HHS Public Access

Author manuscript

Trends Cogn Sci. Author manuscript; available in PMC 2022 May 01.

Published in final edited form as:

Trends Cogn Sci. 2021 May ; 25(5): 342–354. doi:10.1016/j.tics.2021.01.013.

A Decision Architecture for Safety Computations

Sarah M. Tashjian[a,*], **Tomislav D. Zbozinek**[a], **Dean Mobbs**[a,b]

[a]Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[b]Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA.

Abstract

Accurately estimating safety is critical to pursuing non-defensive survival behaviors. However, little attention has been paid to how the human brain computes safety. We conceptualize a model that consists of two components: (1) threat-oriented evaluations that focus on threat value, imminence, and predictability, and (2) self-oriented evaluations that focus on the agent's experience, strategies, and ability to control the situation. Our model points to the dynamic interaction between these two components as a mechanism of safety estimation. Based on a growing body of human literature, we hypothesize distinct regions of the ventromedial prefrontal cortex respond to threat and safety to facilitate survival decisions. We suggest safety is not an inverse of danger, but reflects independent computations that mediate defensive circuits and behaviors.

Keywords

decision making; safety; threat; ventromedial prefrontal cortex

Safety as a distinct computation

Our daily lives are faced with many potential dangers, yet for healthy psychological functioning, we must determine when a **threat** is real versus when we are safe. The potential for danger requires humans to make complex decisions with the goal of promoting survival by acquiring safety. Low-threshold defensive responses (e.g., reactive flight, startle) are important for survival, but safety decisions are critical for satisfying other necessary behaviors. For example, behavioral ecologists have long shown that decreased perceptions of danger result in increased feeding, decreased energy consumption, and increased mating in non-human animals[1]. Safety determinations are important not only for appropriate allocation of resources, but also for understanding threat-related psychopathology, which often involves deficient safety processing (Box 1). Prior threat-focused work examines safety as a linear inverse of danger. Although this is sometimes the case, we argue that safety

*Corresponding author Sarah M. Tashjian (smtashji@caltech.edu), Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, MC 228-77, Pasadena, CA 91125, USA.

can also result from computations by an independent neural system that mediates functioning of canonical **defensive survival circuits**.

In this article, we address how humans determine and respond to safety from a decision neuroscience framework. We propose that integrative weighting of lower-level features (e.g., stimuli value, perceived control) results in an estimate of survival capability that signals safety via the brain's anterior ventromedial prefrontal cortex (vmPFC). Accumulating evidence suggests that across species, the vmPFC plays a role in signaling both danger and safety[2,3]. We further posit that the posterior vmPFC prepares organisms for anticipated danger, but that the anterior vmPFC, a region with distinct functional connectivity from its posterior counterpart, cues safety both in the presence and absence of danger. A key feature of our model is that calculations in the vmPFC alter how stimuli are interpreted in core regions of the brain's defensive circuitry to facilitate behavioral flexibility. This proposal corresponds with rodent models demonstrating the infralimbic (IL) and prelimbic (PL) cortex exert control over fear expression and inhibition in the amygdala as distinct prefrontal systems[4].

Defining safety

We define safety in terms of the perception of present and prospective survival likelihood. Safety exists on a continuum from zero to complete safety. At the zero safety extreme, for example, during predatory attack, impending harm is unavoidable. At the other extreme of complete safety, we identify 'contextual safety', which refers to the absence of threat of an adverse outcome. Contextual safety frequently occurs in a nest or shelter where there is no present or prospective danger. Complete safety represents a healthy organism's 'default' state. Safety can also exist in the presence of a dangerous stimulus. Learning that a threat has become safe is termed 'extinction-based safety'. Extinction can occur via a 'safety signal' or 'prospective safety'. When threat is accompanied by a mitigating factor, we call these mitigating factors 'safety signals'. For example, being near a lion is dangerous, but the danger is mitigated if that lion is caged. The cage thus acts as a safety signal. 'Prospective safety' can occur in the presence or absence of danger, and refers to the expectation that one's behavior will lead to future safety. Locking the door at night is a form of prospective safety in the absence of discernible threat. 'Safety-seeking' behaviors fall under prospective safety. In the face of actual threat, safety seeking is a primary mechanism through which organisms prevent exacerbation of threat. However, safety-seeking responses to threat overestimation (e.g., phobias) contribute to **fear** maintenance through misattribution of safety[5]. Indeed, some of the most salient features of anxiety and compulsive disorders relate to behaviors intended to provide safety[6].

Linking safety to the threat imminence continuum (TIC)

Safety perceptions can be assessed by observing an organism's behavior. Fanselow and Lester's threat imminence continuum (TIC) provides a useful framework for examining how safety perceptions modify typical defensive behaviors[7] (Figure 1A). Four stages of attack probability form the TIC. 'Safe states' involve very low or no threat and often consist of contextual safety. In safe states, non-defensive motivated behaviors like foraging are

prioritized. Prospective safety planning in the form of niche construction may also occur. Increasing likelihood of attack shifts behavior away from flexible non-defensive pursuits toward defensive reactions[8]. The first non-safe TIC stage is the 'Pre-encounter threat' phase. During pre-encounter threat, no perceptible threat is present, but certain defensive behaviors are exhibited in an attempt to prevent having to engage in avoidance behavior. Once a predator is detected, 'Post-encounter threat' avoidance behaviors like freezing and fleeing are activated. At the post-encounter stage, the goal is to prevent transitioning into circa-strike. 'Circa-strike threat' exists when the predator is prepared to attack or is attacking. Importantly, safety can occur at any stage of the TIC. For example, even during circa-strike, carrying a powerful weapon can promote safety (Figure 1B). Thus, safety ultimately relates to coping ability, which can change despite spatiotemporal properties of the threat remaining constant.

Protection taxonomy

One of the most conserved ecological demands across species is the adaptive ability to acquire safety. How safety is secured, however, differs across species and contexts (Supplemental Figure S1). Protection is deployed at different points on the TIC[7] depending on its reliance on prospection. As reliance on prospection increases, so does the ability to employ the protection at an earlier imminence stage. 'Phenotypic' and 'Niche' protections are often deployed during post-encounter and circa-strike phases. Retreat to a nest or vocalization to announce an approaching predator during pre-encounter or early post-encounter phases may actually be counterproductive by drawing unwanted attention[9]. 'Social' and 'Manufactured' protection are used to create safe states or to reduce attack likelihood during pre- and post-encounter phases. The presence of social others can minimize surprise attack as a function of the many eyes effect thereby increasing safe-state behaviors[10]. Social protection can also signal safety to both the agent and the predator. For example, what we term the Chuck Norris effect bidirectionally signals safety via protection conferred by physically or socially competent others[11]. Manufactured safety, like the presence of a weapon, decreases attack likelihood when a threat is present as a result of increased risk to the threatening counterpart[12]. Lastly, 'Tactical' protection may be employed during safe and pre-encounter phases in response to prospective threats. Humans are particularly adept at developing complex and frequently symbolic tactical protections because our primary source of social threat is from conspecifics. For example, learning how to accurately interpret facial expressions can help ward off aversive social rejection[13].

