下的恐惧与厌恶. 心理科学, 46(4), 825-832.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). Author.

Anderson, M. C., & Floresco, S. B. (2022). Prefrontal-hippocampal interactions supporting the extinction of emotional memories: The retrieval stopping model. *Neuropsychopharmacology, 47*(1), 180-195.

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings. Biological Sciences, 277*(1692), 2411-2416.

Asede, D., Doddapaneni, D., & Bolton, M. M. (2022). Amygdala Intercalated Cells: Gate Keepers and Conveyors of Internal State to the Circuits of Emotion. *The Journal of Neuroscience, 42*(49), 9098-9109.

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016). Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *eLife, 5*, e15192.

Badarnee, M., Wen, Z., Hammoud, M. Z., Glimcher, P., Cain, C. K., & Milad, M. R. (2025). Intersect between brain mechanisms of conditioned threat, active avoidance, and reward. *Communications Psychology, 3*(1), 32.

Bauer, E. A., Cooper, S. E., Keller, N. E., Cisler, J. M., & Dunsmoor, J. E. (2025). Encoding-Retrieval Similarity Reveals Distinct Neural Reinstatement of Safety Memories Following Counterconditioning in Posttraumatic Stress Disorder. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 10*(11), 1125-1133.

Bezmaternykh, D. D., Melnikov, M. Ye., Petrovskiy, E. D., Mazhirina, K. G., Savelov, A. A., Shtark, M. B., Vuilleumier, P., & Koush, Y. (2025). Attenuation processes in positive social emotion upregulation: Disentangling functional role of ventrolateral prefrontal cortex. *iScience, 28*(2), Boedhoe, P. S. W., Schmaal, L., Abe, Y., Ameis, S. H., Arnold, P. D., Batistuzzo, M. C., Benedetti, F., Beucke, J. C., Bollettini, I., Bose, A., Brem, S., Calvo, A., Cheng, Y., Cho, K. I. K., Dallaspezia, S., Denys, D., Fitzgerald, K. D., Fouche, J.-P., Giménez, M., ···van den Heuvel, O. A. (2017). Distinct Subcortical Volume Alterations in Pediatric and Adult OCD: A Worldwide Meta- and Mega-Analysis. *The American Journal of Psychiatry, 174*(1), 60-69.

Brühl, A. B., Delsignore, A., Komossa, K., & Weidt, S. (2014). Neuroimaging in social anxiety disorder—A meta-analytic review resulting in a new neurofunctional model. *Neuroscience and Biobehavioral Reviews, 47*, 260–280.

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

Carlton, C. N., Sullivan-Toole, H., Ghane, M., & Richey, J. A. (2020). Reward Circuitry and Motivational Deficits in Social Anxiety Disorder: What Can Be Learned From Mouse Models? *Frontiers in Neuroscience, 14*, 154.

Cho, H., Likhtik, E., & Dennis-Tiwary, T. A. (2021). Absence Makes the Mind Grow Fonder: Reconceptualizing Studies of Safety Learning in Translational Research on Anxiety. *Cognitive, Affective & Behavioral Neuroscience, 21*(1), 1–13.

Christianson, J. P., Fernando, A. B. P., Kazama, A. M., Jovanovic, T., Ostroff, L. E., & Sangha, S. (2012). Inhibition of Fear by Learned Safety Signals: A Mini-Symposium Review. *The Journal of Neuroscience, 32*(41), 14118–14124.

Cooper, S. E., & Dunsmoor, J. E. (2021). Fear conditioning and extinction in obsessive-compulsive disorder: A systematic review. *Neuroscience and Biobehavioral Reviews, 129*, 75–94.

Cooper, S. E., Grillon, C., & Lissek, S. (2018). Impaired discriminative fear conditioning during later training trials differentiates generalized anxiety disorder, but not panic disorder, from healthy control participants. *Comprehensive Psychiatry, 85*, 84–93.

Deng, J., Fang, W., Gong, Y., Bao, Y., Li, H., Su, S., Sun, J., Shi, J., Lu, L., Shi, L., & Sun, H. (2021). Augmentation of fear extinction by theta-burst transcranial magnetic stimulation of the prefrontal cortex in humans. *Journal of Psychiatry & Neuroscience: JPN, 46*(2), E292–E302.

Dong, Q., Li, X., Zhang, Q., Ju, Y., Liao, M., Zhu, J., Li, R., Yao, Z., Zhang, Y., Hu, B., & Zheng, W. (2025). Aberrant functional gradient of thalamo-cortical circuitry in major depressive disorder and generalized anxiety disorder. *Journal of Affective Disorders, 376*, 473–486.

Dou, H., Dai, Y., Qiu, Y., & Lei, Y. (2022). Attachment voices promote safety learning in humans: A critical role for P2. *Psychophysiology, 59*(6), e13997.

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, Concepts, and Conditioning: How Humans Generalize Fear. *Trends in Cognitive Sciences, 19*(2), 73–77.

Duvarci, S. (2024). Dopaminergic circuits controlling threat and safety learning. *Trends in Neurosciences, 47*(12), 1014–1027.

Favero, J. D., Luck, C., Lipp, O. V., Nguyen, A. T., & Marinovic, W. (2023). N1-P2 event-related potentials and perceived intensity are associated: The effects of a weak pre-stimulus and attentional load on processing of a subsequent intense

stimulus. *Biological Psychology, 184*, Fisher, C. T. L., & Urcelay, G. P. (2024). Safety signals reinforce instrumental avoidance in humans. *Learning & Memory, 31*(8), a053914.

Foilb, A. R., Sansaricq, G. N., Zona, E. E., Fernando, K., & Christianson, J. P. (2021). Neural Correlates of Safety Learning. *Behavioural Brain Research, 396*, 112884.

