nature neuroscience

SUPPLEMENTARY INFORMATION

In the format provided by the authors and unedited.

Viewpoints: Approaches to defining and investigating fear

Interviewed by Dean Mobbs[1], Ralph Adolphs[1], Michael S. Fanselow[2], Lisa Feldman Barrett[3,4], Joseph E. LeDoux[5,6,7], Kerry Ressler[8,9] and Kay M. Tye[10]

Supplementary Note 1

Fear: The Awareness that You Are in Harm's Way

Joseph LeDoux

A Snapshot of My view of Fear

Fear is a conscious experience in which you come to believe that you are about to be harmed (LeDoux, 2015, 2017, 2019; LeDoux and Pine, 2016; LeDoux and Brown, 2017; LeDoux and Hofmann, 2018; LeDoux et al, 2018). Such a subjective state of inner awareness, of phenomenological consciousness (LeDoux, 2019; Brown et al, 2019), can result from the presence of an innate or learned threat, but also by misattributing danger to a benign stimulus or situation, or also by imagining the possibility of harm in the near or distal future, even when such a possibility is improbable or even physically impossible. All that is necessary is your expectation or prediction that this is the case, which occurs when a mental model of fear is pattern-completed using your 'fear schema'--the unique collection of memories about threat, fear, danger, that you have accumulated throughout life. But if you don't apprehend that it is you that is to be harmed by the threat, you cannot experience fear—no self, no fear. Your 'self schema,' your repository of 'you,' must therefore also be involved in the mental model for the experience of fear to result.

Some of the factors that contribute to the assembly of this conscious state include nonconscious representations of perceptual, mnemonic, and/or conceptual information in temporary working memory. Additionally, in some, but not all, instances of fear, brain circuits implicated in processing threat and controlling defensive responses to threat (such as the amygdala, extended amygdala, hypothalamus, and periaqueductal gray circuits) will be activated. These initiate brain arousal and trigger body responses (behaviors and supporting physiological changes in body homeostasis) that result in feedback to the brain in the form of somatic and visceral sensations and circulating hormones. The net result is a colaition of nonconscious cortical and subcortical states that can be monitored by working memory and can contribute to the pattern completion the one's fear and self-schema, and thus a momentary mental model of fear. The resulting integrated representation is thus tailored not only to the situation, but also to the individual, and constitutes the cognitive foundation of a conscious fear experience.

To understand fear, we thus have to understand consciousness. Fortunately, the science of consciousness is a vibrant and thriving area of research (Dehaene et al, 2017). Although this field has focused on perceptual awareness, and has not paid much attention to fear and other emotions, approaching emotional experiences in this context is a very promising avenue for scientific investigations of the phenomenology of fear—the essence of what it is like to be afraid (LeDoux and Brown, 2017; LeDoux et al, 2018; Brown et al, 2019; Taschereau-Dumouchel et al, 2018; LeDoux, 2019).

My View of Fear in Relation to Modern Studies of the Emotional Brain.

Emotions are traditionally thought of as mental states that cause certain behavioral and physiological reactions. Darwin put this notion in an evolutionary context, arguing, for example, that fear is an inner mental state inherited from our mammalian ancestors, and

is outwardly visible in the form of behavioral responses (Darwin, 1872). In the twentieth century, Darwin's perspective guided the search for brain loci of specific emotions using brain lesions (Cannon, 1929; Bard, 1928) and electrical stimulation techniques (Hess, 1954; Hilton and Zbrozyna, 1963, Panksepp, 1980) and shaped so-called basic emotions psychological theories of Tomkins (1962), Izard (1971), and Ekman (1972). Panksepp (1998) fused these approaches, arguing that innate circuits in the limbic system (MacLean, 1970) give rise to emotional experiences in mammals, including humans.

Much contemporary research on the search for emotional mechanisms in the human brain has used the basic emotions approach in conjunction with functional imaging. This empirically productive approach, though, is founded on the assumption that the subjective experience of fear and objective responses often measured all depend directly on innate fear circuits inherited from animals. There are numerous issues with this assumption, as reviewed elsewhere (LeDoux 2014, 2015, LeDoux and Pine, 2017; LeDoux, 2017; LeDoux et al, 2018; LeDoux, 2019). One key example is that humans can respond physiologically to subliminal threats that activate brain areas like the amygdala but that do not elicit reportable feelings of fear).

An alterative and parallel area of research came out of the behaviorist tradition (Watson, 1925; Skinner, 1938; Hull, 1943), which marginalized the mental state view of emotion, but retained the mental state terminology to describe the influences of learning on behavior (Miller, 1951; Mowrer,1951; Rescorla and Solomon, 1967; McAllister and McAllister, 1971; Bolles, 1972). For example, fear came to mean a relation between stimuli and responses in light of a history of aversive reinforcement and/or punishment, not a subjective experience. This approach, which fostered much of the work I and others did on the brain mechanisms of so-called fear conditioning, had the advantage of avoiding anthropomorphic assumptions in animals, but at the cost of ignoring human subjective experience.

In the 1960s and 70s, the work of Schachter and Singer (1962), and Mandler (1975), launched an alternative and currently popular tradition in which emotions such as fear are treated as cognitive states resulting from evaluations/appraisals of, and attributions about, the environmental context in which objective behavioral and physiological responses occur (Frijda, 1986; Scherer, 1984; Ortony and Clore, 1989; Barrett and Russell, 2015; Barrett, 2017). This approach to human emotions acknowledges that fear and other subjective experiences are significant factors in human life, and treats them as emerging from cognitive underpinnings. As Matthew Lieberman recently noted, "emotion is emotional experience" (Lieberman, 2019).

The Back Story About My Approach

My approach to emotions has long been to integrate evolutionary and behaviorist traditions into a cognitive framework (LeDoux, 1984, 1987; 1994, 1996, 2008, 2012, 2014, 2015 2017; LeDoux and Pine, 2016; LeDoux and Brown, 2017; LeDoux et al, 2018; Brown et al, 2019; LeDoux, 2019). Specifically, I have argued for a model in which behavioral and physiological responses in biologically or socially challenging situations involve innate (mostly subcortical) circuits inherited from animal ancestors (consistent with basic emotions theory) and that operate nonconsciously (roughly consistent with behaviorist views), while the conscious experience of emotion is a product of cognitive cortical circuits that interpret the meaning of on-going events in light

of the physical and social context (consistent with cognitive theories). In this model, the internal and external consequences of the response elicited by the subcortical states are part of what the cognitive systems evaluate in the assembly of emotional experiences.

Throughout my years of empirical work on emotion, which began in the early 1980s, I have focused how the role of subcortical circuits in controlled behavioral and physiological responses in dangerous situations. I built upon the elegant behavioral work on Pavlovian 'fear' conditioning, a staple of the behaviorist toolbox, by Robert and Caroline Blanchard (Blanchard and Blanchard, 1969, 1972), and by Robert Bolles and his students, especially Mark Bouton and Michael Fanselow (Bouton and Bolles, 1980; Bolles and Fanselow 1980). As a result of this rigorous, yet simple, method, by the early 1990s, studies by me (LeDoux, 1987, 1992), Bruce Kapp (1992) and Michael Davis (1992) had provided compelling evidence that specific circuits within the amygdala play essential and specific roles in the acquisition and storage of the associations required for 'fear' conditioning, allowing the conditioned stimulus to control behavioral and physiological outputs. Through my long-standing collaboration with Liz Phelps, the basic findings I pursed in animals (LeDoux, 2000) were extended to humans (Phelps and LeDoux, 2005; LeDoux and Phelps, 2008).

For many in the animal fear conditioning field, especially for those trained in the behaviorist tradition, consciousness was not part of the program. When they talked about rats "frozen in fear," or about "using freezing as a measure of fear," they were not talking about the everyday notion of fear as a subjective conscious state. Conscious fear was irrelevant. Although I was working in this context, I wanted conscious fear to be part of the discussion.

In the field of memory, implicit (non-conscious) and explicit (consciously accessible) aspect of memory had come to be recognized (Squire, 1987). Perhaps a similar distinction could apply to emotional memory (LeDoux, 1993, 1996). The amygdala could be thought of as controlling behavioral and physiological responses to threats implicitly (consistent with behaviorist approaches), while conscious fear experiences might depend on processing by neocortical circuits (LeDoux, 1984, 1987, 1996) (consistent with cognitive approaches). And I proposed that part of what the cognitive circuits processed was the consequences of implicit processing by the amygdala.

Why was I so interested in consciousness? My PhD work was on consciousness in split-brain patients (LeDoux et al, 1977; Gazzaniga and LeDoux, 1978). Although I did not personally research consciousness after that, I maintained a strong interest in conscious fear experiences throughout my studies of fear conditioning. Working memory (Baddeley, 1992) had arisen as a possible way for non-conscious information processing (Kihlstrom, 1987) to be made conscious (Baars, 1988), and specific regions of prefrontal cortex was emerging as crucial for working memory in humans and other primates (Goldman-Rakic, 1987; Fuster, 1989). Conscious fear, I therefore argued, might be the result of cognitive circuits involving prefrontal working memory circuits (LeDoux, 1996, 2002, 2008). More recently, the prefrontal networks involved in consciousness have been elaborated on, with the frontal pole, a unique region possessed only by humans (Kochilen), also playing a special role (Lau and Rosenthal, 2011; Brown et al, 2019; LeDoux, 2019).

The implicit processing vs conscious fear distinction didn't stick. The behaviorally-oriented folks didn't see the need for it since consciousness was, in a sense, irrelevant to them. But there were other factors at play. Some scientists were vocally advocating for a role of the amygdala in conscious fear (Panksepp, 1980, 1998). Strict adherence to behaviorists principles was declining, and views such as Pankepp's drew less scrutiny. Moreover, many from biological backgrounds were beginning to study brain and behavior with little awareness of, or interest in, the conceptual debates.