Safety is not inverted threat

Safety can be signaled alone or in the context of threat. In the absence of threat, safety may not be consciously processed, but behaviors like foraging and mating indicate safety is computed. In the context of threat, perception of safety can differ as a function of protection despite holding attack probability constant (Figure 1B). As a result, behavior may be differentiated along the TIC depending on safety (Figure 1C). If safety is computed, non-defensive survival behaviors continue even as threats become more imminent. Although the imminence context remains the same, the presence of protection or another safety cue facilitates behavior characteristic of a lower imminence context (i.e., reduced vigilance,

increased exploration). Safety also extends opportunities for learning about the environment. Subjective perception of safety, rather than objective safety, determines exigency of defensive behavior. If safety is inaccurately overestimated, the likelihood of danger may ultimately increase due to inhibited defensive behavior (i.e., risk compensation)[14]. Beyond behavior, safety is also associated with altered subjective emotional experiences (Figure 1C). Safety decisions motivate individuals to explore their environment with the aid of positively-valenced emotions like curiosity[15].

Safety Decision Model

Computationally, humans attempt to solve the problem of determining to what extent they are safe. We propose a solution that partitions the complexity of safety decisions into two main evaluative components, which are then implemented at the neural level. **Threat-oriented evaluation** involves determining whether an external stimulus poses a threat of adverse outcome and is influenced by relative titration of three primary factors: imminence, value, and uncertainty (Figure 2A; Key Figure). The resources at one's disposal for combating threat are computed within **self-oriented evaluation**. This internally triggered evaluation is influenced by policies, experience, and control (Figure 2A).

We assert that safety decisions depend on the interaction between threat- and self-oriented evaluations, which dictate an organism's belief about their ability to cope with threat. We suggest safety evaluative components play interconnected roles in shaping decision-making through reciprocal influences on each other. For example, when on a roller coaster, our threat-oriented defensive circuits sense danger and respond by releasing adrenaline and analgesic opioids. At the same time, our self-oriented evaluation of the situation engages circuits that dampen defensive responding because we know we have safety protections such as harnesses, engineering, and regulations. Metacognitive integration of threat and protection estimates transforms the experience from threatening to rewarding. The anterior vmPFC has been identified as integrating socially relevant safety signals[16], and we propose this region is pivotal to integrating information to compute safety. As safety becomes more certain, features such as uncertainty and control may not be explicitly processed while other features like imminence and experience exert greater influence (Figure 2B).

Threat-oriented evaluation

A requisite step in determining safety is to evaluate external stimuli and environments for threat potential. Imminence, value, and uncertainty shape safety inferences by informing external threat representations. Threat-oriented evaluations are updated and influenced by other threat- and self-oriented components.

Imminence.—Traditionally, imminence is defined as spatiotemporal proximity of a predator[7,17,18] (Figure 1A). The spatiotemporal distance to a threat is critical in organizing defensive behaviors, however, perceptions of safety can also influence these responses. That is, the perception of the external environment is biased depending on the perceiver's needs and goals. Subjective imminence is influenced by perceptions of threat intensity[19] as well as protection, which directly affects the need to avoid danger. As

perception of attack probability increases along the TIC, cognitive appraisals, emotional experiences, and defensive behaviors transition.

Imminence also places constraints on the neural basis of decision-making. Increasing threat imminence triggers a neural shift from safety signaling in the anterior vmPFC to fear signaling in the periaqueductal gray (PAG)[20]. In low imminence contexts, agents can learn which protections confer safety and the extent to which they are effective against various threats. This safety learning process relies on the vmPFC[21]. When under imminent attack, however, innate responses are triggered by defensive circuits including the PAG, hypothalamus, and midcingulate cortex (MCC)[22]. These neural fear circuits support fast, reactive defensive behaviors and are considered a hardwired circuit[22]. When threats are imminent and safety is diminished, strategizing that takes place in the posterior vmPFC is too slow to mount adaptive behavior[23]. Bypassing the vmPFC to more reactive circuits under imminent danger may be one reason for less robust evidence of the posterior vmPFC in response to certain threats such as circa-strike. However, recent work in non-human primates points to a causal role of the posterior vmPFC in generation of defensive behavior in response to distal threat (areas 25, 14)[24,25] and anticipatory responses to threat during **Pavlovian conditioning** (area 25)[24,26].

Value.—Decision processes require value computations to guide behavior[27]. We define value as the weight or level of danger of a stimulus. Value is influenced by basic stimulus features (i.e., size) as well as more elaborate information about context, including imminence, and past experience[28]. Self-oriented features and competing goals also influence threat value. For example, individuals are willing to pay more money to prevent a stranger's pain than their own, demonstrating an increase in threat value when others are involved[29]. Stimuli of higher threat value evoke greater attentional focus and are more likely to result in defensive responses[30].

A flexible network consisting of the amygdala, striatum, insula, hippocampus, and vmPFC computes stimulus value[27]. The basolateral amygdala (BLA) plays a key role in modulating appetitive and aversive behavior through connections with the central amygdala (CeA) and nucleus accumbens (NAcc)[31]. Value signals from subcortical structures converge in the vmPFC, which exerts control over the BLA through reciprocal connections and via the hippocampus to guide decision making[32,33]. In mice, mPFC synchrony with the BLA switches from bidirectional during threat to mPFC-to-BLA directional modulation during safety[34]. In safe contexts, the human vmPFC is considered to play a regulatory role, reducing threat-related value representations through downregulation of aversive neural signals. The anterior vmPFC also couples with the striatum during safety, helping to drive appetitive behavior[35].

Uncertainty.—Uncertainty can originate both from the external world and from internal states. We classify safety-related uncertainty as a threat-oriented feature given that even internal uncertainty (e.g., Can I effectively combat this threat?) is related to external threat features (e.g., How powerful is this predator?). Threat-oriented uncertainty can be reducible through learning[36]. Once previously lacking knowledge about the probability of an aversive outcome is acquired, behavior can be adapted to achieve safety. For example, rock

climbing involves potential danger, but, for an experienced climber, neural signals convey safety. Increased experience (a self-oriented safety feature) can increase safety due to influences on uncertainty: experience decreases estimation uncertainty about impending danger and increases access to more protective safety-promoting solutions. However, irreducible uncertainty may remain regarding inherent randomness in threat behavior. Additionally, if the environment suddenly changes, unexpected uncertainty can decrease safety estimates. For example, an earthquake might render the expert climber's planned protection ineffective. Uncertainty computations vary across individuals, and intolerance of uncertainty is associated with both anxiety and compromised safety associations[37,38].

The anterior MCC (aMCC) processes affective and sensory information to guide learning under conditions of uncertainty[39,40]. Together with more dorsal regions of the anterior cingulate, the aMCC aids action selection by executing appropriate control over downstream regions to maximize positive outcomes[41]. The aMCC has widespread connections to key regions of the defensive circuitry including the amygdala, insula, striatum, and PAG[42,43]. The aMCC is thus well-positioned to modify behavior in response to uncertainty by modulating other threat-oriented evaluations[44,45]. The anterior insula has also been implicated in uncertainty, but is thought to play a greater role in interoceptive predictions than outcome predictions[46].

Self-oriented evaluation

Interactions between systematic policies, prior experience, and perceived control inform safety decisions by collectively representing an organism's competence for successfully coping with threat. Importantly, self-oriented evaluation relies on metacognition, or the awareness of one's own internal thought processes. Poor metacognition can impair perceptual judgment and ability to integrate errors and update safety estimations[47].

Policies.—Humans are uniquely adept at developing formal policies, or guidelines, to systematically combat threats[48]. **Model-based** computations typically occur through prospective simulation of possible future action-outcome contingencies. Safety promotes development of model-based policies by extending the capacity to gather and integrate explicit knowledge of the environment[49]. Because **model-free** policies are fast, they are favored as estimates of danger increase and time horizons for engaging in slower model-based planning decrease. However, overly conservative defensive responses can be counterproductive, increasing the utility of establishing accurate and flexible model-based policies when safe[18]. As experience increases, model-based episodes can become model-free and habitual. Safety computations may thereby contribute to future computational efficiency wherein initial estimates of safety are resource-intensive but future safety computations are performed with greater speed and ease.

