11 Unlearning Fear

In chapter 10, we discussed how brain circuits are key to generating complex behaviors in general terms. But how does it come about in practice? To provide a more concrete example, here we discuss *extinction learning*: After learning the association between a conditioned stimulus (say, a light) and an unconditioned stimulus (say, a shock), how does an animal learn that a light no longer predicts shock when the world has now changed and the one no longer leads to the other?

Much like other nine-month-olds, "little Albert" was not really bothered by the presence of a small white rat. Unlike other infants, however, little Albert was a subject in a study by John Watson, one of the early proponents of behaviorism. In one of the most controversial experiments in psychology, Watson and his assistant, Rosalie Rayner, applied their knowledge of classical conditioning to induce fear in the infant. They presented the boy with a white rat and then loudly clanged an iron rod. Not surprisingly, little Albert responded by crying. After multiple paired presentations, Watson and Rayner presented the white rat by itself, which led to a "fear response" (the boy cried). They had conditioned an initially neutral stimulus, which now evoked a response originally triggered by the loud noise. Priding himself on his ability to shape people's emotions, Watson later went into advertising and published an influential book on infant psychological care—yes, gasp.

In what now can only be seen as a perverse experiment, Watson and Rayner were building on knowledge about classical conditioning, most notably on the experiments of the Russian physiologist Ivan Pavlov. Today, in school, we all learn about Pavlov's dogs and how they came to salivate on hearing the bell. His discovery of the conditioned response was one of his most significant contributions to physiology and psychological science. (Pavlov earned the Nobel Prize in 1904 "in recognition of his work on the

physiology of digestion, through which knowledge on vital aspects of the subject has been transformed and enlarged.")[1] Pavlov was also very interested in what he called the "internal inhibition of conditioned reflexes." He noted that the absence of reinforcement resulted in a weakening or disappearance of acquired behaviors; for example, in dogs, the discontinuation of food delivery on hearing the bell led to the weakening of salivation. More generally, when a conditioned stimulus (CS) no longer predicts the unconditioned stimulus (UCS; say, a shock) to which it was paired in the past, the CS gradually stops eliciting the conditioned response. This process is called *extinction*.

Fear conditioning has been among the most influential paradigms in all of psychology. Whereas extinction has not been so popular, it has also attracted a lot of attention. The importance of both phenomena, which include an extended family of related paradigms, transcends the laboratory, of course. Four out of five Americans will be exposed to a trauma during their lifetime, and many of them will develop a form of anxiety disorder, including phobia, social anxiety, and posttraumatic stress disorder. These conditions can be extremely debilitating and substantially impair quality of life. Not surprisingly, an array of psychological therapies has been developed to try to cure or at least ameliorate these conditions.

Consider trauma-focused approaches, such as "exposure therapy." When people are highly fearful of something, they tend to avoid the feared objects, activities, or situations. A person may avert parties if they experience social anxiety, for instance. Avoidance is initially beneficial; after all, the feared object or situation induces feelings that can be very unpleasant. But in the long term, avoidance can have rather negative consequences by excluding the person from the feared objects/situations. In exposure therapy, one is exposed to the fear-triggering event but in a safe environment. The idea is that if the event is not followed by an aversive experience, after multiple such pairings, the individual will be desensitized and the event will no longer evoke an unpleasant outcome. The general logic of exposure therapy is therefore that of the extinction processes. So, a CS no longer followed by a UCS *extinguishes* the conditioned response, which is the one that is so negative. When exposure therapy works, what makes it successful? When does it work best? At present, we don't know the answers to these questions because extinction has turned out to be fiendishly challenging to dissect, both at the behavioral and the neuroscientific levels.

Fear Extinction as Inhibition of Emotion by Cognition

Let's delve more deeply into the mechanisms of *fear extinction* (figure 11.1). When a conditioned stimulus no longer predicts the unconditioned stimulus to which it was paired at some point in the past, the CS stops eliciting the conditioned response. Pavlov himself believed this change involved the "development of internal inhibition," but was quite vague about the mechanisms involved, aside from alluding to cortical cells "entering into a state of inhibition" (Pavlov 1927). But the notion that the CS acquires *inhibitory* properties that allow it to suppress the conditioned response has played a central role in thinking about extinction.

The *acquisition* of fear itself during classical conditioning relies on the amygdala as well as several of its targets in the brainstem (chapter 5). But how about extinction? The role of the prefrontal cortex in the regulation of behavior in general, and emotion in particular, is a perennial theme in neuroscience. If investigating the question in a rat, a natural candidate to look at would be the medial sector of the frontal lobe, as rats don't have a prominent lateral prefrontal cortex (some researchers even question whether they have brain areas that are comparable to the lateral prefrontal cortex of primates). In the early 1990s, Joseph LeDoux and colleagues reported that