All of this combined to make scientific conceptions of fear looser and looser, more colloquial--fear simply became fear. Scientists began to freely talk about conditioning in rats as the way to understand human fear (i.e. conscious fear experience) and to find treatments for uncontrollable conscious feelings of fear-- the pharmaceutical industry had done this since the 1960s, but the success of fear conditioning attracted more and more academics. The press and lay public, needless to say, always thought fear was fear. The idea that the amygdala was the brain's fear center thus took off, divorced from any distinction between conscious and non-conscious aspects, and is today a cultural meme.

I contributed to the confusion by not always being clear about what I meant when I used the word fear, despite the fact that I had laid out the idea of the amygdala being part of an implicit (non-conscious) processing system, with neocortical circuits contributing to conscious fear, many times. I am in a sense trying to make amends for not being clearer. I have therefore been using the term" defensive survival circuit" when discussing the amygdala's role in detecting and responding to threats, saving "fear" as referent of the actual experience one has, and knows they are having, when in harm's way (LeDoux, 2012, 2014, 2015, 2017, 2019).

Some feel that my raising a red flag about how we talk about fear will have the effect of undermining animal research (Fanselow and Pennington, 2018). I disagree. I think such discussions will clarify what we can and can't learn from animal research—clearly we can learn a lot, but maybe not everything we need to know. And we should reach conclusions about all this by considering the issues, rather than presuming we know what is going on. This special topic discussion is thus very useful for defining the direction, and the health, of our field, in the years to come.

References

Baars BJ (1988) A Cognitive Theory of Consciousness. New York: Cambridge University Press.

Baddeley A (1992) Working memory. Science 255:556-559.

Bard P (1928) A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. American Journal of Physiology 84:490-515.

Barrett LF (2017) How emotions are made. New York: Houghton Mifflin Harcourt.

Barrett LF, Russell JA (eds.) (2015) The psychological construction of emotion. New York: Guilford Press.

Blanchard DC, Blanchard RJ (1972) Innate and conditioned reactions to threat in rats with amygdaloid lesions. J Comp Physiol Psych 81:281-290.

Blanchard RJ, Blanchard DC (1969) Crouching as an index of fear. J Comp Physiol Psych 67:370-375.

Bolles RC (1972) The avoidance learning problem. In: The psychology of learning and motivation, vol. 6 (Bower, G. H., ed), pp 97-145 New York: Academic Press.

Bolles RC, Fanselow MS (1980) A perceptual-defensive-recuperative model of fear and pain. Behavioral and Brain Sciences 3:291-323.

Bouton ME, Bolles RC (1980) Conditioned fear assessed by freezing and by the suppression of three different baselines. Animal Learning and Behavior 8:429-434.

Brown R, Lau H, LeDoux JE (2019) The Understaing the Higher-Theory Approach to Consciosness. Trends in Cognitive Science (in press).

Cannon WB (1929) Bodily changes in pain, hunger, fear, and rage. New York: Appleton.

Darwin C (1872) The expression of the emotions in man and animals. London: Fontana Press.

Davis M (1992) The role of the amygdala in conditioned fear. In: The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction (Aggleton, J. P., ed), pp 255-306 NY: Wiley-Liss, Inc.

Dehaene S, Lau H, Kouider S (2017) What is consciousness, and could machines have it? Science 358:486-492.

Ekman P (1972) Universals and Cultural Differences in Facial Expressions of Emotions. In: Nebraska Symposium on Motivation, 1971 (Cole, J., ed), pp 207-283 Lincoln, Nebraska: University of Nebraska Press.

Fanselow MS, Pennington ZT (2018) A return to the psychiatric dark ages with a two-system framework for fear. Behav Res Ther 100:24-29. PMC5794606.

Frijda N (1986) The Emotions. Cambridge: Cambridge University Press.

Fuster JM (1989) The prefrontal cortex. New York: Raven.

Gazzaniga MS, LeDoux JE (1978) The Integrated Mind. New York: Plenum.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: Handbook of Physiology Section 1: The Nervous System Vol V, Higher Functions of the Brain (Plum, F., ed), pp 373-418 Bethesda: American Physiological Society.

Hess WR (1954) Functional organization of the diencephalon. New York: Grune and Stratton.
Hilton SM, Zbrozyna AW (1963) Amydaloid region for defense reactions and its efferent pathway to the brainstem. J Physiol 165:160-173.
Hull CL (1943) Principles of behavior. New York: Appleton-Century-Crofts.
Izard CE (1971) The Face of Emotion. New York: Appleton-Century-Crofts.
Kapp BS, Whalen PJ, Supple WF, Pascoe JP (1992) Amygdaloid contributions to conditioned arousal and sensory information processing. In: The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction (Aggleton, J. P., ed), pp 229-254 New York: Wiley-Liss.
Kihlstrom JF (1987) The Cognitive Unconscious. Science 237:1445-1452.
Koechlin E (2011) Frontal pole function: what is specifically human? Trends Cogn Sci 15:241; author reply 243.
Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness. Trends Cogn Sci 15:365-373.
LeDoux JE (1984) Cognition and emotion: processing functions and brain systems. In: Handbook of Cognitive Neuroscience (Gazzaniga, M. S., ed), pp 357-368 New York: Plenum Publishing Corp.
LeDoux JE (1987) Emotion. In: Handbook of Physiology 1: The Nervous System Vol V, Higher Functions of the Brain (Plum, F., ed), pp 419-459 Bethesda: American Physiological Society.
LeDoux JE (1992) Emotion and the Amygdala. In: The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction (Aggleton, J. P., ed), pp 339-351 New York: Wiley-Liss, Inc.
LeDoux JE (1993) Emotional memory systems in the brain. Behavioral Brain Research 58:69-79.
LeDoux JE (1994) Emotion, memory and the brain. Sci Am 270:50-57.
LeDoux JE (1996) The Emotional Brain. New York: Simon and Schuster.
LeDoux JE (2000) Emotion circuits in the brain. Annu Rev Neurosci 23:155-184.
LeDoux JE (2008) Emotional colouration of consciousness: how feelings come about. In: Frontiers of Consciousness: Chichele Lectures (Weiskrantz, L. and Davies, M., eds), pp 69-130 Oxford: Oxford University Press.
LeDoux J (2012) Rethinking the emotional brain. Neuron 73:653-676.
LeDoux JE (2014) Coming to terms with fear. Proc Natl Acad Sci U S A 111:2871-2878.
LeDoux JE (2015) Anxious: Using the brain to understand and treat fear and anxiety. New York: Viking.
LeDoux JE (2017) Semantics, Surplus Meaning, and the Science of Fear. Trends Cogn Sci 21:303-306.
LeDoux JE (2019) The Deep History of Ourselves: The Four-Billion Year Story of How We Got Conscious Brains. New York, Viking.
LeDoux JE, Brown R (2017) A higher-order theory of emotional consciousness. Proc Natl Acad Sci U S A 114:E2016-E2025.
LeDoux J, Brown R, Pine DS, Hofmann SG (2018) Know Thyself: Well-Being and Subjective Experience. In: Cerebrum New York: The Dana Foundation.
LeDoux JE, Hofmann SG (2018) The subjective experience of emotion: a fearful view. Current Opinion in Behavioral Sciences 19:67-72.

LeDoux JE, Phelps EA (2008) Emotional Networks in the Brain. In: Handbook of Emotions (Lewis, M. et al., eds), pp 159-179 New York: Guilford Press.
LeDoux JE, Pine DS (2016) Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework. Am J Psychiatry 173:1083-1093.
LeDoux JE, Wilson DH, Gazzaniga MS (1977) A divided mind: observations on the conscious properties of the separated hemispheres. Annual Neurology 2:417-421.
Lieberman M (2019) Boo! The consciousness problem in emotion. Cognition and Emotion 33: 24-30.
MacLean PD (1970) The triune brain, emotion and scientific bias. In: The Neurosciences: Second Study Program (Schmitt, F. O., ed), pp 336-349 New York: Rockefeller University Press.
Mandler G (1975) Mind and Emotion. New York: Wiley.
McAllister WR, McAllister DE (1971) Behavioral measurement of conditioned fear. In: Aversive Conditioning and Learning (Brush, F. R., ed), pp 105-179 New York: Academic Press.
Miller NE (1951) Learnable drives and rewards. In: Handbook of Experimental Psychology (Stevens, S. S., ed), pp 435-472 New York: Wiley.
Mowrer OH (1951) Two-factor learning theory: summary and comment. Psychol Rev 58:350-354.
Ortony A, Clore GL (1989) Emotions, moods, and conscious awareness. Cognition and Emotion 3:125-137.
Panksepp J (1980) Hypothalamic integration of behavior: rewards, punishments, and related psychological processes. In: Handbook of the Hypothalamus Vol 3, Behavioral Studies of the Hypothalamus (Morgane, P. J. and Panksepp, J., eds), pp 289-431 New York: Marcel Dekker.
Panksepp J (1998) Affective Neuroscience. New York: Oxford U. Press.
Papez JW (1937) A proposed mechanism of emotion. Archives of Neurology and Psychiatry 79:217-224.
Phelps EA, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175-187.
Rescorla RA, Solomon RL (1967) Two process learning theory: Relationships between Pavlovian conditioning and instrumental learning. Psychological Review 74:151-182.
Schachter S, Singer JE (1962) Cognitive, social, and physiological determinants of emotional state. Psychol Rev 69:379-399.
Scherer KR (1984) Emotion as a multicomponent process: A model and some cross-cultural data. Review of Personality and Social Psychology 5:37-63.
Skinner BF (1938) The behavior of organisms: An experimental analysis. New York: Appleton-Century-Crofts.
Squire LR (1987) Memory and the Brain. New York, Oxford University Press.
Taschereau-Dumouchel V, Cortese A, Chiba T, Knotts JD, Kawato M, Lau H (2018) Towards an unconscious neural reinforcement intervention for common fears. Proc Natl Acad Sci U S A 115:3470-3475. PMC5879705.
Tomkins SS (1962) Affect, Imagery, Consciousness. New York: Springer.
Watson JB (1925) Behaviorism. New York: W.W. Norton.