The hippocampus, in concert with the vmPFC, contributes to development of model-based representations by providing memory and prospection inputs to predict future threat behavior[50]. Neural shifts reflecting utility of model-based versus model-free processes mirror the shift in defensive circuitry with imminence. As time horizons for defensive responding are lengthened through safety, model-based planning opportunities are available

(Figure 1C). Safety also promotes experiential diversity, which in turn allows formation of model-based constructions of the environment and maximization of prospection flexibility [51].

Experience.—Optimal decision making depends on flexibly adapting behavior based on experience[52]. **Bayesian decision theory** provides a useful framework for understanding how humans integrate prior experience and current likelihood information to make predictions about possible future outcomes[53]. For example, repeatedly encountering a snake with no adverse outcome should reduce expectations of future risk. Safety expectations should only be updated if new information becomes available. However, past experience is not objectively integrated. Salience of past experience is influenced by contextual and situational factors, including emotions[54,55].

The hippocampus is part of a flexible decision circuit that promotes strategic decisions in safe states[23]. Together with the vmPFC, the hippocampus supports non-defensive behavior by computing a model-based representation of the environment that incorporates internal representations of prior experience[56]. The hippocampus can also facilitate safety learning by promoting safety memories associated with experiences of successfully confronting threatening situations[57]. Generalization of prior threat experiences to perceptually similar stimuli and contexts is a common failure in safety processing characteristic of threat-related psychopathology and linked to diminished threat-safety discrimination in the anterior vmPFC[58]. Relevance of stimulus-outcome associations also differentially evokes safety neural circuitry. Safe stimuli never previously associated with threat activate the vmPFC[59] and hippocampus[60] to a lesser extent than cues indicating cessation of previously encountered danger. Recent work suggests a combination of direct and vicarious learning enhances safety associations[61]. The benefit of direct experience may be one reason safety computations differ across development (Box 1).

Control.—Perceived control refers to the belief that one can influence the outcome of an event[62]. In response to continued threat, agents inhibit defensive responding to reallocate energy to other survival processes such as boosting immune response to combat infection from attack wounds[63]. The presence of control reverses this inhibition so that active safety-oriented behaviors can continue[64]. Controllability also increases proactive, model-based behavior[65] which may influence development of safety-decision policies, again likely promoting survival. Conversely, underestimations of control reduce error-related learning[47], resulting in a problematic feedback loop whereby prediction errors do not update future safety estimations. Underestimation of perceived control is a transdiagnostic vulnerability factor across anxiety disorders[66], making control a promising target for boosting safety decisions in clinical populations.

Control is detected by the anterior vmPFC, which responds by outputting to the limbic system and brain stem to inhibit stress-induced behavioral and psychological responses[67]. The human vmPFC plays a role in calibrating behavioral responses based on controllability estimates[68], and vmPFC inhibition of stress-induced limbic activation[65] may account for the reduction in fear expression in response to controllable threats[69]. Experiences of threat-related control also bolsters resilience to future threats, in part because of plasticity in

vmPFC-limbic circuits that increases vmPFC response to later uncontrollable threats[70]. As such, the anterior vmPFC has been identified as a key locus of behavioral control and coping[71].

Observable safety decision responses

Threat- and self-oriented evaluations are continuously integrated to result in neural, behavioral, and cognitive responses (Figure 2A). We propose that threat-oriented features are computed neurally in a bottom-up fashion such that they are primarily driven by sensory processes. Self-oriented features are integrated through top-down metacognitive processes to update representations of threat-related sensory input[72]. Although crudely represented as a dual system, each lower-level feature also iteratively influences other features via integration in the vmPFC[73]. In addition to the Safety Decision Model components, time, emotions, and goals exert influence. For example, anxiety about negative consequences may manifest as preoccupation with threat, as is often observed in anxiety, and likely increases prioritization of model-free policies, emphasis of negative prior experiences, and underestimations of control[74]. Behaviorally, increased safety estimates result in non-defensive behavior whereas reduced safety estimates result in defensive responses. Behavioral outputs depend on relative comparison of current states to past states (and, in some cases, prospective states to current states). This comparative process relies on accurate estimation of both present and previous safety, which taps into memory, self-projection, and conflict monitoring[75]. Prediction errors and reinforcement signals are then used to develop models of the environment and update future safety decisions.

Safety neural circuitry

We propose distinct contributions of regions in the human defensive neural circuitry responsible for safety and threat decisions (Figure 3A). We suggest adaptive coding from mixed selectivity neurons in core defensive regions like the BLA, aMCC, hippocampus, striatum, and insula allows the brain to function differently depending on signals from the vmPFC[75]. For example, anterior vmPFC regulation of the insula can produce an analgesic effect in response to pain[76]. Safety signaling also alters BLA output projections, redirecting to the striatum to facilitate approach behavior rather than to the CeA during threat which facilitates defensive behavior[31].

Central to our neural model is the proposal that the anterior vmPFC (Brodmann area (BA) 10r) and posterior vmPFC (BA 25 and 32pl) contribute to safety and threat computation, respectively [77]. We present three lines of evidence to support this proposal: (1) studies supporting safety-related anterior vmPFC activation and threat-related posterior vmPFC activation (Figure 3B), (2) Neurosynth meta-analytic decoding of Safety Decision Model components by vmPFC subregion (Figure 3C), (3) Neurosynth meta-analytic coactivations by vmPFC subregion (Figure 3D). See Supplemental Materials for Figure 3C–D details.

Anterior/posterior vmPFC identification

Extensive prior work identifies a role for the human anterior vmPFC in extinction and reversal of Pavlovian conditioning. More recent evidence from lesion patients suggests the

vmPFC also plays a causal role in threat acquisition[78]. Despite general consensus of vmPFC involvement in response to threat, a clear delineation of vmPFC subregions in threat and safety processing is lacking. Affective processing in the vmPFC provides initial insight into a posterior-anterior split with the generation of negative and positive affect following a similar divide[79]. We examined vmPFC activation for a selection of studies on neural response to safety and threat, which revealed consistent concentration in the anterior and posterior vmPFC, respectively (Figure 3B). These studies included a range of paradigms (Supplemental Table S1) supporting generalizability of the proposed vmPFC fractionation. Notably, our model extends beyond classic threat models to include second-order representation of one's own thoughts in safety decisions. A recent meta-analysis of metacognition identifies the anterior vmPFC as a hub for metacognitive and mentalizing processes, which may drive the involvement of the anterior vmPFC in response to safety[80]. The vmPFC is also a key hub of the default mode network (DMN), a network characterized by a distinct function profile of higher resting activity. The DMN reliably engages during metacognition and inferences about others' mental states, both of which are key components of safety processing.

Decoding meta-analysis

Using meta-analytical decoding with Neurosynth, we probed correlations between vmPFC subregions and Safety Decision Model components (Figure 3C). The anterior vmPFC subregion evinced positive correlation with model features whereas the posterior vmPFC was correlated with fear.

Coactivation meta-analysis

We performed an automated meta-analysis on coactivation patterns using the Neurosynth database for the vmPFC subregions (Figure 3D)[81,82]. Results demonstrated important differences in coactivation and support the assertion that involvement of the anterior vmPFC alters responding of other defensive regions. The anterior vmPFC coactivates to a greater extent with the hippocampus, ventral striatum, thalamus, and hypothalamus, whereas the posterior vmPFC coactivates to a greater extent with the PAG, aMCC, and dorsal anterior insula. Coactivation maps for subcomponents of the Safety Decision Model are provided in Supplemental Figure S2.