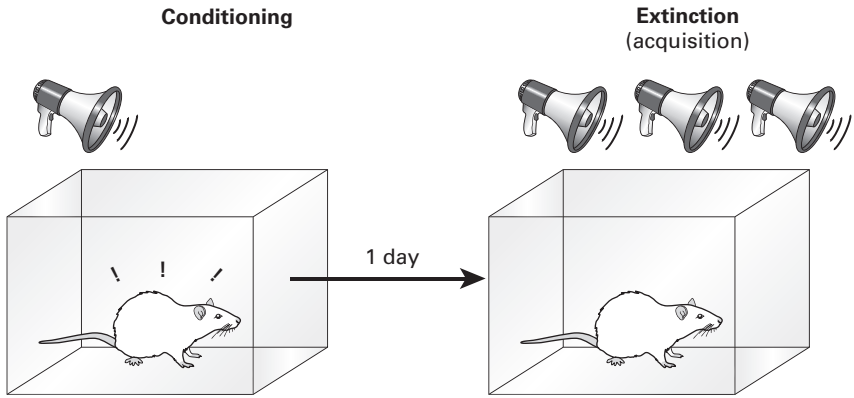


Figure 11.1
Unlearning fear. After aversive conditioning, the conditioned stimulus is presented alone multiple times until a conditioned response is no longer produced. Whereas after conditioning the animal freezes after the tone, once extinction occurs, the animal moves around normally.

the medial prefrontal cortex plays an important role during fear extinction (Morgan, Romanski, and LeDoux 1993). Animals with a lesion of the medial prefrontal cortex (PFC) took considerably longer to extinguish learned associations. Because the medial PFC is extensively interconnected with the amygdala, as well as with several of its brainstem targets that generate conditioned responses, the findings resonated with the idea that the medial PFC inhibits the amygdala, thereby halting the conditioned response. This mechanism of fear extinction fit the old formula: cognition, tied to the medial PFC, controlling emotion, itself tied to the amygdala and other subcortical structures (figure 11.2).

Behaviorally, what is extinction? In the experiment by LeDoux's group, following the extinction procedure, they tried to determine the general fearlessness of the rat. Had the animals simply become unusually bold and the presumed fear extinction a manifestation of their new personality? No, the general fearfulness of rats didn't seem altered. But when presented with the prior CS, they didn't mind it as much as before, and so the stimulus's ability to produce freezing behavior was diminished.

Let's delve deeper into the extinction process. When the CS no longer predicts an aversive outcome, it behooves the animal to take into account

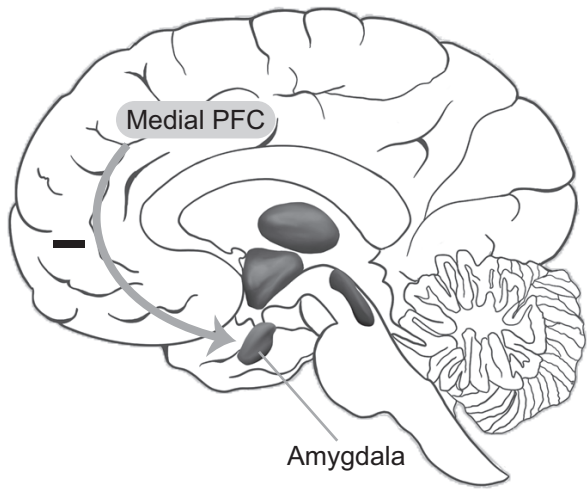


Figure 11.2
Early proposed mechanism for fear extinction. The medial prefrontal cortex (PFC) inhibits the amygdala, thereby preventing the conditioned response from being generated when the conditioned stimulus no longer predicts it.

this information, learning features of the now-safe environment that signal safety. But when the CS no longer predicts the UCS, why exactly is this case? Is the CS being presented in a completely different environment or context? What if the original CS (call it CS1) now appears simultaneously with another environmental stimulus (call it CS2). If no UCS ensues, should safety be deemed thanks to the CS1 (perhaps it no longer predicts the UCS) or to the CS2 (CS2's appearance now makes the world safe)? Experiments show that if, during the extinction process (when the CS1 is presented without the concomitant UCS), another stimulus (CS2) is presented alongside the original CS1 will *not* be treated as safe. This situation is at times called "protection from extinction." In other words, the relationship between the original CS and the UCS is *maintained*, and when the CS is presented alone, it produces a conditioned response—"fear" continues. The absence of the UCS is being attributed to the additional factor (the CS2), and the animal had better be careful (about CS1). A similar protection from extinction takes place when a new action concurrently performed by the animal leads to safety (that is, prevents the occurrence of the UCS). Here, the action is attributed with the power to ward off the punishment. So the animal will still fear CS1.

The intelligence of the learning processes is further highlighted by a scenario called "backward blocking." Suppose a compound stimulus, CS1 + CS2 (such as a light and a tone), is associated with a UCS. Once learning occurs, by definition the presentation of the pair CS1 + CS2 will generate a conditioned response. When either CS1 or CS2 is presented alone, some amount of conditioned responding ensues, although weaker than when the pair is presented. Now, if from this point forward one of the elements of the compound stimulus (say, the light) is consistently paired with the UCS, the second element (the tone) will *cease* to generate a conditioned response. It seems that since the light can fully predict the UCS, the tone is regarded as unrelated to the UCS, indeed *retrospectively*. This is all the more striking because the predictive value of a CS can be updated despite its absence. In the present case, the value of the tone is updated when the light is presented by itself. Whatever extinction is, it's not dumb!