Constructing Fear

Lisa Feldman Barrett[1,2,3]

[1] Department of Psychology, Northeastern University
[2] Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School
[3] Department of Radiology, Massachusetts General Hospital

The word "fear" refers to a *category* of events -- a grouping of actions and experiences that are similar to one another in some way. Fear is a psychological category. Instances of fear, like instances of any category, can be described as having a set of *features*. The features belonging to any instance of fear in an organism include some set of changes to the physical state of that organism (e.g., in a mammal, this would include autonomic nervous system changes, neurochemical changes, motor actions, etc.) and the brain states that control them, which also represent the resulting sensory changes (in the body) and some set of sensory events (in the world). Depending on the nature of the brain and body of the organism in question, instances of fear might also include psychological features, such as conscious feelings of pleasantness or unpleasantness and activation or quiescence (together called "affect") and a conscious experience of the surrounding world (e.g., an object looming overhead). In humans, instances of fear might sometimes include self-awareness of their experience of the world (such as whether the situation feels safe or threatening, novel or familiar, etc., often called "appraisals"), self-awareness of interoceptive changes (such as an awareness of one's heart beating), self-awareness of what one's sensory and motor changes are "for" (referred to as a "function" or "goal") and self-awareness of oneself as being in an emotional state (called an "experience of emotion"). All of these features extend over some temporal window and occur in a particular context.

A number of prominent research programs attempt to understand fear by creating taxonomies of categories, drawing a boundary around the events that qualify as a certain type of fear, separating them from events that are presumed to be some other type of fear. They then proceed to search for the biological and/or computational basis [e.g., 1-3] of these various types of fear. The existing taxonomies differ in various ways, but they all share a common assumption: a mammalian brain contains some number of innate, dedicated circuits – fear circuits – each of which triggers a specific action for a specific type of fear. From this perspective, each type of fear (usually associated with a specific behavior) is thought to be a specific adaptation associated with a specific state caused by specific neural circuitry. This approach is grounded in an ontological commitment that fear and its circuits are *species-general*, causing fundamentally conserved states across flies, rats and humans, and associated in each with species-specific defensive actions. Ontological commitments are axiomatic assumptions about what must be true in the world for a scientific theory to be true: they are the truth conditions of a scientific theory.

My scientific investigations into the nature of emotion are guided by several observations that fundamentally question some of these ontological commitments, ultimately leading me to ask different questions about the nature of fear. The first observation is that, in humans, instances of fear are *highly* context-dependent and variable in the physical features that scientists typically measure, such as associated autonomic nervous system changes [4] and expressive movements [5]. The situated variation observed in humans is also observed in many other species that are studied by behavioral ecologists; these species have varying ecological niches, revealing that defensive actions necessarily depend on *situation-specific* considerations about an animal's state and the state of the environment. Recent research from evolutionary robotics reinforces these observations, suggesting that defensive behaviors are shaped by the ways in which predators and prey co-evolve via their interactions with one another and with environmental variations (for discussion, see [6]). By acknowledging and attempting to explain this situated variation in instances of fear, my scientific approach is similar to functional

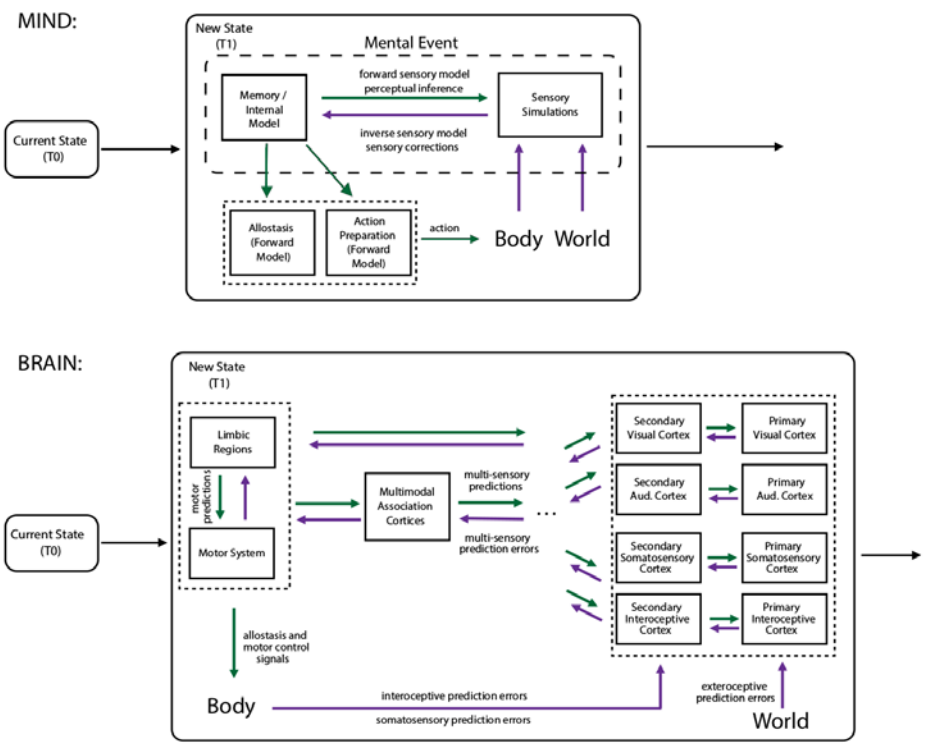
approaches [e.g., 7], descriptive appraisal approaches [e.g., 8], psychological construction approaches [e.g., 9], and the survival circuits approach [e.g., 10] to studying the nature of fear.

Situational factors not only influence *which* defensive action is executed, but also *how* any given action is implemented neurally. Defensive actions are purposeful motor actions (even those that appear to be reflexive; again, see [6] for discussion). This makes it highly unlikely that defensive behaviors are triggered by fixed, preprogrammed circuits, as proposed by existing taxonomies of fear, and more likely that they are assembled compositionally from *widely distributed populations* embedded in synchronized neural activity. "Survival circuits" [10] are part of these neural assemblies, which stretch from association cortices (such as premotor cortex, which is important for action planning and sensory predictions) and primary motor cortex (important for execution of motor actions) all the way down to the motor neurons in the ventral horn of the spinal cord, which contains the modules that direct a specific pattern of muscle fiber activity and joint angles.

The neural architecture of the motor system is structured so that a *single* intention, such as an intention to run (i.e., an action goal), can be translated into *many* patterns of muscle movements (i.e., a one-to many relationship). This translation of a single motor intention into many different patterns of muscle movements is called *motor equivalence*, because one behavior can be implemented by more than one plan of muscle contractions and joint movements, each with some prior probability of being functional in a given situation or context. In biology, motor equivalence is described as *degeneracy*: the ability of structurally different elements to perform the same function [11]. A variety of biological systems display degeneracy [11], and there is growing evidence that degeneracy is a property of brain function at multiple scales [e.g., 12-13], allowing different distributed assemblies of neurons to represent the same information in a way that is generative and robust to damage. Brain imaging studies in humans suggest degenerate patterns for the category fear [14].

Accumulating evidence also suggests that parts of the motor system are routinely involved in perception [e.g., 15, 16] and cognition [e.g., 17], a finding that has been observed for other large-scale brain networks. This is consistent with the hypothesis that brain networks may be domain-general [18] (even the same neuron may participate in many functions, or *many-to-one* mappings).

Empirical observations like these, integrated with advances in cognitive science that have shed light on the nature of concepts and categories, as well as the emerging evidence that brains guide action and construct experiences via *predictive* processes [see 19], led me to propose the theory of constructed emotion, which integrates hypotheses and research findings from neuroconstruction, psychological construction, social construction and rational constructivism [e.g., 20-24]. Within this framework, the current roster of fear circuits or survival circuits [in 1, 2, 10] are a small part of a much richer, more flexible, context-sensitive complex system for assembling and controlling all mental events (Figure 1). Comparably complex combinatorial systems in biology exist for genes, the immune system, the retina, and the autonomic nervous system.



The theory of constructed emotion proposes that the circuitry for constructing an instance of fear is assembled by a brain as needed, via the interplay of evolved ingredients (25), some of which are *species-general* and others which are *species-specific*.

The circuitry controlling survival-related actions such as those used for defense, foraging, reproduction, thermoregulation, and fluid intake – *survival circuits* – can be considered one type of *ingredient* in constructing an instance of fear. But understanding the brain basis of fear requires more than just the careful mapping of the circuitry that supports survival-related actions: it also requires understanding the neurobiological underpinnings of how these actions and their sensory consequences are made meaningful in a brain.