We propose safety can vary despite threat-oriented features remaining constant. If this is the case, it is likely that neural systems independent from those involved in threat detection contribute to safety computations. Rodent work supports this proposal identifying distinct neuronal ensemble patterns for danger and safety stimuli. The prelimbic subdivision of the rodent mPFC monitors danger, while the infralimbic subdivision conveys safety[83]. Our neural model suggests anterior vmPFC signals alter the way canonical defensive regions respond to safety computations. Although we identify converging evidence for a role of the anterior vmPFC in safety decisions, reliance on threat paradigms with co-occurring safe states is insufficient. Threat-based paradigms infrequently manipulate self-oriented features of safety processing. Thus, identifying the precise population coding dynamics involved in computing safety requires investigation of the spatial and temporal tuning properties of the anterior vmPFC using paradigms designed to elicit safety decisions[84].

Concluding remarks

We propose that successful safety decisions rely on interpretation and weighting of threat-oriented and self-oriented evaluative factors that indicate survival probability in the face of threat (see Outstanding Questions). We link our cognitive model to processing in the vmPFC, a hub of the brain's DMN. This raises an intriguing possibility that safety is a fundamental aspect of baseline human cognition. If the default state involves computing safety, it is further probable that safety seeking is an attempt at restoring equilibrium. Metacognitive processes subserved by the DMN may be necessary for computing safety, and DMN-related psychopathologies like anxiety may result from metacognitive deficits necessary for safety computing. Our framework provides discrete cognitive processes that can be empirically manipulated to identify boundary contexts that shift perceptions from threat to safety. Although we provide evidence for anterior and posterior vmPFC involvement in safety and threat, evidence supporting this proposal is mixed. Further work is needed to determine when the vmPFC is necessary to promote survival and under what conditions vmPFC supported cognition is unavailable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

DM and SMT are supported by the US National Institute of Mental Health grant no. 2P50MH094258 and Templeton Foundation grant TWCF0366. TDZ is supported by a National Science Foundation grant no. 1911441.

Glossary

Bayesian Decision Theory
A decision theory based on the concept that knowledge about the probability of a possible outcome is estimated from a posterior distribution formed by integrating a prior distribution and a likelihood function. The prior distribution reflects past knowledge whereas the likelihood function summarizes current sensory information

Defensive survival circuits
Neural circuitry triggered by threat that determines which defensive action to choose depending on threat-oriented features.

Fear
Fear is a present emotional state associated with a perceptible threat. Fear is distinct from anxiety, which is a future-oriented emotional state associated with uncertain threats. Fear and anxiety can be distinguished based on the level of certainty about the likelihood, imminence, and value of future threat

Model-based policy
A (typically) learned process by which organisms exploit a model of the environment to prospectively estimate the likely outcome of actions.

Model-free policy
A process by which organisms estimate the value of actions by encoding immediate action-outcome contingencies. Similar to a reflex

Pavlovian conditioning
Also known as classical or respondent conditioning. A learning procedure in which a biologically salient stimulus (e.g., appetitive food; aversive shock) is paired with a previously neutral stimulus (e.g., a tone or image).

Self-oriented evaluation
A component of safety computation that entails evaluating one's preparedness to combat a potential threat. Policies, experience, and control are integrated to predict available coping resources

Threat
An increase in the probability of danger or reference to a stimulus that, through its presence or action, increases the probability of danger. An example of a threatening stimulus is a snake. The danger the snake poses is physical harm

Threat-oriented evaluation
A component of safety computation that entails evaluating the nature of a potentially threatening stimulus. Imminence, value, and uncertainty are integrated to predict stimulus danger

References

1. Lima SL and Dill LM (1990) Behavioral decisions made under the risk of predation: a review and prospectus. Can. J. Zool. 68, 619–640
2. Gonzalez ST and Fanselow MS (2020) The role of the ventromedial prefrontal cortex and context in regulating fear learning and extinction. Psychology & Neuroscience 13, 459–472
3. Sangha S et al. (2020) Know safety, no fear. Neuroscience & Biobehavioral Reviews 108, 218–230 [PubMed: 31738952]
4. Sotres-Bayon F and Quirk GJ (2010) Prefrontal control of fear: more than just extinction. Current Opinion in Neurobiology 20, 231–235 [PubMed: 20303254]
5. Thwaites R and Freeston MH (2005) Safety-Seeking Behaviours: Fact or Function? How Can We Clinically Differentiate Between Safety Behaviours and Adaptive Coping Strategies Across Anxiety Disorders? Behavioural and Cognitive Psychotherapy 33, 177–188
6. Angelakis I and Austin JL (2018) The effects of the non-contingent presentation of safety signals on the elimination of safety behaviors: An experimental comparison between individuals with low and high obsessive-compulsive profiles. Journal of Behavior Therapy and Experimental Psychiatry 59, 100–106 [PubMed: 29291473]
7. Fanselow MS and Lester LS (1988) A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior. In Evolution and learning pp. 185–212, Lawrence Erlbaum Associates, Inc
8. Gray JA and McNaughton N (2003) The Neuropsychology of Anxiety: An enquiry into the function of the septo-hippocampal system, Oxford University Press.
9. Fanselow MS et al. (2019) Timing and the transition between modes in the defensive behavior system. Behavioural Processes 166, 103890 [PubMed: 31254627]
10. Bertram BCR (1980) Vigilance and group size in ostriches. Animal Behaviour 28, 278–286