In sum, extinction is more than a simple form of inhibition. It is a sophisticated form of *learning*, and as such the formation of an "extinction memory" involves processes akin to those observed in learning in general: acquisition, consolidation, and retrieval. What is being learned is safety.

The manner by which this memory influences behavior depends on how it was established (acquisition), how it was strengthened (consolidation), and how it will be reactivated (retrieval) in particular situations. The chief goal is to learn what should be feared, and therefore avoided, and what is safe and doesn't call for special measures and might be even approached. The factors that drive this process fall into two categories: those that promote defensive responding (such as "fearing" the stimulus) and those that do not ("safety" responding). To successfully accomplish extinction, the nature of the CS-UCS relationship needs to be unraveled: the CS might no longer predict the UCS, the CS might predict the UCS less reliably than before, or the CS might predict the UCS just as well as before, but something else is preventing the UCS from occurring. The conclusion that is favored will determine if the animal will express fear or not on encountering the stimulus in the future.[2]

Diving Deeper into the Mechanisms of Extinction

We saw that the medial prefrontal cortex plays an important role during extinction. Given that the region is extensively interconnected with the amygdala, it was natural to think that the former inhibits the latter. Nevertheless, since the early studies in the mid-1990s, the picture that the medial PFC *controls* the amygdala has been muddied considerably. For example, chemical blockage of the basolateral amygdala either impairs or entirely prevents extinction in the first place.[3] Furthermore, morphological changes in synapses in the amygdala itself support the consolidation of extinction. These findings strongly counter the notion that the amygdala is simply inhibited by the medial prefrontal cortex. Instead, it is a critical site for the *formation* of safety memories, very much like it is important for processes that establish fear memories themselves.

Anatomically, the amygdala isn't only the target of pathways from the medial PFC but also projects *to* it. Together with the findings above, it becomes untenable to place the amygdala as "down" from the medial PFC, and in fact some investigators propose that the amygdala actually should be viewed as "upstream" of the medial PFC.[4] Put another way, the amygdala and the medial PFC interact in complex ways during extinction. It is well established that the amygdala plays a critical role in establishing the

association between a CS and a UCS (fear learning). It is increasingly clear that it participates in a major way in learning safety (extinction), too.

When a CS no longer predicts a UCS, the specific environment where extinction learning takes place is paramount. The animal learns that the CS *in this environment* is now safe. Indeed, if the CS now reappears in a novel context, the animal displays defensive behaviors—the CS does not signal safety there. Studies have shown that the hippocampus keeps track of the context in which extinction occurs. This type of learning is essential; after all, it could be disastrous for the animal to generalize the safety of a CS to situations unlike those from where extinction took place. The contributions of the hippocampus to contextual learning will be discussed below, but first let's consider some of the functions of this structure—some of which were learned the very hard way.

Hippocampus: A Brief Detour into a Tragic Neurosurgery

At the age of 27, Henry Molaison (known as patient HM when he was alive) was referred to William Scoville, a neurosurgeon at Hartford Hospital. Despite maximum medication of various forms, he could not lead a regular life.[5] He had worked for a time at an assembly line as a motor winder but had become so incapacitated by his seizures that working was no longer possible. The year was 1953, and a radical clinical approach was taken. With the understanding and approval of the patient and his family, the "frankly experimental" operation was undertaken. The surgery didn't eliminate the seizures, but they became less debilitating than before. A standard IQ test indicated that his score was unaltered compared to the test taken before surgery. His personality also appeared stable. So far, so good. Yet, Molaison exhibited a profound, indeed devastating, memory impairment. For example, just before his psychological examination in April 1955, he had been talking to a famous neuroscientist, Karl Pribram, but he formed no memory of this event and denied that anyone had spoken to him. During conversations, he constantly recounted boyhood events and, eerily, didn't appear to realize that he'd had an operation. Something was clearly amiss.

This is how Scoville, the surgeon, and Brenda Milner, the neuropsychologist we encountered in chapter 4, summarized his status in a watershed paper published in 1957:

> After operation this young man could no longer recognize the hospital staff nor find his way to the bathroom, and he seemed to recall nothing of the day-to-day events of his hospital life. There was also a partial retrograde [that is, of the past] amnesia, inasmuch as he did not remember the death of a favourite uncle three years previously, nor anything of the period in hospital, yet could recall some trivial events that had occurred just before his admission to the hospital. His early memories were apparently vivid and intact. (Scoville and Milner 1957, 14)

Given the Hippocratic oath of "do no harm," one can only imagine how the surgeon must have felt. In fact, Scoville and Milner stated that one of the goals of their report was to provide a much-needed warning to others about the risk of the procedure. Scientifically, their paper proposed that the hippocampus is "critically concerned in the retention of current experience." In neuroscience, few other reports have spurred such an enormous literature involving both human and animal research. To say that thousands of papers have their origin with their publication is no exaggeration. It might thus come as a surprise to the reader that, six decades later, the exact contributions of the hippocampus to memory remain a matter of intense and heated debate.