My core hypothesis is that a brain is continually running an internal model of an animal's body, its niche and the relation between the two. The model is generative, meaning that past experiences are recombined in novel ways as they are reimplemented (i.e., remembered). Actions and experiences are constructed by the brain's internal model in its effort to navigate its body in the world. Efficient navigation entails predictively controlling the autonomic nervous system, the endocrine system and other systems of an animal's internal milieu, to anticipate the needs of the body and to meet those needs in the service of future motor actions and learning, a process called *allostasis* [26]. The goal is not to minimize energy expenditure but to maximize energy efficiency [27]. The model contains plans for upcoming motor actions and the allostatic changes in the internal milieu that support those actions. The model also *predicts*, or infers, the sensory consequences that are expected to result from those movements, as well as the *causes* of those sensory changes. I hypothesize that experiences of affect and of the world emerge from those perceptual inferences (again, see Figure 1). Prediction signals can be regarded as anticipatory causal explanations for sensations and actions that are mapped, inversely, to those sensations and action [for discussion, see 19]. A brain's internal model is continuously maintained, updated or refined by comparisons with incoming sensory information from the body and the world, referred to as processing *prediction error signals* or, simply, *learning*. The hypothesis is not that animals are deliberately remembering and deciding which objects are in their niche and which can safely be ignored, but that actions and experiences in the future are automatically conditioned on and filtered through the past.

A further hypothesis is that efficient energy regulation and its affective consequences are core features of all mental events, not just features of emotional events like fear [18, 19, 28]. Behavioral choices (including those that allow an animal to avoid or escape a predator) can be regarded as *economic choices* about energetics and other biological resources. Learning vs. ignoring unpredicted sensory inputs is also an economic choice. Indeed, all potential movements, as well as learning, have an energy cost, and an animal's brain must weigh these costs against potential rewards and revenues in the service of balancing its global energy budget. Economic choices about actions and experience, therefore, are necessarily influenced by a number of situation-specific considerations about an animal's current physical state and the state of the environment (i.e., context). Via its internal model, along with deviations from that model (i.e., prediction error signals), an animal's brain "decides" whether and how to spend energy resources to move (i.e., execute an action). Correspondingly, the brain must also decide whether and how to invest those resources to learn any unanticipated sensory inputs to improve the internal model of its body within its niche, thereby enhancing future predictions and allowing the animal to survive, thrive, and reproduce.

This set of hypotheses integrates recent findings about predictive brain dynamics with anatomical models of information flow in the brain. These hypotheses implicate neurons from a variety of brain regions in the neural basis of fear, including (in mammals) the cerebral cortex, the hippocampus and medial temporal lobe, the cerebellum, and the striatum, along with other subcortical regions that make up the brain's dopaminergic systems [for discussions, see 19, 22, 28, 29]. They also implicate a variety of neurotransmitters that impact energy regulation, as well as neurochemicals whose impact on energy regulation may be underappreciated, such as serotonin (e.g., [30]).

Furthermore, I hypothesize that the prediction signals which constitute a brain's internal model can be described, psychologically, as ad hoc concepts [6, 21-23]. An animal's brain is constantly trying to solve a *reverse inference* problem: it must actively infer the causes of sensory inputs when all it has access to are the effects (i.e., the information coming from the retina, cochlea, and other sensory systems of the body). A brain

solves this reverse inference problem by constructing multiple prediction signals [e.g., 31] using past experiences that were similar in some way to present conditions. Recall that a group of events with some similar features is a category, and the mental representation of a category is a *concept*. Following this line of reasoning yields the hypothesis that a brain is constantly faced with a category construction problem, which it solves by assembling ad hoc concepts (i.e., prediction signals or inferences) that represent the possible causal relationships between events in the world and in the body, as they are right now, and their sensory and motor consequences in a moment from now (see Figure 1). Consistent with recent developments in cognitive science, concepts, in this view, are embodied, whole brain representations, and not mere abstractions corresponding to propositional or semantic knowledge. Once the errors of prediction are sufficiently minimized, the incoming sensory inputs and associated motor actions are *categorized* by the ad hoc concept, making them meaningful.

This line of reasoning, when applied to understanding the nature of fear, yields the following hypothesis: an instance of fear is assembled when the brain remembers similar instances from the past and combines them in a generative way, on the fly, to create an ad hoc concept of fear. This concept of fear functions as an internal model (sometimes called a forward model) that contains a behavioral intention -- a cascade of potential visceromotor and motor patterns, descending through subcortical and brainstem nuclei, sometimes (but not always) resulting in a defensive behavior. The forward model also contains *efferent copies* of the behavioral intention, which function as prediction signals that simulate the expected sensory consequences of the expected internal and skeletomotor movements (also called a corollary discharge). Once prediction errors are sufficiently minimized, the sensory inputs, visceromotor changes and motor actions are categorized as an instance of fear, realized as an animal's experience of its internal milieu (referred to as affect [32]), as well as its experience of the world. Ad hoc concepts for fear are context-dependent, varying with the state of the animal and the current conditions of its environment, suggesting that there is no unified, central state of fear, nor a single concept for fear, but a population of variable ad hoc concepts of fear.

If a brain runs an internal model of an animal's body in its ecological niche, powered by ad hoc concepts, then this begs the million-dollar question: which species have brains that can construct fear concepts to control their survival circuits? The more familiar version of this question is: is fear homologous across species? Can all animals be fearful when faced with a predator? The implicit part of this contentious and hotly debated question is: do non-human animals become afraid *like humans do*? I translate this question into: which non-human species can construct human-like concepts of fear? The answer to this question depends on the architecture of an animal's brain.

All brains categorize [33]. What distinguishes human and non-human animals is not the computational principles that govern prediction and the assembly of psychological events (i.e., the construction of actions and their associated experiences), but the *content* of the ad hoc concepts that are created by those computations [6]. The computational role of most major brain parts remains stable across the vertebrate lineage. What appears to differ among species, because of general brain-scaling functions, is the *type of concepts* that a brain can construct [34]. As information is learned, neural activity propagates along a lamination gradient in the cerebral cortex (in layers 2 and 3 of the cortical sheet), from primary sensory cortices containing smaller neurons with fewer connections to cortices containing progressively larger neurons with more connections, representing shared information with progressively more compressed, more efficient multimodal summaries [35] that are referred to as *abstractions*. The largest neurons, found in association cortices in the front of the brain, integrate across sensory modalities by summarizing their shared information (i.e., the statistical relationships in their patterns of activity), effectively achieving dimensionality reduction. The human brain has expanded association cortices in the frontal lobes, parietal cortex and inferotemporal cortex when compared to other primates, even other great apes, particularly in cortical layers 2 and 3 [35]. This expansion potentially allows for increased information compression and dimensionality reduction during the processing of prediction error (i.e., during learning), suggesting the possibility that human brains may be able to craft multimodal summaries (i.e.,

concepts) characterized by greater degrees of abstraction. Such abstraction allows ad hoc concepts (i.e., prediction signals) to be constructed according to the *functional similarities* of their instances, rather than the statistical regularities of their physical features. That is, human fear is best thought of as an abstract concept, populated with context-dependent instances of fear that have few statistical regularities in their physical features. A neurotypical human brain's capacity for abstraction allows it to impose a similar function on variable changes in the cardiovascular system, in the respiratory system, in facial movements, in skeletomotor movements like freezing and fleeing, and so on, to create ad hoc, human concepts for fear. This implies that no bodily change has an emotional meaning in and of itself. Humans make bodily changes meaningful as emotions by categorizing them as such. Sometimes instances of an emotion (constructed from an abstract concept) may entrain survival circuits, but often they may not. (Likewise, survival circuits may be entrained by ad hoc, abstract concepts that are not constructed by re-establishing prior instances of emotion). Abstract emotion concepts, grounded in functional, rather than physical, similarities, allow a brain to be more generative, giving it enhanced flexibility and inductive capacity. This, combined with the human propensity for social learning, may be why the human niche is so expansive in space and time.

Whether non-human animal brains can construct ad hoc concepts for fear that are rooted in functional rather than physical similarities depends on the architecture of those brains and their capacity for abstraction. A rat's brain, like a human's brain, creates ad hoc concepts using past instances, in the service of controlling the animal's body and creating its immediate experience of the surrounding world, and this concept construction process can involve neural cascades that mobilize survival circuits. Without the same capacity for abstraction as a human brain, however, the rat's concepts may be rooted more in physical similarities between past instances and present conditions, rather than in functional similarities. This hypothesis, if correct, has important implications for what we can learn about human fear by studying non-human animals in laboratories. Typical laboratory settings have intentionally removed the spatial, temporal and biological variation that is inherently present in animals' normal ecological contexts, making it difficult to learn exactly how much abstraction an animal's brain is actually capable of.

Whether human and non-human animals share similar states of fear, then, depends on *the degree of similarity in the ad hoc concepts in human and non-human animal brains*. Are they physically (i.e., biologically) similar or are they functionally similar? Ironically, only a human brain can answer this question. Human scientists have human brains that are capable of impressive feats of abstraction. When a human scientist observes a rat that is immobile in one context, a rat that is retreating in another context, and a rat that is even approaching another animal in yet a third context, the scientist can, each time, construct an abstract, ad hoc concept of fear. These concepts allow the scientist to *infer* that each rat is fearful -- even infer that the rats are in the same state of fear – despite the fact that the observable features differ across instances. Human scientists, with their capacity for abstraction, can even infer similarities when observing different behaviors in different species. Across various observations, the contexts can be different, the brain states of the observed animals can be different, and even the consequent behaviors can be different, but a scientist's brain can nonetheless impose an abstract similarity (e.g., a function, such as seeking protection from a predator), leading them to hypothesize that the same underlying fear state is present on each occasion. In effect, my scientific approach to the nature of fear explains how scientists come to believe that fear is homologous across species, despite observable species-specific differences.

My scientific approach suggests that solving the puzzle of fear may require reconsidering the automatic mental inferences that are embedded in the ontological commitments underlying much of the existing research on fear. Still, the approach I am suggesting in no way diminishes the importance of studying human survival behaviors and their neural assemblies. Nor does it invalidate the importance of studying survival-related behaviors in animal models. Both are necessary ingredients for a full understanding of human fear, even if they are not sufficient.