11. Tedeschi E et al. (2015) Inferences of Others' Competence Reduces Anticipation of Pain When under Threat. Journal of Cognitive Neuroscience 27, 2071–2078 [PubMed: 26102229]
12. Emlen DJ (2008) The Evolution of Animal Weapons. Annual Review of Ecology, Evolution, and Systematics 39, 387–413
13. Bernstein MJ et al. (2008) Adaptive Responses to Social Exclusion: Social Rejection Improves Detection of Real and Fake Smiles. Psychol Sci 19, 981–983 [PubMed: 19000206]
14. Hedlund J (2000) Risky business: safety regulations, risk compensation, and individual behavior. Inj Prev 6, 82–90 [PubMed: 10875661]
15. Maner JK and Gerend MA (2007) Motivationally selective risk judgments: Do fear and curiosity boost the boons or the banes? Organizational Behavior and Human Decision Processes 103, 256–267
16. Qi S et al. (2018) A Collaborator's Reputation Can Bias Decisions and Anxiety under Uncertainty. J. Neurosci. 38, 2262–2269 [PubMed: 29378862]
17. McNaughton N and Corr PJ (2004) A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. Neurosci Biobehav Rev 28, 285–305 [PubMed: 15225972]
18. Mobbs D et al. (2020) Space, Time, and Fear: Survival Computations along Defensive Circuits. Trends in Cognitive Sciences 24, 228–241 [PubMed: 32029360]
19. Cole S et al. (2013) Affective Signals of Threat Increase Perceived Proximity. Psychological science DOI: 10.1177/0956797612446953
20. Mobbs D et al. (2010) Neural activity associated with monitoring the oscillating threat value of a tarantula. Proc. Natl. Acad. Sci. U.S.A. 107, 20582–20586 [PubMed: 21059963]
21. Savage HS et al. (2020) Clarifying the neural substrates of threat and safety reversal learning in humans. NeuroImage 207, 116427 [PubMed: 31801684]
22. Mobbs D et al. (2018) Foraging for foundations in decision neuroscience: insights from ethology. Nat Rev Neurosci 19, 419–427 [PubMed: 29752468]
23. Qi S et al. (2018) How cognitive and reactive fear circuits optimize escape decisions in humans. Proc Natl Acad Sci USA 115, 3186–3191 [PubMed: 29507207]
24. Alexander L et al. (2020) Over-activation of primate subgenual cingulate cortex enhances the cardiovascular, behavioral and neural responses to threat. Nature Communications 11, 5386
25. Stawicka ZM et al. (2020) Ventromedial prefrontal area 14 provides opposing regulation of threat and reward-elicited responses in the common marmoset. Proc Natl Acad Sci USA DOI: 10.1073/pnas.2009657117
26. Wallis CU et al. (2017) Opposing roles of primate areas 25 and 32 and their putative rodent homologs in the regulation of negative emotion. Proc Natl Acad Sci USA 114, E4075–E4084 [PubMed: 28461477]
27. Rangel A et al. (2008) A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience 9, 545–556 [PubMed: 18545266]
28. Mogg K et al. (2000) Selective attention to threat: A test of two cognitive models of anxiety. Cognition & Emotion 14, 375–399
29. Crockett MJ et al. (2014) Harm to others outweighs harm to self in moral decision making. Proc Natl Acad Sci U S A 111, 17320–17325 [PubMed: 25404350]
30. Wise T et al. (2019) A computational account of threat-related attentional bias. PLOS Computational Biology 15, e1007341 [PubMed: 31600187]
31. Piantadosi PT et al. (2017) Contributions of basolateral amygdala and nucleus accumbens subregions to mediating motivational conflict during punished reward-seeking. Neurobiol Learn Mem 140, 92–105 [PubMed: 28242266]
32. Grabenhorst F and Rolls ET (2011) Value, pleasure and choice in the ventral prefrontal cortex. Trends in Cognitive Sciences 15, 56–67 [PubMed: 21216655]
33. Levy DJ and Glimcher PW (2012) The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. 22, 1027–1038 [PubMed: 22766486]
34. Stujenske JM et al. (2014) Fear and safety engage competing patterns of theta-gamma coupling in the basolateral amygdala. Neuron 83, 919–933 [PubMed: 25144877]

35. Qi S et al. (2020) The Role of the Medial Prefrontal Cortex in Spatial Margin of Safety Calculations. bioRxiv DOI: 10.1101/2020.06.05.137075
36. Payzan-LeNestour E and Bossaerts P (2011) Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. PLoS Comput Biol 7, e1001048 [PubMed: 21283774]
37. Jensen D et al. (2016) Clarifying the Unique Associations among Intolerance of Uncertainty, Anxiety, and Depression. Cogn Behav Ther 45, 431–444 [PubMed: 27314213]
38. Morriss J et al. (2016) Nothing is safe: Intolerance of uncertainty is associated with compromised fear extinction learning. Biol Psychol 121, 187–193 [PubMed: 27178640]
39. Shackman AJ et al. (2011) The Integration of Negative Affect, Pain, and Cognitive Control in the Cingulate Cortex. Nat Rev Neurosci 12, 154–167 [PubMed: 21331082]
40. Morriss J et al. (2019) The uncertain brain: A co-ordinate based meta-analysis of the neural signatures supporting uncertainty during different contexts. Neurosci Biobehav Rev 96, 241– 249 [PubMed: 30550858]
41. Shenhav A et al. (2016) Dorsal anterior cingulate cortex and the value of control. Nature Neuroscience 19, 1286–1291 [PubMed: 27669989]
42. Grupe DW and Nitschke JB (2013) Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. Nat Rev Neurosci 14, 488–501 [PubMed: 23783199]
43. Milad MR et al. (2007) A role for the human dorsal anterior cingulate cortex in fear expression. Biol. Psychiatry 62, 1191–1194 [PubMed: 17707349]
44. Soltani A and Izquierdo A (2019) Adaptive learning under expected and unexpected uncertainty. Nature Reviews Neuroscience 20, 635–644 [PubMed: 31147631]
45. Stern ER et al. (2010) Updating Beliefs for a Decision: Neural Correlates of Uncertainty and Underconfidence. J Neurosci 30, 8032–8041 [PubMed: 20534851]
46. Paulus MP and Stein MB (2006) An Insular View of Anxiety. Biological Psychiatry 60, 383–387 [PubMed: 16780813]
47. Rigoni D et al. (2013) When errors do not matter: Weakening belief in intentional control impairs cognitive reaction to errors. Cognition 127, 264–269 [PubMed: 23466640]
48. Bach DR and Dayan P (2017) Algorithms for survival: a comparative perspective on emotions. Nat Rev Neurosci 18, 311–319 [PubMed: 28360419]
49. Wang O et al. (2018) Model-based and model-free pain avoidance learning. Brain and Neuroscience Advances 2, 239821281877296
50. Moscarello JM and Maren S (2018) Flexibility in the face of fear: hippocampal–prefrontal regulation of fear and avoidance. Current Opinion in Behavioral Sciences 19, 44–49 [PubMed: 29333482]
51. Heller AS et al. (2020) Association between real-world experiential diversity and positive affect relates to hippocampal-striatal functional connectivity. Nat. Neurosci. 23, 800–804 [PubMed: 32424287]
52. Baglan H et al. (2017) Learning in mosquito larvae (Aedes aegypti): Habituation to a visual danger signal. Journal of Insect Physiology 98, 160–166 [PubMed: 28077263]
53. Berger JO (2013) Statistical Decision Theory and Bayesian Analysis, Springer Science & Business Media.
54. Xu C et al. (2016) Distinct Hippocampal Pathways Mediate Dissociable Roles of Context in Memory Retrieval. Cell 167, 961–972.e16 [PubMed: 27773481]
55. Paulus MP and Yu AJ (2012) Emotion and decision-making: affect-driven belief systems in anxiety and depression. Trends in Cognitive Sciences 16, 476–483 [PubMed: 22898207]
56. Rigoli F et al. (2016) Neural processes mediating contextual influences on human choice behaviour. Nature Communications 7, 12416
57. Micale V et al. (2017) Extinction of avoidance behavior by safety learning depends on endocannabinoid signaling in the hippocampus. J Psychiatr Res 90, 46–59 [PubMed: 28222356]
58. Greenberg T et al. (2013) Ventromedial prefrontal cortex reactivity is altered in generalized anxiety disorder during fear generalization. Depress Anxiety 30, 242–250 [PubMed: 23139148]