Graham, B. M., & Milad, M. R. (2011). The Study of Fear Extinction: Implications for Anxiety Disorders. *American Journal of Psychiatry, 168*(12), 1255–1265.

Grasser, L. R., & Jovanovic, T. (2021). Safety learning during development: Implications for development of psychopathology. *Behavioural Brain Research, 408*, 113297.

Grill, F., Guitart-Masip, M., Johansson, J., Stiernman, L., Axelsson, J., Nyberg, L., & Rieckmann, A. (2024). Dopamine release in human associative striatum during reversal learning. *Nature Communications, 15*(1), 59.

Gründahl, M., Retzlaff, L., Herrmann, M. J., Hein, G., & Andreatta, M. (2022). The skin conductance response indicating pain relief is independent of self or social influence on pain. *Psychophysiology, 59*(3), e13978.

Gunther, K. E., Petrie, D., Pearce, A. L., Fuchs, B. A., Pérez-Edgar, K., Keller, K. L., & Geier, C. (2022). Heterogeneity in PFC-amygdala connectivity in middle childhood, and concurrent interrelations with inhibitory control and anxiety symptoms. *Neuropsychologia, 174*, 108313.

Haaker, J., Lonsdorf, T. B., Schümann, D., Menz, M., Brassen, S., Bunzeck, N., Gamer, M., & Kalisch, R. (2015). Deficient inhibitory processing in trait anxiety: Evidence from context-dependent fear learning, extinction recall and renewal. *Biological Psychology, 111*, 65–72.

Harb, M., Jagusch, J., Durairaja, A., Endres, T., Leßmann, V., & Fendt, M. (2021). BDNF haploinsufficiency induces behavioral endophenotypes of schizophrenia in male mice that are rescued by enriched environment. *Translational Psychiatry, 11*(1), 233.

Harrison, B. J., Fullana, M. A., Via, E., Soriano-Mas, C., Vervliet, B., Martínez-Zalacaín, I., Pujol, J., Davey, C. G., Kircher, T., Straube, B., & Cardoner, N. (2017). Human ventromedial prefrontal cortex and the positive affective processing of safety signals. *NeuroImage, 152*, 12–Heldt, S. A., & Falls, W. A. (2006). Posttraining lesions of the auditory thalamus, but not cortex, disrupt the inhibition of fear conditioned to an auditory stimulus. *The European Journal of Neuroscience, 23*(3), 765–779.

Hennings, A. C., Mason McClay, Drew, M. R., Lewis-Peacock, J. A., & Dunsmoor, J. E. (2022). Neural reinstatement reveals divided organization of fear and extinction memories in the human brain. *Current Biology, 32*(2), 304-314.e5.

Herringa, R. J. (2017). Trauma, PTSD, and the Developing Brain. *Current Psychiatry Reports, 19*(10), 69.

Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biological Psychiatry, 83*(8), 638–647.

Holland, P. C. (1989). Transfer of negative occasion setting and conditioned inhibition across conditioned and unconditioned stimuli. *Journal of Experimental Psychology. Animal Behavior Processes, 15*(4), 311–328.

Holland, P. C., & Lamarre, J. (1984). Transfer of inhibition after serial and simultaneous feature negative discrimination training. *Learning and Motivation, 15*(3), 219–243.

Holtz, K., Pané-Farré, C. A., Wendt, J., Lotze, M., & Hamm, A. O. (2012). Brain activation during anticipation of interoceptive threat. *NeuroImage, 61*(4), 857–865.

Hur, J., Smith, J. F., DeYoung, K. A., Anderson, A. S., Kuang, J., Kim, H. C., Tillman, R. M., Kuhn, M., Fox, A. S., & Shackman, A. J. (2020). Anxiety and the Neurobiology of Temporally Uncertain Threat Anticipation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 40*(41), 7949–7964.

Jeong, H., Taylor, A., Floeder, J. R., Lohmann, M., Mihalas, S., Wu, B., Zhou, M., Burke, D. A., & Namboodiri, V. M. K. (2022). Mesolimbic dopamine release conveys causal associations. *Science (New York, N.Y.), 378*(6626), eabq6740.

Jovanovic, T., Kazama, A., Bachevalier, J., & Davis, M. (2012). Impaired Safety Signal Learning May be a Biomarker of PTSD. *Neuropharmacology, 62*(2), 695–704.

Kausche, F. M., Carsten, H. P., Sobania, K. M., & Riesel, A. (2025). Fear and safety learning in anxiety- and stress-related disorders: An updated meta-analysis. *Neuroscience & Biobehavioral Reviews, 169*, 105983.

Kenwood, M. M., Kalin, N. H., & Barbas, H. (2022). The prefrontal cortex, pathological anxiety, and anxiety disorders. *Neuropsychopharmacology, 47*(1), 260–275.

Kleine, R. A. D., Hutschemaekers, M. H. M., Hendriks, G. J., Kampman, M., Papalini, S., Minnen, A. V., & Vervliet, B. (2023). Impaired action-safety learning and excessive relief during avoidance in patients with anxiety disorders. *Journal of Anxiety Disorders, 96*, 102698.

Ko, B., Yoo, J.-Y., Yoo, T., Choi, W., Dogan, R., Sung, K., Um, D., Lee, S. B., Kim, H. J., Lee, S., Beak, S. T., Park, S. K., Paik, S.-B., Kim, T.-K., & Kim, J.-H. (2023). Npas4-mediated dopaminergic regulation

Note: Figure translations are in progress. See original paper for figures.

ChinaRxiv 中X

Source: ChinaXiv —Machine translation. Verify with original.