References

1. Gross, C.T. & N.S. Canteras. (2012). The many paths to fear. *Nature Reviews Neuroscience*,13, 651-8.
2. Fanselow, M.S. (2018). Emotion, motivation and function. *Current Opinion in Behavioral Sciences*, 19, 105-109.
3. Bach, D.R. & P. Dayan (2017). Algorithms for survival: a comparative perspective on emotions. *Nature Reviews Neuroscience*, 18, 311-319.
4. Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., Quigley, K. S., & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, 144(4), 343-393.
5. Barrett, L. F., Adolphs, R., Martinez, A., Marsella, S., & Pollak, S. (commissioned article). Emotional expressions reconsidered: Challenges to inferring emotion in human facial movements. *Psychological Science in the Public Interest.*
6. Barrett, L. F. & Finlay, B. L. (2018). Concepts, goals and the control of survival-related behaviors. *Current Opinion in the Behavioral Sciences, 24*,172-179.
7. Adolphs, R. (2017). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Soc Cogn Affect Neurosci*, 12(1), 24-31.
8. Clore, G. L., & Ortony, A. (2008). Appraisal theories: How cognition shapes affect into emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 628-642). New York, NY: Guilford Press.
9. Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.
10. LeDoux, J. & N.D. Daw. (2018). Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, 19(5), 269-282.
11. Edelman, G.M. & J.A. Gally. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98, 13763-13768.
12. Marder, E.T. & Taylor, A. L. (2011). Multiple models to capture the variability in biological neurons and networks. *Nature Neuroscience*, 14,133-138.
13. Sporns, O. (2011). *Networks of the Brain*. Cambridge MA: MIT press.
14. Clark-Polner, E., Johnson, T., & Barrett, L. F. (2017). Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. *Cerebral Cortex, 27*, 1944-1948.
15. Keck, T., Keller, G. B., Jacobsen, R. I., Eysel, U. T., Bonhoeffer, T., & Hübener, M. (2013). Synaptic Scaling and Homeostatic Plasticity in the Mouse Visual Cortex In Vivo. *Neuron*, 80(2), 327–334.
16. Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C., Carandiniy, M., & Harrisy, K. D. (2018). *Spontaneous behaviors drive multidimensional, brain-wide neural activity*. bioRxiv preprint, doi: http://dx.doi.org/10.1101/306019
17. Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9), 458-470.
18. Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., Quigley, K.S., Dickerson, B. C., & Barrett, L. F. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behavior*, 1, 0069.
19. Hutchinson, J. B., & Barrett, L. F. (2019). The power of predictions: An emerging paradigm for psychological research. *Current Directions in Psychological Science*, 28, 280-291.
20. Barrett, L.F. (2011). Emotions are real. *Emotion*, 12(3), 413-29.
21. Barrett, L.F. (2017). *How emotions are made: The secret life the brain*. New York, NY: Houghton-Mifflin-Harcourt.

22. Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Soc Cogn Affect Neurosci*, 12(1), 1-23.
23. Hoemann, K., Xu, Fei, & Barrett, L. F. (in press). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*.
24. Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nature Human Behavior*, 2, 624–636.
25. Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives in Psychological Science*, 4, 326-339.
26. Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiology & Behavior*, 106(1), 5-15.
27. Sterling, P. & S. Laughlin. (2015). *Principles of neural design*. Cambridge MA: MIT Press.
28. Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16, 419-429.
29. Chanes, L., & Barrett, L. F. (2016). Redefining the role of limbic areas in cortical processing. *Trends in Cognitive Sciences, 20*, 96-106.
30. Namkung, J., Kim, H., & Park, S. (2015). Peripheral serotonin: A new player in systemic energy homeostasis. *Mol. Cells*, 38(12), 1023-1028.
31. Gallivan, J.P., et al. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nat Neurosci*, 19(2), 320-6.
32. Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology, 41*, 167-218.
33. Mareschal, D., Quinn, P. C., Lea, S. E. G. (Eds.) (2010). *The making of human concepts*. New York: Oxford.
34. Workman, A.D., Charvet, C. J., Clancy, B., Darlington, R. B., & Finlay, B. (2013). Modeling transformations of neurodevelopmental sequences across mammalian species. *Journal of Neuroscience*, 33(17), 7368-7383
35. Finlay, B.L. & R. Uchiyama. (2015). Developmental mechanisms channeling cortical evolution. *Trends in Neurosciences*, 38(2), 69-76.

Supplementary Note 3

Investigating fear as a functional state

Ralph Adolphs
HSS 229-77, Caltech, Pasadena, CA 91125
radolphs@caltech.edu
626-395-4486

My approach to emotions is motivated by the observations that Charles Darwin made in his book, "The expression of the emotions in man and animals" [1]. There are patterns of behaviors seen across animal species from which we regularly infer fear. There are environmental circumstances that regularly accompany these behaviors. These observed regularities are indisputable, and noticed by the ethologist and biologist as well as the pet owner. What is disputed is how to explain them.

The layperson's common explanation proceeds by inferring conscious emotional experiences, to humans and animals alike. There are well-known problems with this commonsense explanation, first and foremost the strong resistance, at least among most scientists, against inferring conscious experiences in animals. So what are the alternative explanations? Taken together, there seem to be four reasonable positions to take, which I briefly enumerate here:

1. Deny that the observed fear behaviors and environmental contexts in fact show any regularity that needs to be explained, or that multiple different regularities could be picked out. This position argues that the patterns Charles Darwin noted are in some sense illusory, and just in the eye of the beholder. If I understand her correctly, this is, at least in part, Lisa Barrett's view [2, 3]. While it seems obviously wrong to me in the case of fear, I think it may indeed apply for some other states we call emotions.

2. Bite the bullet and stick with the commonsense view. It is not an obviously wrong view, but it is generally thought to be an unnecessary view, since the same explanation could be made without involving any conscious experiences (see views #3 and #4 below). If I understand him correctly, Joe LeDoux holds a version of this view, except that he thinks conscious experiences should only be attributed to humans and therefore only humans have emotions [4, 5]. My own view here is to leave this option #2 open as a viable future alternative, but to begin with option #4 below because it avoids the contentious issue of conscious experiences in animals [6].

3. Behaviorism. This view acknowledges the behavioral fear patterns, agrees that they are regularly caused by specific environmental stimuli, but simply links the two without inferring any further internal states, conscious or not. The well-known problem with this view is that it cannot explain the flexibility of emotions and their interaction with other cognitive states. This is not to say that simple reflex-like circuits, and associations built on them, don't constitute a part of an emotion like fear. But Behaviorism cannot account for many other features of emotions, so it is an incomplete position.

4. Functionalism. This is my own view, albeit in a form strongly modified from its historical inception. Classical functionalism is a position in the philosophy of mind, of which there are a number of technical versions. In a nutshell, functionalism says that emotions are states of an organism that are defined by what they do, not by how they are constituted. A common modern analogy is a computer, whose states are also functionally (computationally) defined, and can be implemented in many different physical devices (but I have doubts that the computer analogy is a good one). Like position #2, functionalism posits internal states that provide causal explanations of the observed fear behavior. Like position #3, it does not (at least not necessarily) require attributing conscious experiences. Like position #1, it emphasizes that you need an observer (scientists, in this case) to provide the functional explanation that makes sense of the behavior (and it is possible that more than one functional explanation could explain the same behavior; there need not be a unique one). Unlike position #1, it does not conclude that therefore there are arbitrarily many explanations or that the explanations are not objective.

My own version of a functionalist view of emotions like fear requires an interdisciplinary approach [7]. In particular, it borrows broadly the idea of "levels" of explanation that is attributed to David Marr [8]. At the top level is a functional explanation of behavior in terms of evolution: what does the fear behavior do for the animal, how is it adaptive? This level is required to ground all the others, and it also explains how an emotion can be maladaptive, as happens in anxiety disorders. Lower levels explain more proximally the functions of neural systems, or circuits within them, and how they contribute to the larger behavioral function. (David Marr described three levels of explanation, but there can be as many levels as you like).

In the case of fear, then, we are faced with a collection of behavioral regularities in how animals and humans respond to various threats. As neuroscientists we want to explain these in terms of the systems, structures, circuits and transmitters that are involved. I think we should begin with careful behavioral studies, in the natural environment, to parse fear at the ethological level. I am certain that this will yield varieties of fear, which likely can be related to varieties of threats. Work by Dean Mobbs and Michael Fanselow provides some initial inventory [9, 10]. Importantly, this initial characterization of fear types will require incorporation with our inventory of other cognitive states: we not only need to describe how threats cause behaviors, but also how they cause changes in memory, perception, attention, and other cognitive states.

The neural basis of particular types of fear can then be investigated at many levels, as long as we take care to note the particular functional level under investigation. Thus, the amygdala will certainly play a functional role in fear, but its function will not be to link stimuli to behavior. Instead it will provide a narrower functional piece of the story, explaining how particular sensory information is processed in terms of inputs from, and outputs to, other brain structures. Ditto for the neuropeptide CRF, subtypes of cells in the periaqueductal gray, the saliency network, or completely different structures in an octopus brain: all of these neurobiologically defined entities can play a role in processing fear, but none of them can be identified with fear. Fear isn't located anywhere in the brain, just like time isn't located anywhere in a clock: fear, like timekeeping, is functionally defined. We initially find it based on the broad behavioral function we observe without taking anything apart (the clock keeps time, the animal has fear), and we can then investigate the constituent processes (also functionally defined, but at a narrower level) that

explain this broad function (particular algorithms or computations, and particular physical implementations, depending on the interests of the investigator).