59. Schiller D et al. (2008) From fear to safety and back: reversal of fear in the human brain. J. Neurosci. 28, 11517–11525 [PubMed: 18987188]
60. Meyer HC et al. (2019) Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. PNAS 116, 26970–26979
61. Pan Y et al. (2020) Social safety learning: Shared safety abolishes the recovery of learned threat. Behaviour Research and Therapy DOI: 10.1016/j.brat.2020.103733
62. Bollini AM et al. (2004) The influence of perceived control and locus of control on the cortisol and subjective responses to stress. Biological Psychology 67, 245–260 [PubMed: 15294384]
63. Frank MG et al. (2013) Stress-induced glucocorticoids as a neuroendocrine alarm signal of danger. Brain Behav Immun 33, 1–6 [PubMed: 23459026]
64. Maier SF and Seligman MEP (2016) Learned Helplessness at Fifty: Insights from Neuroscience. Psychol Rev 123, 349–367 [PubMed: 27337390]
65. Moscarello JM and Hartley CA (2017) Agency and the Calibration of Motivated Behavior. Trends in Cognitive Sciences 21, 725–735 [PubMed: 28693961]
66. Gallagher MW et al. (2014) Perceived control and vulnerability to anxiety disorders: A meta-analytic review. Cognitive Therapy and Research 38, 571–584
67. Maier SF et al. (2006) Behavioral control, the medial prefrontal cortex, and resilience. Dialogues Clin Neurosci 8, 397–406 [PubMed: 17290798]
68. Collins KA et al. (2014) Taking Action in the Face of Threat: Neural Synchronization Predicts Adaptive Coping. Journal of Neuroscience 34, 14733–14738 [PubMed: 25355225]
69. Hartley CA et al. (2014) Stressor controllability modulates fear extinction in humans. Neurobiol Learn Mem 113, 149–156 [PubMed: 24333646]
70. Maier SF and Watkins LR (2010) Role of the medial prefrontal cortex in coping and resilience. Brain Res 1355, 52–60 [PubMed: 20727864]
71. Sinha R et al. (2016) Dynamic neural activity during stress signals resilient coping. Proc Natl Acad Sci USA 113, 8837–8842 [PubMed: 27432990]
72. Kveraga K et al. (2007) Top-down predictions in the cognitive brain. Brain and Cognition 65, 145–168 [PubMed: 17923222]
73. Roy M et al. (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci 16, 147–156 [PubMed: 22310704]
74. Wells A (1995) Meta-Cognition and Worry: A Cognitive Model of Generalized Anxiety Disorder. Behav. Cogn. Psychother. 23, 301–320
75. Euston DR et al. (2012) The Role of Medial Prefrontal Cortex in Memory and Decision Making. Neuron 76, 1057–1070 [PubMed: 23259943]
76. Wang Y et al. (2020) Altruistic behaviors relieve physical pain. Proc Natl Acad Sci U S A 117, 950–958 [PubMed: 31888986]
77. Ongür D et al. (2003) Architectonic subdivision of the human orbital and medial prefrontal cortex. J Comp Neurol 460, 425–449 [PubMed: 12692859]
78. Battaglia S et al. (2020) Revaluing the Role of vmPFC in the Acquisition of Pavlovian Threat Conditioning in Humans. J. Neurosci. 40, 8491–8500 [PubMed: 33020217]
79. Myers-Schulz B and Koenigs M (2012) Functional anatomy of ventromedial prefrontal cortex: implications for mood and anxiety disorders. Mol. Psychiatry 17, 132–141 [PubMed: 21788943]
80. Vaccaro AG and Fleming SM (2018) Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. Brain and Neuroscience Advances 2, 2398212818810591 [PubMed: 30542659]
81. Qi S et al. (2020) The Role of the Medial Prefrontal Cortex in Spatial Margin of Safety Calculations, Neuroscience.
82. Yarkoni T et al. (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8, 665–670 [PubMed: 21706013]
83. Corches A et al. (2019) Differential fear conditioning generates prefrontal neural ensembles of safety signals. Behav Brain Res 360, 169–184 [PubMed: 30502356]
84. Panzeri S et al. (2015) Neural population coding: combining insights from microscopic and mass signals. Trends in Cognitive Sciences 19, 162–172 [PubMed: 25670005]

85. Jovanovic T et al. (2009) Posttraumatic stress disorder may be associated with impaired fear inhibition: relation to symptom severity. Psychiatry Res 167, 151–160 [PubMed: 19345420]
86. Lis S et al. (2020) Generalization of fear in post-traumatic stress disorder. Psychophysiology 57, e13422 [PubMed: 31206738]
87. Gagne C et al. (2018) When planning to survive goes wrong: predicting the future and replaying the past in anxiety and PTSD. Current Opinion in Behavioral Sciences 24, 89–95
88. Koenigs M et al. (2008) Focal brain damage protects against post-traumatic stress disorder in combat veterans. Nat Neurosci 11, 232–237 [PubMed: 18157125]
89. Lambert HK and McLaughlin KA (2019) Impaired hippocampus-dependent associative learning as a mechanism underlying PTSD: A meta-analysis. Neuroscience & Biobehavioral Reviews 107, 729–749 [PubMed: 31545990]
90. Shechner T et al. (2014) Fear conditioning and extinction across development: Evidence from human studies and animal models. Biological Psychology 100, 1–12 [PubMed: 24746848]
91. Morris TL and March JS (2004) Anxiety Disorders in Children and Adolescents, Guilford Press.
92. Weil LG et al. (2013) The development of metacognitive ability in adolescence. Conscious Cogn 22, 264–271 [PubMed: 23376348]
93. Lau JY et al. (2011) Distinct neural signatures of threat learning in adolescents and adults. Proceedings of the National Academy of Sciences 108, 4500–4505
94. Gee DG et al. (2013) A developmental shift from positive to negative connectivity in human amygdala-prefrontal circuitry. J. Neurosci. 33, 4584–4593 [PubMed: 23467374]
95. Tashjian SM et al. (2019) Multivoxel Pattern Analysis Reveals a Neural Phenotype for Trust Bias in Adolescents. J Cogn Neurosci 31, 1726–1741 [PubMed: 31322468]
96. Tottenham N et al. (2011) Elevated Amygdala Response to Faces Following Early Deprivation. Dev Sci 14, 190–204 [PubMed: 21399712]
97. Whissell CM (1989) The dictionary of affect in language. In The Measurement of Emotions (Plutchik R and Kellerman H, eds), pp. 113–131, Academic Press

Box 1.

Special Considerations: Psychopathology and Development

Individuals with skewed safety estimates may meet criteria for clinical disorders. For example, underestimated safety is a hallmark of post-traumatic stress disorder (PTSD) [85]. After experiencing trauma, individuals with PTSD overgeneralize fear to new situations. Safety misalignment in PTSD reflects errors in both threat- and safety-oriented computations. Individuals with PTSD demonstrate threat-oriented distortion whereby non-threatening stimuli become tagged as aversive. The more closely a safe stimulus resembles a threatening stimulus, the more impaired defensive responses become for individuals with PTSD[86]. Regarding self-oriented experiences, problems arise when the system overemphasizes negative prior experiences[87]. Metacognitive failure in recognizing these biases may be particularly relevant for the maintenance and exacerbation of PTSD[87]. Canonical defensive circuitry as critically involved in PTSD pathogenesis. vmPFC hypoactivation, amygdala hyperactivation, and altered hippocampal responding have been identified in individuals with PTSD[88,89]. Our Safety Decision Model provides a framework for future work to determine precise mechanisms of safety distortion in threat-related psychopathology.

Neural systems involved in threat and safety learning dramatically differ across development. With age, safety learning transitions from heavy reliance on the amygdala to a more distributed network including the vmPFC and hippocampus[90]. Adolescence poses a critical time to study safety computations for several reasons. First, prevalence of anxiety disorders increases from childhood to adolescence, putting youth at risk for threat-related psychopathology in adulthood[91]. Second, metacognitive abilities improve during adolescence, creating potential vulnerabilities in accurate self-oriented evaluations[92]. Third, adolescents demonstrate poorer threat-oriented evaluation compared to adults, with greater difficulty discriminating between threat and safety stimuli[93]. Earlier subcortical development and protracted PFC development during adolescence[94] point to the amygdala as having an outsized influence on safety computations during this period. Imbalance in amygdala-vmPFC contributions may result in overemphasized threat-oriented features and impaired self-oriented feature processing. Beyond activation differences, the way the adolescent amygdala represents valence is associated with individual differences in appraisal of uncertain stimuli, potentially acting as a phenotype of threat bias[95]. These phenotypic differences may result from adverse early-life experiences, which can have long term consequences for safety computations[96]. Targeting safety processing mechanisms outlined here may be a promising intervention for developmental populations.