In 1971, John O'Keefe and Jonathan Dostrovsky reported that neurons in the hippocampus of the rat respond selectively to places in the environment. Animals were placed in a rectangular box, and when they were in particular locations—say, at the middle of the southmost wall facing south—specific cells in the hippocampus fired vigorously. O'Keefe and Dostrovsky speculated that the region provides the rest of the brain with a "spatial reference map" given that some neurons are particularly attuned to the spatial location, or the place, of the animal in its environment—these neurons would later be popularized as "place cells" (figure 11.3). The implications of these findings were developed extensively in a book by O'Keefe and Lynn Nadel, published in 1978 and now considered a classic—*The Hippocampus as a Cognitive Map*. Since then, intense debate contrasting spatial versus memory functions of the hippocampus has raged in neuroscience.

By treating the hippocampus as a cognitive map, O'Keefe and Nadel attempted to build on the theoretical framework developed in the 1930s by the psychologist Edward Tolman.[6] At a time when learning was viewed as a simple process of passive accumulation of associations imposed on the animal by the environment, Tolman viewed learning as an active process of extracting information from the world. To him, animals track the

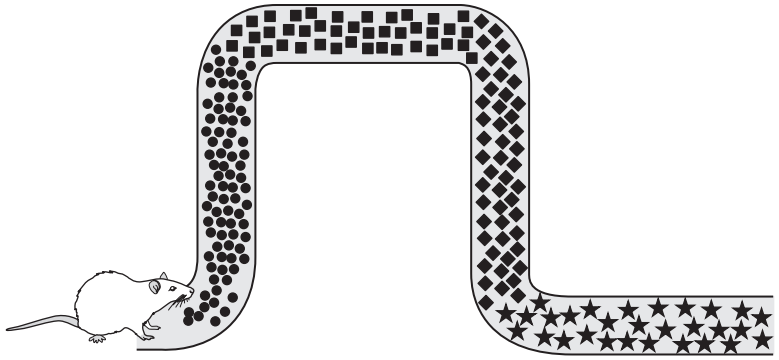


Figure 11.3
"Place cells" in the hippocampus. As the rat navigates down the path, specific neurons fire more vigorously at certain locations. Thus, one neuron fires strongly at the locations marked with stars, another at the locations marked with circles, and so on.

underlying structure of the world through a maplike representation of causal associations (that is, what leads to what?). Its central concept, the cognitive map, allowed the combination of causal information to produce novel ways of achieving outcomes, much like a physical map allows the planning of novel routes to a previously visited destination.

Cognitive maps provide a summary of the places visited by the animal, together with information about distances and directions between them. As an animal moves about its environment, researchers have found that the hippocampus helps encode information about it, including establishing distance and direction vectors: How much distance has it covered and in what directions? Indeed, researchers have by now uncovered various hippocampus neuronal properties that are summarized by catchy descriptors, including "grid cells," "border cells," "head direction cells," "speed cells," and "time cells." (Research on the hippocampus landed John O'Keefe and the Norwegian investigators Edvard Moser and May-Britt Moser the Nobel Prize in 2014; O'Keefe received half the prize and the Moser couple shared the other half.)

As the cell monikers indicate, hippocampal responses are attuned to the spatial and temporal properties during an animal's navigation through its environment. But the more these cells are studied, the clearer it becomes that their activity is very nuanced. Hippocampal firing to places, borders, direction, speed, and so on is influenced by a gamut of factors, including

the presence or absence of objects, the presence of a stimulus previously paired with aversive outcomes (as in conditioning paradigms), as well as generally factors such as novelty, attention, and even an animal's internal state (is it hungry?). Other motivational information also plays a role, as cells fire more vigorously near "task goal" locations, including places where an animal receives reward. Together, the firing of hippocampal cells reflects spatial knowledge in extremely rich and multifaceted ways.

Why does the memory-versus-space debate persist to this day? On the surface, it is difficult to appreciate how two such different views of hippocampal function can be reconciled. Is one of them just plain wrong and waiting to be debunked? As often is the case in science, when groups hold opposing views for a long time, both are probably right, at least to some extent. In the present context, one way to square the contrasting views is to think that the brain uses space as a way to organize memories, or what are called episodic memories. Returning to a place, or thinking about it, helps retrieve memories of things and events that happened at that location. A related idea is that the hippocampus generates a "memory map," at once a map of space and a map of memories, together with the links between them.

Despite decades of vigorous work, determining the contributions of the hippocampus to memory remains very much a matter of current scientific interest. However, a noticeable change in today's research is that it is less and less centered on a sole region—it's not all about the hippocampus anymore. Instead, researchers try to understand how the hippocampus interacts with neighboring areas in the temporal lobe (the so-called medial temporal lobe), as well as how it participates in broader interactions with regions across the brain. Episodic memory and spatial navigation are not carried out by single regions—no mental process is.

The Place of Extinction

Extinction is not simply forgetting, or inhibiting, the link between a CS and an aversive event. Considering the environment where the CS and the UCS become uncoupled is a must. In fact, associating environments, rather than just specific cues within them, with safety versus danger is necessary for survival. What anatomical pathways support these processes?