A main remaining challenge is to actually articulate the functions that make a state an emotion (as opposed to some other type of cognitive state), and that make a state a particular emotion (fear, as opposed to anger or disgust or some other emotion). This is a hard problem. David Anderson and I have taken the approach of listing the operating characteristics of emotions [7]. This list of emotion properties could be taken as a functional description at an algorithmic level. It lists properties such as scalability, valence, generalizability and so forth. Specifying particular emotions, or subtypes of them, requires a more fine-grained functional account. Examples would be the varieties of fear that threat imminence theory proposes, or varieties of disgust.

For readers who want a more comprehensive account of my view, I would recommend the following. Begin with [11], a short and broad overview. Then read Chapters 2,3,10 in [7], which provide an in-depth explanation of a functional account of emotion, and compare this view with others. Readers wishing a more formal philosophical account could then read [6]; those wishing a more neurobiological account specifically about fear could read [12].

References

1. Darwin, C. (1872/1965). The Expression of the Emotions in Man and Animals, (Chicago: University of Chicago Press).
2. Barrett, L.F. (2014). The conceptual act theory: a precis. Emotion Review 6, 292-297.
3. Barrett, L.F. (2017). How Emotions are Made: The Secret Life of the Brain, (New York: Houghton Mifflin Harcourt).
4. LeDoux, J.E. (2012). Rethinking the emotional brain. Neuron 73, 653-676.
5. LeDoux, J. (2017). Semantics, surplus meaning, and the science of fear. Trends Cogn Sci 21, 303-306.
6. Adolphs, R., and Andler, D. (2017). Investigating emotions as functional states distinct from feelings. Emotion Review in press.
7. Adolphs, R., and Anderson, D.J. (2018). The Neurobiology of Emotion: A New Synthesis, (Princeton, NJ: Princeton University Press).
8. Marr, D. (1982). Vision: a computational investigation into the human representation and processing of visual information, (New York: W. H. Freeman and Co.).
9. Mobbs, D., Hagan, C., Dalgleish, T., Stilson, B., and Prevost, C. (2015). The ecology of human fear: survival optimization and the nervous system. Frontiers in Neuroscience: Evolutionary Psychology and Neuroscience 9, 55.
10. Fanselow, M. (2018). Emotion, motivation and function. Current Opinion in Behavioral Sciences 19, 105-109.
11. Adolphs, R. (in press). Emotions are functional states that cause feelings and behavior. In The Nature of Emotion, 2nd Edition, R.J. Davidson, A. Shackman, A. Fox and R. Lapate, eds. (New York: Oxford University Press).
12. Adolphs, R. (2013). The Biology of Fear. Current Biology 23, R79-R93.

Supplementary Note 4

Six Requirements for a Definition of Fear

Michael S. Fanselow
Department of Psychology
Department of Psychiatry & Biobehavioral Sciences
University of California, Los Angeles

Fear is a concept and like any concept it requires a set of defining characteristics to make the term scientifically useful. My interest in the concept of fear grows from a broader interest in basic motivational processes. Like Tolman, I see motivation as the process that stirs us to action when a particular demand must be met, behavior serves a purpose (Tolman, 1932). Fear, like hunger and thirst, reflects a biological problem that must be solved promptly to ensure future reproductive success. Fear relates to defensive behavior, in much the same way that hunger relates to feeding and thirst to drinking. A full appreciation of any of these concepts concerns biology in general and at the least spans the subfields of ecology, ethology, physiology, neuroscience and psychology and cannot be limited to any one of them. Fear, hunger and thirst have a similar set of fundamental questions: 1) What are the external and internal signals that call for action? 2) How are the current signals integrated with each other and with past experience? 3) What are the specific measurable behaviors associated with the signals and their integration? But the starting point in the analysis is biological function; all purposive behavior requires some expenditure of energy and that investment must be justified by the benefit it brings to fitness.

There are also translational questions of importance. Given the prediction that 1/3 of all Americans will have an anxiety disorder some time in their life, why are anxiety disorders so prevalent? Some anxiety disorders are characterized by fear but others by extreme panic or a general sense of anxiety. Valuable scientific theories of motivated behavior will provide insight into these translational issues as well.

By posing these 6 questions I have set a high bar for defining the concept of fear. The best theory will set us on a clear path toward answers of each of the questions.

Question 1: What is the biological function of fear?

While defining an adaptive function is not a formal requirement of a theory, I feel that understanding function forces us to place a theory within the context of biology more generally. This requirement helps constrain the theory yet also provides key insights. I use the following functional definition: Fear is a functional Behavior System that generates defensive behaviors which, over phylogenetic history, have protected individuals of that species from predation. As an example, freezing is clearly effective because visual systems are very sensitive to movement and movement is the releasing stimulus for predatory attack. Justifiably, explanations of phenomena in terms of adaptive function are often criticized for being untested, or even untestable, speculation. This criticism is readily parried by two frequent observations. It occurs in prey animals in the presence of their predators and freezing animals are demonstrably less likely to be attacked. An interesting example of this is that pharmacologically suppressing movement, essentially making the prey a super-freezer, decreases the probability of a lethal attack (Herzog & Burghardt, 1974).

In analyzing hunger, George Collier made brilliant insights by deconstructing the behavior into a series of individual sequential problems that needed to be accomplished before

calories could be utilized (e.g., search, procurement, handling, consumption and digestion; Collier et al., 1972). This is a temporally organized sequence; you must search for food before you can procure it. Timberlake (1994) called these stages 'modes' and made considerable strides in understanding conditioning with food reinforcers by recognizing what mode a particular conditioning experiment modelled. In essence, the different modes of feeding describe the predator's behavior when hungry. A major problem in the science of fear is specifying how fear maps into specific behaviors. It was Collier's description of predatory behavior that gave me the insight to develop the predatory, or threat, imminence continuum model of the topography of defensive behaviors (Fanselow & Lester, 1988 see Figure 1). The goal of defense is to thwart predatory behavior but the antipredator behavior that will be successful depends on where the prey is with respect to the sequence of predatory behaviors. I proposed 3 distinct modes of defense where specific behaviors occur according to where the prey places itself with respect to the predator. I called these, ranging from furthest to closest to being consumed, pre-encounter, post-encounter and circa-strike defensive behaviors. This behavior system approach has proven successful, not only for feeding and defense but also sexual behavior (Domjan, 1994). While less developed, similar principles can be applied to thirst (Marwine & Collier, 1979). The functional behavior systems approach acts as a general organizing principle, or metatheory, on how to conceptualize basic motivational processes.

Question 2: What are the external and internal signals that call for action?
This is the question of what causes fear. The question should be recognized as the antecedent semantic of a theory, which is a formal requirement of any empirical theory because it links the concept to the measurable events or manipulations that cause the concept (Bolles, 1975). It tells us what are the independent variables of experiments and how they should influence the level of the concept, in this case the level of fear. It is the cause in a cause-effect relationship. For fear it is the threat of predation and obviously there are a plethora of examples of defensive behaviors being provoked by the presence of a predator. Elsewhere, I have argued that because it is difficult to encode innate recognition of all possible predators in all their possible forms, a unique form of learning, fear conditioning, evolved to allow instantaneous *recognition* of *potential* threats. Conditional fear stimuli produce reactions virtually identical to predators (Fanselow, 2018).

Ideally the antecedent stimulus should be titratable producing orderly changes in indexes of fear. As examples, the speed of the predator is a major determinant of flight initiation distance in Thompsons gazelle (Gunther, 1961). One major benefit of fear conditioning is that it offers several quantifiable parameters to test cause in a highly refined manner. Increasing the current of shock monotonically increases the amount of freezing caused by a conditional stimulus (Young & Fanselow, 1992). Elsewhere I have described how continuous manipulations of shock probability moves behavior from pre-encounter to post-encounter to circa-strike defenses (Fanselow, 1989).

Question 3: How are the current signals integrated with each other and with past experience?
In terms of a formal theory, the answer to this question describes the theory's syntax, which links the antecedent cause with the behavioral consequence (Bolles, 1975). Predatory imminence links environmental threat to defensive behavioral action. Because predatory imminence is multiply determined it results from a synthesis of many stimuli. A predator's identity, distance, direction of movement, and speed all combine to determine the level of predatory imminence (Gunther,

1961). Learning, which is a natural synthesizer of environmental stimuli, plays a key role in determining predatory imminence (Fanselow, 2018a). I view Predatory imminence as a continuous variable. At a certain level it activates a particular behavioral mode (pre-encounter, post-encounter or circa-strike). The mode confines the behavioral repertoire to a limited set of actions. These actions become more intense or probable as predatory imminence increases but only to a point. At some point a specific behavior ceases as the next mode is engaged. Thus predatory imminence links threat to behavior via a piecewise function. By analogy, predatory imminence is the land speed of a car and each gear corresponds to a mode. Engine speed in a particular gear corresponds to the strength (magnitude, probability) of the behavior in that mode. This is illustrated in Figure 2 (Fanselow, 1989). The continuous environmental variable is probability of shock, each mode-relevant behavior is a piece. Although the exact piecewise function has yet to be determined a specific level of predatory imminence produces a specific behavior at a specific level.

Question 4: What are the specific measurable behaviors associated with the signals and their integration?

In terms of empirical theory construction, this question asks for the consequent semantic (Bolles, 1975). It anchors the model in terms of observable actions and is the effect in a cause-effect relationship. Therefore, it is essential that each mode be tied to its own set of behaviors and the last row of Table 1 lists mode-specific behaviors.

Question 5: Why are anxiety disorders so prevalent?

Understanding fear in terms of its biological utility provides an immediate answer to this translational question. The cost of a missed opportunity to defend is far greater than the cost of a false alarm to nonexistent threat; so evolution favors false alarms (Fanselow & Sterlace, 2014). Anxiety disorders, experiencing fear when there is no danger, is essentially a false alarm.