Outstanding Questions

- Is the anterior vmPFC necessary for safety? The vmPFC is a heterogenous structure implicated in a myriad of affective and cognitive processes. These processes are not exclusive to safety, raising the question of whether the vmPFC is necessary for safety computing. Additionally, discrepancy exists concerning the role of the rodent vmPFC in safety. However, significant differences exist between the rodent and human frontal cortex. New safety paradigms and primate translational work are needed to establish a causal role for the vmPFC in safety.
- Are certain sources of safety observational learning more effective across development? Success of observational learning may depend on the salience of social others. The utility of different sources of information may differ across development. For example, in adolescence, when peers are highly salient, are peers also perceived as more reliable sources for safety learning than parents? Does the type of stimulus influence who is seen as a credible source of information?
- How does self-oriented evaluation affect safety generalization? Humans rarely repeatedly encounter the same stimuli in identical form and context. Deficits in safety generalization contribute to anxiety disorder development. Pavlovian fear conditioning studies suggest external features are generalizable during safety learning. However, these paradigms exclude self-oriented components that often accompany human safety decisions.
- How is protection evaluated? Neural and psychological mechanisms of protection evaluation are obscured by focus on extinction-based safety. We propose different types of protection rely on varying degrees of explicit and implicit processes. It is plausible that different protection types tap distinct neural systems associated with action generation, sensory perception, and learning. Protection is an important, yet understudied, component of safety.

Highlights

Survival decisions often hinge on the perception that one is safe. Without safety, defensive behaviors are prioritized at the expense of other behaviors. Underestimations of safety can therefore hinder survival and self-actualization.

Extant accounts of threat-related decision making implicitly suggest safety is an inverse of threat, or, in other words, as threat reduces, safety increases. While this may be true under certain conditions, we argue that safety computations are often distinct from threat systems due to separable metacognitive processes.

We propose the human ventromedial prefrontal cortex (vmPFC) integrates sensory information about threats with top-down predictions about self-oriented coping ability to drive safety decisions.

Building on evidence from threat-focused paradigms, we identify the anterior vmPFC as a candidate hub of safety decision making.

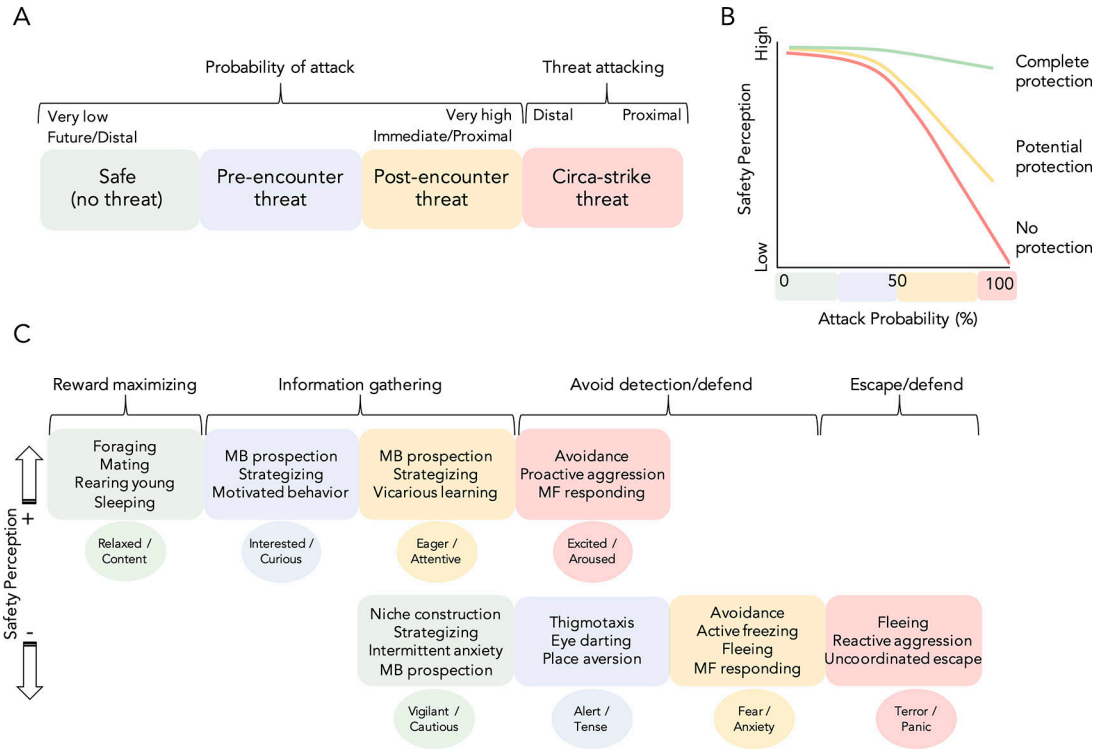

Figure 1. Protection and imminence influence safety estimates.

A. Threat imminence continuum (TIC). Based on Fanselow and Lester (1988). **B.** Protection alters safety perceptions along the TIC even as attack probability remains constant. **C.** Behavioral and emotional responses to threat imminence vary as a function of safety perception. Emotion terms for low safety contexts were based on Mobbs et al. (2020) and positively valenced emotions with similar arousal at the same level were selected for high safety contexts using the Dictionary of Affect in Language (DAL; [97]). MB=model based, MF=model free.