The hippocampus provides context-related information that guides extinction.[7] Animals need to keep track of where extinction took place,

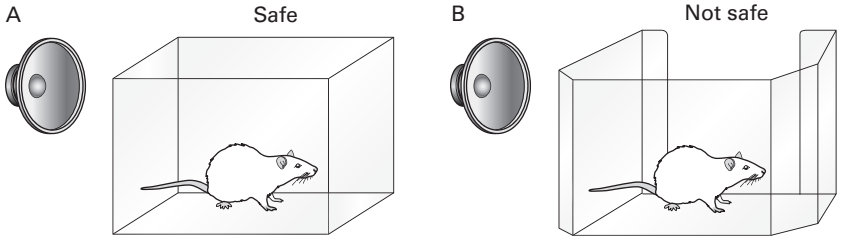


Figure 11.4
Context information is essential for extinction. (a) Original environment context where extinction took place. If the conditioned stimulus (CS) reappears here, the animal is likely safe. (b) If the CS occurs in a new environment, the animal must register this difference and treat it as unsafe. In other words, it is unwise to generalize the safety of the CS across contexts.

because if the CS reappears there, the animal is probably safe. But when the context changes, it makes sense to treat it as dangerous if the CS appears (figure 11.4). The hippocampus has direct connections to the amygdala, and the targeted neurons in the amygdala promote defensive responding ("fear"). Through this pathway, the hippocampus signals that the CS is now happening in a novel context—fear is renewed. The hippocampus also has dense projections to the medial prefrontal cortex, and the hippocampus can engage the medial PFC to indicate that the environment has *not* changed. Here, the original context of extinction is the same experienced presently, so it is likely safe.

Changing Values

Animals learn that some environmental cues are positive signs and predict reward. The pattern of earth around a burrow may indicate to a fox that a mouse has just entered it and that quickly excavating the hideout may lead to catching the prey. After some experience, the animal learns to associate the cue (the earth pattern) with the reward (the mouse). But now suppose that this contingency changes, and the cue is no longer predictive of reward; perhaps mice no longer enter the burrow in a manner that leaves such clear traces. More consequently, the cue may now be associated with a negative outcome. For instance, what if some burrows now contain snakes? In such cases, the animal needs to learn that the cue no longer predicts reward.

Learning to reverse an association, which is known as *reversal learning*, engages the orbitofrontal cortex. A lesion study in the early 1970s demonstrated that monkeys with damage to this area were impaired in their ability to switch or reverse behavior.[8] During reversal learning, the animal first learns that an item is good and predicts subsequent reward, while another item is bad and predicts punishment (or at least is not followed by reward). Training continues until it is clear that the animal has learned the mapping, at which point the association is reversed by the ornery human. Thus, when the first reversed trial is experienced, the animal experiences a complete mismatch between expectation and what is delivered. Without the orbitofrontal cortex, animals had considerable trouble learning the new contingency.

In the early 1980s, investigators managed to record from neurons in the orbitofrontal cortex while monkeys actively performed tasks (Thorpe, Rolls, and Maddison 1983). In some of the first experiments, spiking activity was recorded during reversal learning. Some neurons fired vigorously when the monkey saw a syringe used to deliver black currant juice. But when the contents of the syringe now contained saline (which is mildly aversive, especially in comparison with a favored juice), the monkey's activity declined sharply on seeing the syringe. The discovery of neurons that responded to the *meaning* of the stimulus was quite exciting and led to a wave of experiments trying to sort out the functions of this part of the cortex. In the coming decades, it became increasingly clear that it encodes *value* (figure 11.5): In other words, the firing is not due to particular object features (like the syringe) but to the outcome that it predicts (sweet juice). A corollary of this finding is that distinct objects that signal the same reward generate very similar responses. Reinforcing the notion of value coding, neurons in the orbitofrontal cortex integrate information about the magnitude and the probability of reward. This is critical, because if outcomes are not always certain, one must take into account their probability of occurrence. As one can intuit, a potential reward of $100 that has a probability of 10 percent is not as attractive as an intermediate reward of $50 and a good probability of occurring, say, 70 percent (or even a moderate reward of $25 that is more certain with, say, a 80 percent chance).

Even more broadly, neuronal activity in the orbitofrontal cortex represents the *expectation* of the value of the outcome (Roesch and Schoenbaum 2006). According to this view, the firing of neurons is not simply a result of

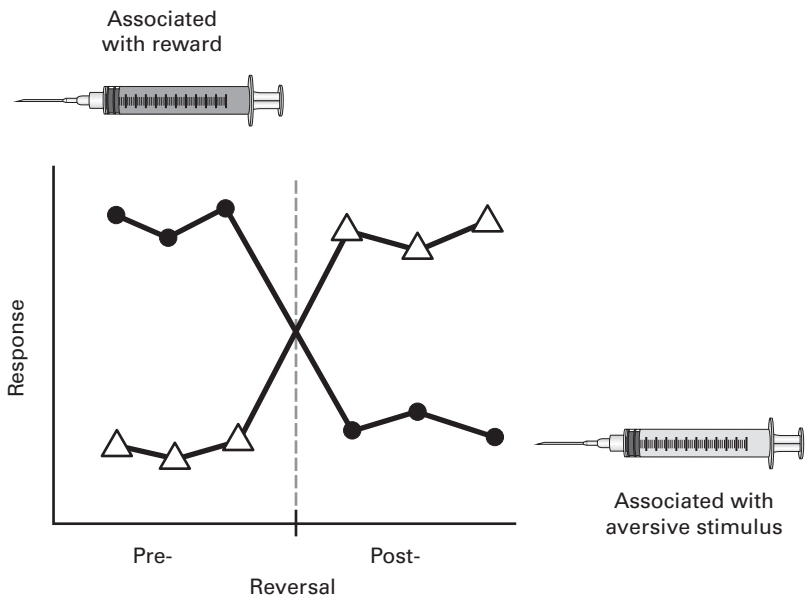


Figure 11.5
Reversal learning and cell responses in the orbitofrontal cortex. Responses to cues predicting "good" (associated with reward) stimuli shift drastically when their meaning is reversed, and vice versa. The vertical line marks when the previously rewarded stimulus now predicts an aversive outcome, and vice versa. The *x* axis indicates blocks of trials before or after reversal.
Source: Results based on Rolls et al. (1996).