Question 6: What insight does the theory provide to differentiate the states of anxiety, fear and panic?

Each mode is related to an experiential state (see Figure 2 and Table 1). From a biological perspective such states must exist for a reason and I believe the reason is that the states influence cognition in a way to benefit the mode (Fanselow, 2018b). Since Pre-encounter defense occurs when there is no immediate peril, the state affords an opportunity for planning in advance. Anxiety may help focus the organism on how to best plan activities to minimize risk (e.g., better grab that last meal before daylight because there were visually guided predators during the last week). Fear occurs when peril is immediate and post-encounter defense must occur immediately; it's a time when action, not thought, is called for. The experiential state of fear may be the way the amygdala tells the prefrontal cortex shut up and let me takeover with phylogenetically proven actions. The protean movement that occurs during circa-strike defense allows evasion of capture. The autonomic arousal and chaotic actions characteristic of panic may support such ballistic and energetic reactions. Like fear, panic should eliminate a delay in responding that could be caused by trying to think your way out of the situation.

A problem in terminology:

I've used fear in two ways. Sometimes level of fear is used to refer to the overall system and its 3 modes. Other times, fear is synonymous with just the post-encounter mode. Elsewhere, I've referred to **F**ear vs **f**ear (Fanselow, 2018b). The larger dimension across the entire continuum would be better called threat imminence but I often use fear to maintain contact with the target literature. Perhaps the word fear should be restricted to the post-encounter mode.

Conclusions:

Fear, like hunger or memory, is a theoretical construct. Such constructs must be constrained and in the form of questions I've imposed 6 constraints. Three of these constraints, antecedent semantic, consequent semantic and syntax are an absolute requirement of any formal theory in any empirical science. They are what make the theory useful, predictive and testable.

I also imposed a requirement of biological utility. This adds an additional constraint and brings the concept of fear into the realm of biology. It also aids in the development of the structure of the theory by connecting it with the behavior systems framework. Theories are judged by their usefulness in organizing and explaining data. The more data explained the more useful the theory. Understanding the biological function of fear immediately helps explain two of the most significant translational questions about fear. One is why anxiety disorders are so prevalent. Additionally, because behavior systems are organized around different functional modes the model provides a ready distinction between anxiety, fear, and panic.

Figure 1: The Predatory Imminence Continuum Model.

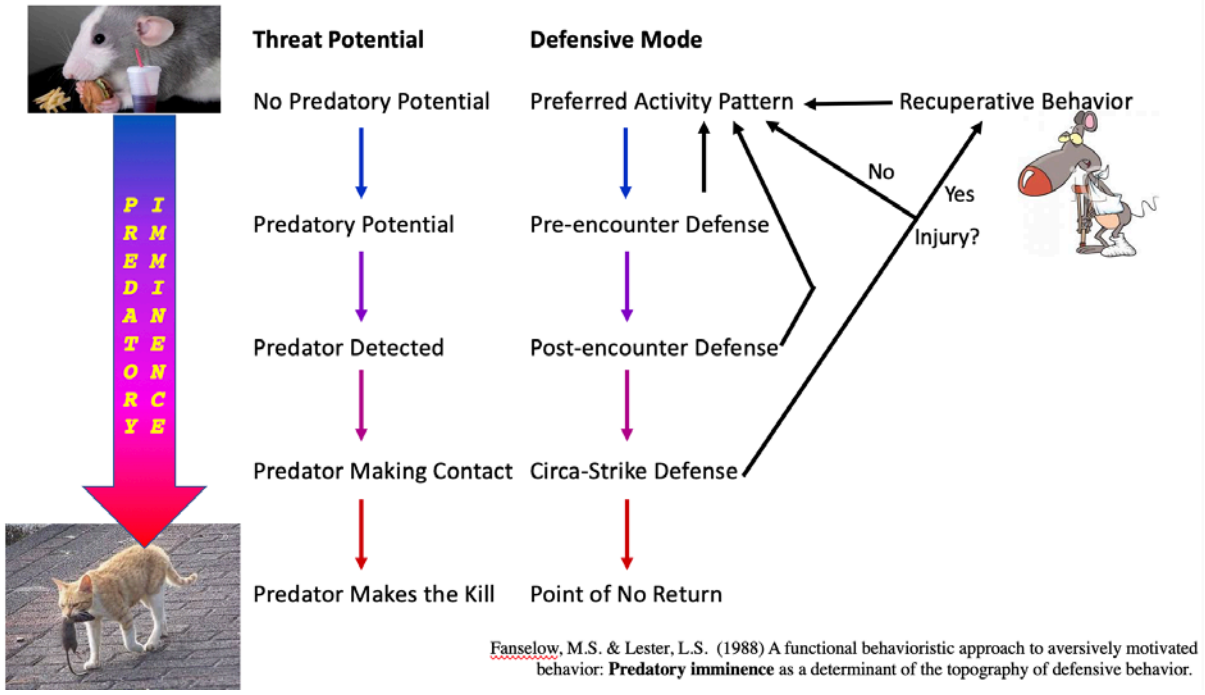


Figure 2: The relationship between predatory imminence and the strength of defensive behavior:

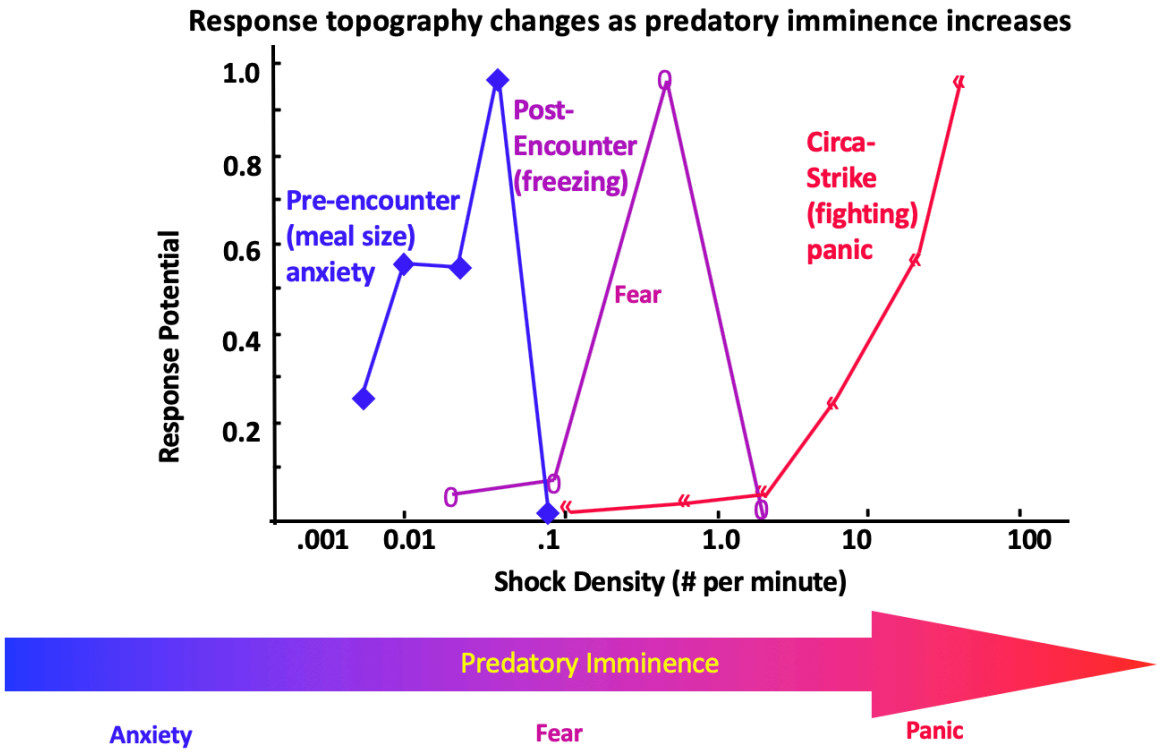


Table 1:

The organization of defense.	**Pre-Encounter**	**Post-Encounter**	**Circa-Strike**
Predatory Behavior	Foraging	Search & Procure	Handling & Consumption
Function of Defensive Mode	Reduce the likelihood of encountering a predator	Decrease the likelihood of detection and attack	Survive direct contact with a predator
State	**anxiety**	**fear**	**panic**
Antecedent Stimuli	Past experiences with predation or threats	Detection of a predator or imminent threat	A striking predator is making or is about to make physical contact
Consequent Behaviors	Stretched approach, alterations in meal patterns (less frequent larger meals), retreat to nest	Freezing and thigmotaxis	Audible vocalizations (scream), vigorous escape attempts. Protean movement.
Neural Circuit includes:	Prefrontal cortex, ventral hippocampus	Amygdala, bed nuclei of the stria terminalis, ventral periaqueductal gray	Dorsal periaqueductal gray with sensory inputs from the superior colliculus.

Supplementary Note 5

Essay on Fear and its Underlying Circuits: The Promise of Translational Neuroscience

Kerry J. Ressler, MD, PhD
McLean Hospital, Harvard Medical School

The work of my lab focuses on genetic and neurobiological approaches to fear-related disorders, in particular post-traumatic stress disorder (PTSD) in humans, and the neural, molecular, and pharmacological mechanisms of Pavlovian Fear or Threat Conditioning in mouse models. Our goal is to utilize this highly conserved circuitry and resultant behavior to understand and develop new approaches for treatment of fear-related clinical disorders.