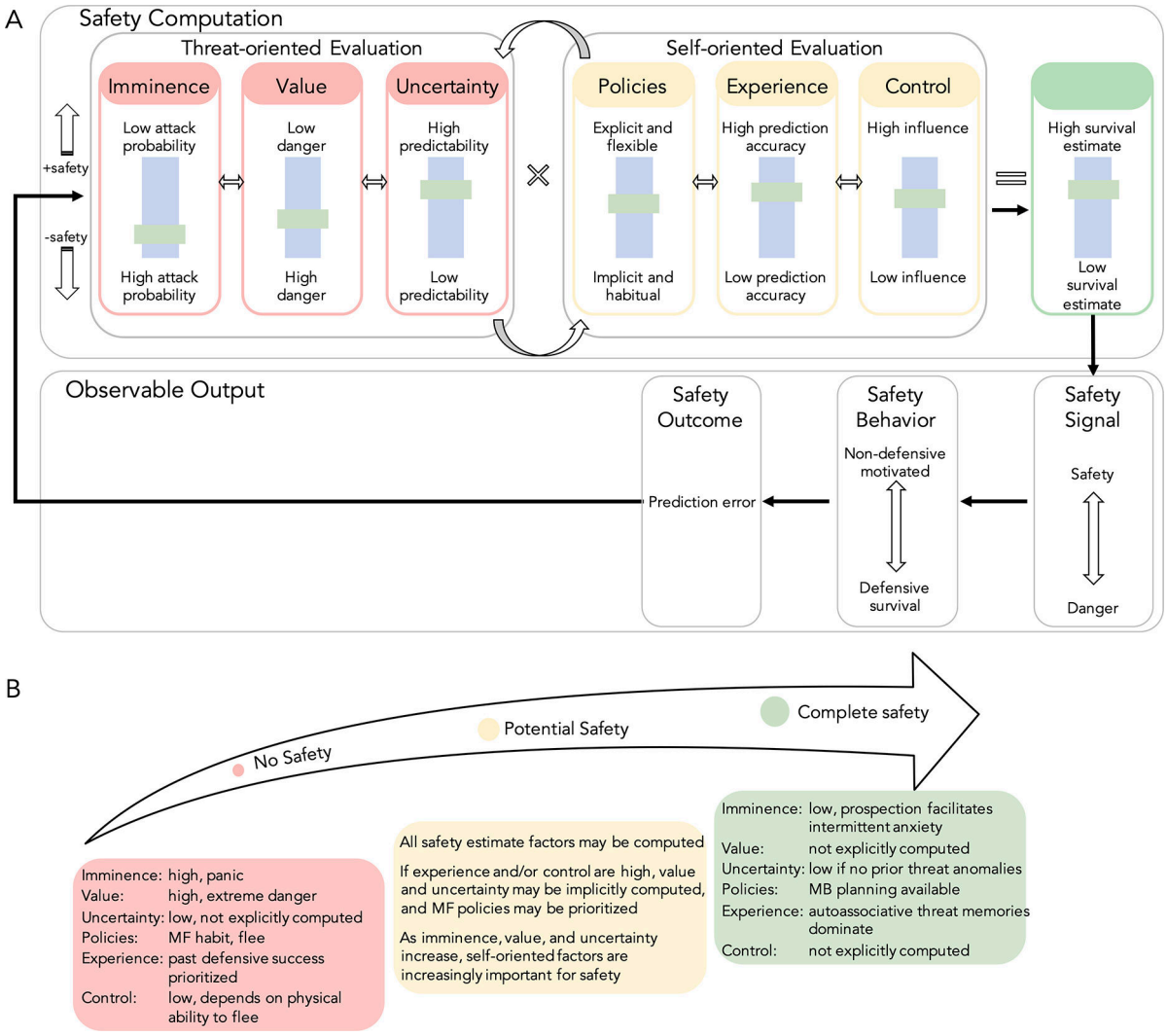

Figure 2, Key Figure. Safety Decision Model components and computational flexibility across the safety continuum.

A. A hypothetical rendering of the Safety Decision Model components underlying safety computation. Solid black arrows indicate temporal influence. Evaluative components exert bidirectional influence on other lower-level features. **Safety computation:** Relative titrating of threat- and self-oriented evaluative components are integrated in estimates of survival to compute perceived safety. **Safety signal:** Distinct predictions about survival success are encoded at the neural level. **Safety behavior:** As safety increases, non-defensive survival behaviors are prioritized. As safety decreases, threat monitoring and defensive behaviors increase, suppressing non-defensive motivated behavior. **Safety outcome:** Prediction errors are integrated to update future safety computations. **B.** Safety computation factors are proposed to vary in the extent to which they are processed along the safety continuum.

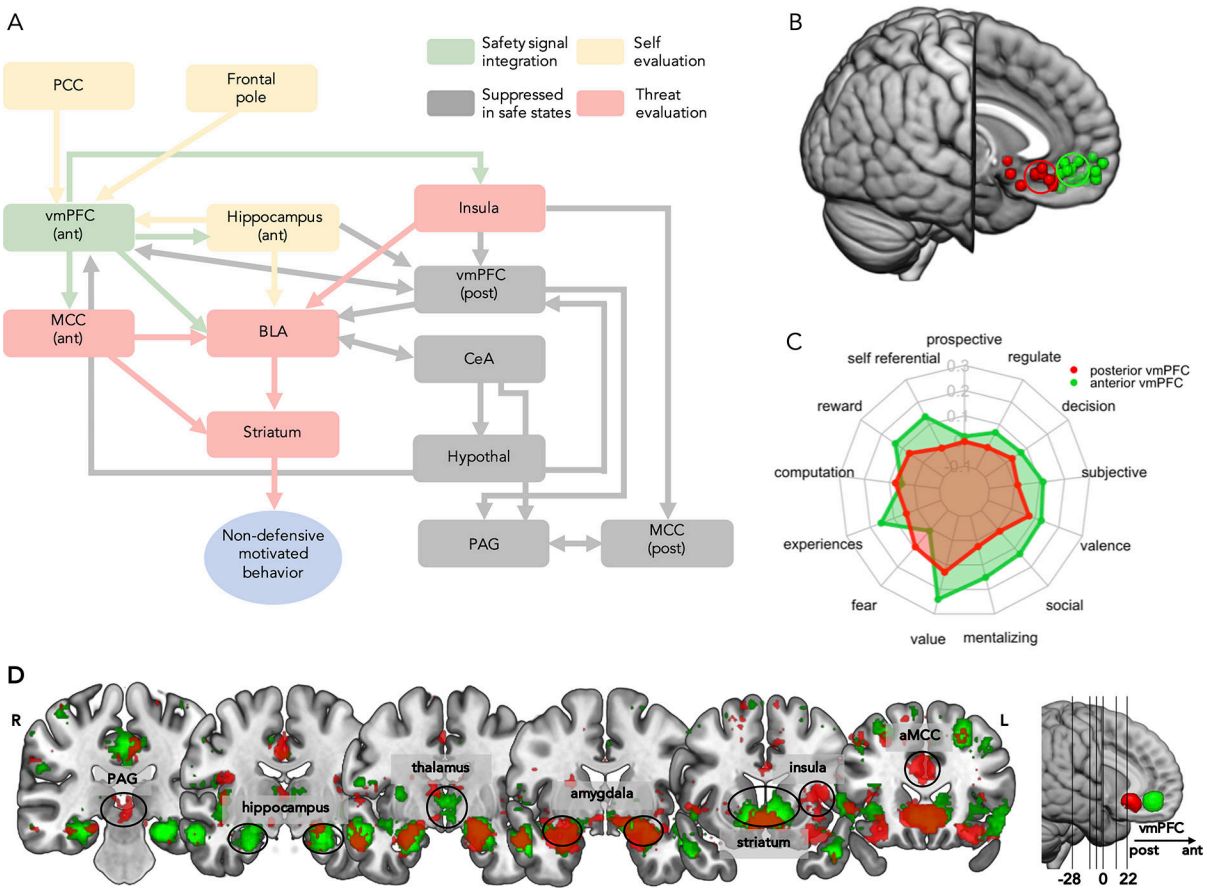

Figure 3. Safety Decision Neural Model and evidence supporting a posterior-to-anterior threat-to-safety distinction.

A. Simplified model of the proposed flow of threat- and self-oriented processes from the Safety Decision Model resulting in safety computation. Green illustrates anterior vmPFC safety signal. Regions in yellow (underlying self-oriented evaluations) and red (underlying threat-oriented evaluations) alter responding depending on safety signaling. In low safety contexts these regions are involved in defensive responding whereas in safe contexts they promote non-defensive behavior. Grey regions are part of the canonical defensive circuit that are suppressed when safety is computed. Arrows represent the flow of functional safety pathways, but do not suggest comprehensive functional or anatomical connectivity pathways. **B.** Peak coordinates from representative human functional magnetic resonance imaging studies reporting neural response to safety (green) and threat (red) (see Supplemental Table S1 for study details). Posterior-vmPFC (red) and anterior vmPFC (green) seeds are overlayed as circles. **C.** Meta-analytic decoding with Neurosynth. Red and green radar bars represent correlation strength between key words representing components of the Safety Decision Model and the anterior (green) and posterior vmPFC (red) ROIs. **D.** Meta-analytic coactivations from Neurosynth including 14371 studies for ROIs in the anterior (green; x=−2, y=46, z=−10) and posterior (red; x=0, y=26, z=−12) vmPFC. vmPFC=ventromedial prefrontal cortex, BLA=basolateral amygdala, CeA=central

amygdala, PAG=periaqueductal gray, MCC= midcingulate cortex, hypothal= hypothalamus, ant=anterior, post=posterior.