the association between an object and its outcome in the past. Instead, it reflects a *prediction* about potential outcomes generated on the fly, at that moment in time. Consider the following "devaluation" experimental paradigm. An animal first learns that different objects go with different food rewards. Say object 1 predicts high reward (a favored food, such as raisins) and object 2 predicts low reward (a less preferred piece of fruit). When the animal sees object 1, neurons in its orbitofrontal cortex fire vigorously, and less so for object 2. When given a choice, the animal will pick object 1 as a means to obtain the favored food. But are the responses to seeing the objects related to the value of the foods? To get at that, experimenters let the animal overfeed on its favored food. What does the monkey do when offered a choice between objects 1 and 2? It will go for object 2 (after all, we've all experienced the decrease in pleasure after overeating sweets!).

Interestingly, monkeys with orbitofrontal cortex lesion do not show a bias for the less favored food; they fail to update the new value of the objects based on their current state (being satiated with the preferred food). They reach out to object 1 even though the food it predicts is not really that desired after consuming so much of it.

We saw how neurons in the hippocampus are particularly sensitive to spatial locations and other navigational information. But cells there are affected by reward and the overall goal relevance of a place, too. Given the involvement of the orbitofrontal cortex in valuation processes, might these two structures work together to integrate spatial information and value? Researchers have started to uncover how interactions between them might be involved. In one experiment, a rat had to navigate its environment in particular ways to receive a reward.[9] In the setup, the animal navigated down an alley and, at certain points, was forced to decide whether to turn right or left (the contraption is called a T-maze given right/left choice points at the top of T-like bifurcations). By trial and error, the rat had to discover which behavior led to a reward—for instance, always turn left at a junction or alternate right and left turns. At the beginning, when rats hadn't figured out the pattern yet, they often paused at the choice point before making a left or right turn. What were they doing? Analysis of both their behavior and cell responses suggests that they were simulating the consequences of potential actions before deciding which turn to take. When they paused just prior to turning, hippocampal cell firing encoded information about pathways ahead of the animal (along the potential left- and right-turn paths), in a manner consistent with trying to determine the consequences of particular actions. Intriguingly, orbitofrontal neurons fired along an entire path based on the probability that the path in question would lead to reward. So, if a segment of the course was part of a pathway leading eventually to reward, firing was vigorous, and vice versa. As the hippocampus has direct connections to the orbitofrontal cortex, the former likely conveys spatial information to the latter so that its behavioral significance can be ascertained—namely, will it lead to a reward?

We've seen that the hippocampus and orbitofrontal cortex encode somewhat overlapping information. Clearly the hippocampus is more attuned to space while the orbitofrontal cortex is more linked to rewards. But instead of thinking of them as implementing different functions—hippocampus:

navigation; orbitofrontal cortex: value—by focusing on their interactions we can see how they support behaviors that are meaningful in natural habitats.

Contrasting Explanations

Let's return to extinction (figure 11.6a). The medial prefrontal cortex in involved in learning that the CS no longer signals threat. However, viewing this region's contribution as the inhibition of emotion by cognition doesn't do justice to the behavior as well as the neuronal interactions at play. One possibility is to conceptualize extinction in terms of the multiple influences discussed so far (figure 11.6b). The final result—extinction—is the result of the multiple contributions, which result in a behavior that is flexible. (We haven't discussed the thalamus in relation to extinction, but based on recent studies, it is one more region that contributes to the process.) Whereas this description broadens the spectrum of influences considerably, it invites thinking that is too linear and region-oriented. The orbitofrontal cortex provides reward information, the hippocampus context, and so on; furthermore, the amygdala responds by combining its inputs to arrive at an answer: Should the animal be wary or not? This type of "boxes and arrows"

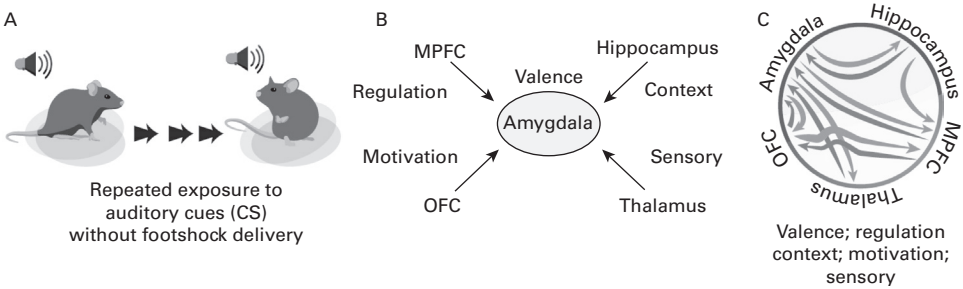


Figure 11.6
Contrasting explanations: (a) Fear extinction. (b) Fear extinction in terms of the top-down regulation of the amygdala by the medial prefrontal cortex, with additional variables influencing the process. (c) Schematic representation of the anatomical connections between some of the brain regions involved, emphasizing a nonhierarchical view of the processes leading to fear extinction. The descriptors "valence," "regulation," and so on are not tied to brain areas in any straightforward one-to-one fashion. MPFC, medial prefrontal cortex; OFC, orbitofrontal cortex.
Source: Reprinted with permission from Pessoa (2018b).