The Human subjective experience of Fear: We have all felt afraid, scared or terrified. Fear makes some feel anxious, alone, embarrassed, or ashamed, while others may become angry, defensive, or hostile. Socially, group experiences of fear of the unknown and xenophobia can drive negative political movements, mass hysteria, and even war. Throughout human history, fear has driven our actions and often unwanted symptoms and behaviors – such as anxiety, PTSD, and aggression. As a society, fear-based societal movements have been unethical, inhumane, and destructive. Ironically, fear - which exists for our individual survival - may be the strongest emotional driver of risk for our own self-destruction as a species. While we can debate the exact meaning of fear, its subjective vs. objective definitions and fear-related internal experiences vs. external behaviors, the emotion of fear is one that is fascinating but also of critical importance to understand and learn to control, as its dysregulation is an important cause of individual morbidity and irrational behavior at a societal level.

Clinical Implications of Fear: Anxiety and related disorders, from Generalized Anxiety Disorder, Panic Disorder, Social Anxiety, and Phobia to PTSD, are among the most common of Psychiatric maladies, affecting hundreds of millions of people world-wide. Combined, they are also among the highest in terms of morbidity, loss of work, comorbid medical disorders, and mortality from suicide. Despite these unfortunate statistics, we understand these disorders moderately well and have reasonable treatments.[1] These disorders all share the core emotion of fear and threat-related symptoms[2]. The diagnosis of a panic attack, shared among all of these disorders,[3] includes racing heartbeat, sweats, chest pains, breathing difficulties, feelings of loss of control and a sense of terror, fear, impending doom and death – basically the 'fear reflex' run amok!

The same reflexes and symptoms that are 'normal' in a threatening situation are experienced by those with anxiety disorders all the time – as if they can't 'turn-off' the fear switch. Furthermore, the most well-supported, empirically-validated treatments for these disorders rely on repeated exposure, now understood as the process of 'fear extinction'.[4] Advances in our understanding of mechanisms of fear and threat-processing, its underlying neural circuitry and molecular biology, and improved methods of fear inhibition and extinction will contribute to advancing treatment and prevention for these devastating disorders.

Mammalian Neural Circuits of Fear: Why is the emotion of fear so strong? All mammals, in fact almost all vertebrates, share a very similar evolutionarily conserved neural circuitry of threat processing[5]. The fear reflex is arguably the strongest of primitive instincts for survival and a critical driver of evolution. Begging

the question: can this most base, most central emotion be controlled? In recent years, neuroscientists have made amazing advancements in our understanding of the role of the amygdala,[6-9] the best known brain area underlying the fear response, and its interactions with other regulatory brain components.

'The' Amygdala is actually a cluster of several cell clusters or 'nuclei' that together make up the amygdala complex, which is conserved from the most primitive mammals and in most vertebrates. It receives neural projections from essentially all sensory areas of the brain, as well as memory processing areas – such as hippocampus and entorhinal cortex, in addition to association and cognitive brain regions[10]. It sends projections back to many of these areas, but most interestingly, also communicates with an array of brainstem and other subcortical areas. Progress in dissecting these connections has contributed to our understanding of how they regulate the autonomic, physiological, and behavioral activity patterns that together comprise the 'fear reflex' which appears to be highly conserved across species. Furthermore, work in human neuroimaging studies, has shown that the amygdala complex is reproducibly activated when subjects are viewing emotionally activating cues, such as pictures of fearful faces.[11]. Combining work across mice and humans, progress in this area has been both rapid and fascinating.

Microcircuits of Fear within the Amygdala: In all areas of science, accomplishing a more detailed or precise understanding generally leads to new, often previously unimaginable, questions. Recent fascinating work has similarly shown that even within the same subregion of the amygdala, neighboring cells can have opposing functions. An array of fantastic new molecular tools, from optogenetics to chemogenetics to *in vivo* dynamic imaging, has allowed a functional dissection of cells, molecules and pathways that underscore threat processing and inhibition.

Such findings now suggest that parallel information pathways, for example different cells encoding 'fear-on' vs. 'fear-off' information, flow through basolateral and central amygdala nuclei.[12-16]. Furthermore, the same cells that 'turn off' a fear response may be responsible for activating positive emotions – such as appetitive or even addictive behavior. Thus, these information channels may be better appreciated as underlying approach vs. avoidance related behaviors and drives. While it is clear that the different functions of these opposing cells will surely be more complex than such a dichotomous perspective,[17-19] such new insights are demonstrating a remarkably specific level of behavioral control by even a small number of precisely connected neurons. Understanding these processes will provide novel and robust insights into control of specific kinds of emotional responses, in particular fear and threat.

From a translational perspective, such a cellular level of precision of behavioral control leads to remarkable possibilities. Through single-cell RNA sequencing, we can now determine if similar cell types and microcircuits are conserved from mouse to human. Furthermore, we can ask if these conserved pathways also share molecular targets, so that one could apply data analytics and bioinformatics towards understanding combinations of drugs that might specifically inhibit conserved fear circuits or enhance extinction circuits[20]. Could we use brain stimulation techniques or even gene therapy to target fear circuits in reliable, therapeutic ways?

Conclusion: Fear is a remarkable emotion – it is associated with a range of psychiatric disorders and social difficulties, while it also represents a conserved neural circuitry of threat-related behavior that is among the best understood circuits in mammalian neuroscience. Although much progress is needed, the tools now exist for the possibilities of neurobiologically-driven rational approaches to pharmacology and neural circuit manipulation that could transform the way we approach anxiety disorders, PTSD, and behavioral disorders in general.

References:

1. Ross DA, Arbuckle MR, Travis MJ, Dwyer JB, van Schalkwyk GI, Ressler KJ. An Integrated Neuroscience Perspective on Formulation and Treatment Planning for Posttraumatic Stress Disorder: An Educational Review. *JAMA Psychiatry.* 2017 Apr 1;74(4):407-415.
2. Lang PJ, Bradley MM, Cuthbert BN. Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biol Psychiatry.* 1998 Dec 15;44(12):1248-63.
3. The Diagnostic and Statistical Manual of Mental Disorders (*5th ed.;DSM–5*; American Psychiatric Association, 2013)
4. Rothbaum BO, Davis M. Applying learning principles to the treatment of post-trauma reactions. *Ann N Y Acad Sci.* 2003 Dec;1008:112-21.
5. Davis M. Pharmacological and anatomical analysis of fear conditioning using the fear-potentiated startle paradigm. *Behav Neurosci.* 1986 Dec;100(6):814-24.
6. Pitkänen A, Savander V, LeDoux JE. Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. *Trends Neurosci.* 1997 Nov;20(11):517-23.
7. Fanselow MS, LeDoux JE. Why we think plasticity underlying Pavlovian fear conditioning occurs in the basolateral amygdala. *Neuron*. 1999 Jun;23(2):229-32.
8. Janak PH, Tye KM. From circuits to behaviour in the amygdala. *Nature*. 2015 Jan 15;517(7534):284-92.
9. Fenster RJ, Lebois LAM, Ressler KJ, Suh J. Brain circuit dysfunction in post-traumatic stress disorder: from mouse to man. *Nat Rev Neurosci*. 2018 Sep;19(9):535-551.
10. Pitkänen A, Pikkarainen M, Nurminen N, Ylinen A. Reciprocal connections between the amygdala and the hippocampal formation, perirhinal cortex, and postrhinal cortex in rat. A review. *Ann N Y Acad Sci*. 2000 Jun;911:369-91.
11. Rauch SL, Shin LM, Wright CI. Neuroimaging studies of amygdala function in anxiety disorders. *Ann N Y Acad Sci*. 2003 Apr;985:389-410.
12. Herry C, Ciocchi S, Senn V, Demmou L, Müller C, Lüthi A. Switching on and off fear by distinct neuronal circuits. *Nature*. 2008 Jul 31;454(7204):600-6.
13. Beyeler A, Namburi P, Glober GF, Simonnet C, Calhoon GG, Conyers GF, Luck R, Wildes CP, Tye KM. Divergent Routing of Positive and Negative Information from the Amygdala during Memory Retrieval. *Neuron*. 2016 Apr 20;90(2):348-361.
14. Namburi P, Beyeler A, Yorozu S, Calhoon GG, Halbert SA, Wichmann R, Holden SS, Mertens KL, Anahtar M, Felix-Ortiz AC, Wickersham IR, Gray JM, Tye KM. A circuit mechanism for differentiating positive and negative associations. *Nature*. 2015 Apr 30;520(7549):675-8.
15. McCullough, K. M. et al. Molecular characterization of Thy1 expressing fear-inhibiting neurons within the basolateral amygdala. *Nature communications* 7, 13149, doi:10.1038/ncomms13149 (2016).
16. Andero R, Dias BG, Ressler KJ. A role for Tac2, NkB, and Nk3 receptor in normal and dysregulated fear memory consolidation. *Neuron*. 2014 Jul 16;83(2):444-454.
17. Paré, D. & Quirk, G. J. When scientific paradigms lead to tunnel vision: lessons from the study of fear. *NPJ Science of Learning* 2, 6, doi:10.1038/s41539-017-0007-4 (2017).
18. Kim JJ, Jung MW. Fear paradigms: The times they are a-changin'. *Curr Opin Behav Sci.* 2018 Dec;24:38-43.
19. Campese VD, Sears RM, Moscarello JM, Diaz-Mataix L, Cain CK, LeDoux JE. The Neural Foundations of Reaction and Action in Aversive Motivation. *Curr Top Behav Neurosci.* 2016;27:171-95.
20. McCullough KM, Daskalakis NP, Gafford G, Morrison FG, Ressler KJ. Cell-type-specific interrogation of CeA Drd2 neurons to identify targets for pharmacological modulation of fear extinction. *Transl Psychiatry*. 2018 Aug 22;8(1):164.