diagram has been used for over a century in neuroscience (Pessoa 2017a), so what could be the alternative?

The diagram in figure 11.6c tries to convey the idea that the brain regions *collectively* determine the extinction process. This type of description provides a different springboard to reasoning about the mapping between function and structure in the brain—how brain regions bring about behaviors. Conceptually, part of the shift is due to the fact that the putative underlying processes (valence, regulation, context, etc.) are *not* separable, so they don't encode stable variables (such as "valence") that are simply pushed up or down by the other factors. Put another way, these variables are so intertwined that they are *jointly determined*: to understand the system, we need to consider the *integration* of the signals.

The scheme diagrammed in figure 11.6c doesn't mean that all regions contribute in the same way to the behavior in question. At first glance, it appears that a lot is missed by diagramming things this way; the description appears too vague. But simplification for the sake of simplification won't help us out of our problem. Go back to figure 11.6a, which says that the behavior in question (exhibiting fear or not) is a function of activity in the amygdala, itself influenced by inputs from the other areas. Figure 11.6b, instead, tells another story: Behavior is a function of all the regions when taken *together*. And their bidirectional interactions imply that the flow of information is far from straightforward (such as the inhibition of the amygdala by the medial PFC).

The boxes-and-arrows arrangement (figure 11.6a) also invites the interpretation that the mechanisms in question are somewhat static. Indeed, in the laboratory, one generally investigates behavior at a specific point in time, or perhaps during a narrow temporal window. But natural behaviors are dynamic, evolving as the animal interacts with its environment. Whereas in the lab an animal may be placed in a chamber where it received shocks in the past, in nature the time frame of experiencing negative (or positive) scenarios is more gradual (unless the animal is surprised by a very sudden attack). Accordingly, a dynamic description of the underlying neural processes is not only beneficial but necessary in general.

Let's try to motivate further the idea of integration of signals and appreciate the nuances of signal flow. As a simple scenario, take a predator-prey system (chapter 8) involving foxes and hare. The number of foxes, F, and

the number of hare, H, will fluctuate jointly with time. The number of foxes grows based on predation and decays based on death. We can write this as

$$dF(t) = \alpha FH - \beta F,$$

where dF(t) signifies the change in the number of foxes as a function of time. The first term says that the number grows in proportion to the number of foxes times the number of hare. That is, the increase in foxes is proportional to the number of foxes (the ones that give birth to more of them) and the prey population that supports it. The multiplier α is a constant that specifies the "efficacy" of this growth process (based on predation efficiency and turning food into offspring, for example). The second term says that the number of foxes will decrease from death at a rate given by β.

Now, we need to specify how the hare population changes with time. The number of hare increases exponentially (all that mating!), except that the presence of predators puts that in check:

$$dH(t) = \gamma H - \delta FH,$$

where dH(t) is the change in the number of hare as a function of time. The first term says that the number will grow based on the number of hare present, and the second that it will decrease based on the product of the number of foxes and hare (the more foxes and the more hare, the more encounters and potential death). The constants γ and δ are efficacies of the growth rate (which also accounts for factors such as food availability, and so on) and consumption rate (by foxes), respectively.

These two equations define a "system": To know the number of foxes we need to know the number of hare, and vice versa—they are interdependent. Translating this into the extinction scenario, we can think of the activity of cells in the amygdala, hippocampus, thalamus, medial PFC, and orbitofrontal cortex as jointly interdependent. A computational neuroscientist can then specify equations for how these signals change as a function of time. But how about the experimentalist? How should the experimental scientist proceed? After all, training in neuroscience is not very mathematical. A potential direction is to move research efforts toward studying the *multiregion temporal evolution* of brain data. Here, the focus is on studying multiple regions simultaneously and trying to characterize the joint state of brain regions and how the state evolves temporally (we develop these ideas further in chapter 12). Tools from networks science, among many others,

are needed. In the end, experimental scientists need to learn more technical skills or collaborate in larger teams—more likely both.

From Foraging to Escaping

Prey are, by definition, at risk of predation. But their lives aren't always made of the dramatic moments seen in nature documentaries, such as a seal evading a white shark or a gazelle escaping a cheetah, both managing to do so through a series of dazzling twists and turns. Fortunately, they spend a considerable amount of time engaging in positive motivated behaviors, such as foraging, maintaining a nest, nursing, feeding, and mating. They undertake these behaviors when the risk of predation is minimal; obviously, they can't engage in them when they are about to be struck by a predator. But between these two extremes, they exhibit a range of behaviors that depend on their distance to predators. The distance is often the perceived one rather than based on a measuring tape; after all, a predator evades detection exactly to bring its physical separation to the prey within striking distance.

In the late 1980s, Michael Fanselow and Laurie Lester proposed that prey behaviors are structured around a *continuum of predatory imminence* with particular key stages: pre-encounter, post-encounter, and circa-strike.[10] In the absence of predators, animals will engage in their preferred activities, including the positive behaviors mentioned above. How they behave the rest of the time takes into account predator distance. During pre-encounter, behavior is based on the assessment of the probability of encountering a predator. So, although a predator hasn't been detected, foraging may proceed more cautiously in places where predators have been spotted in the past. During post-encounter, the animal's behavior shifts quite strikingly; they may suppress behavior, taking stock of the situation (should they dash away or can they continue grazing for a little longer?). Circa-strike behaviors are often a last-ditch attempt to escape from capture and are often unusual and highly energy consuming. For example, a deer mouse freezes in the presence of a gopher snake but attempts a last second, spectacular vertical leap as the snake strikes.

Investigators originally described these stages based on stereotypical and relatively fixed patterns; yet actual behaviors are quite flexible. One can think more broadly in terms of computations of *threat detection* and *threat*

escape.[11] The central goal of threat detection is to evaluate sensory information to determine if a threat is present. This assessment of threat is flexible and dynamic and calibrated by expectations built on experience. If the risk of predation is low, prey adjust the threshold for reacting to threats to a higher level (more evidence is needed) compared to when the risk is higher (less evidence is needed). Remarkably, animals quickly learn to suppress escape responses if they are repeatedly challenged but no adverse outcome ensues, even if the stimuli are potent and innately threating. In all, the choice of action when threat is spotted is highly context dependent. Threat detection has not been studied extensively in mammals, and little is known about the underlying mechanisms (more research has been conducted in invertebrates, as well as fishes, frogs, and birds). But we know that the superior colliculus (chapter 3) participates in the detection process. For example, visual signals from the retina engage the neurons in the superior layers of this structure, which are tuned to looming stimuli resembling a predator coming from above.

The ability to handle threats expands considerably with learning, and in particular, animals learn to avoid locations where predators were encountered previously. For example, when exploring an arena where they saw threats before, mice attempt to escape. (Most of the focus in rodent research has been on freezing responses instead of active escaping because the chambers used are small and don't provide a possible escape route.) Alternatively, mice generate other defensive behaviors, such as an increase in stretch postures or reduced exploratory locomotion, which indicates their altered risk assessment of the situation (Silva et al. 2013).

Naturally, after a stimulus is detected and deemed a threat, an action is required, and fast—this is the escape computation referred to before. Indeed, evolution has shaped some neural circuits to permit just that. Specialized cells in fish, for example, allow responses to start a mere 5 to 10 milliseconds after threat is surmised (this fast response allows fish to change direction and be propelled forward). But animals do not necessarily flee immediately once predators are detected. Why? In a nutshell, running away is not always a good idea. Field studies reveal that animals attempt to get away when the costs of remaining (such as the risk of injury or capture) are higher than the cost of fleeing (such as loss of foraging opportunities). There's also a close link between an animal's internal state and its decision to escape. For example, mice exhibit risky behaviors when hungry, such

as spending more time in threatening environments. Sexual receptiveness also influences the escape calculus.

A frequent alternative to fleeing is, as discussed in chapter 3, freezing in place. The choice between staying or going is determined, in part, by knowledge about the spatial properties of the environment. For example, mice memorize an escape location based on a single and brief (less than 20 seconds) visit to a shelter, and changes to the spatial environment lead to a rapid update of the defensive action chosen: fleeing versus freezing. Additional variables considered by mice include how safe the shelter is, the distance and relative position of the predator, and potential competition for shelter access.

Animals thus confront a *detection-response dilemma*: both responding too early and too late are costly. Escaping is metabolically expensive as getting away requires energy. But it is also costly in terms of opportunity losses, including those related to food and mating. In effect, triggering a full-blown escape response on detecting a threat is a rather poor survival strategy and essentially nonexistent in the natural kingdom. The decision to take flight is not just triggered by threat detection and involves computations that rely on multiple external and internal variables. Together, escape behaviors are far from simple stimulus-driven, stereotypical reactions. The mechanisms involved engage specialized circuits refined by eons of evolutionary time. Whereas some circuit components play special roles, they exchange information with multiple areas across the brain. The behavioral complexity points to solutions involving the integration of signals across distributed circuits so as to promote behavioral flexibility and survival.

As mentioned, little is known about the brain circuits involved in escape in mammals. What we do know has focused on a few usual suspects: the superior colliculus, periaqueductal gray (PAG), hypothalamus, and amygdala. To understand why knowledge is so limited, we need to consider the inherent limitations of current experimental setups. A typical rodent experiment will take place in a small cage, where the animal is exposed in a controlled fashion to stimuli such as tones, lights, or a foot shock. The animal will also have levers to press and simple decisions to make (choose food A versus B). If brain recordings are being made, the animal is tethered so that electrode signals can be conveyed to a computer for data analysis. To be sure, behavioral experiments without recordings are performed in larger setups, including T- or radial-shaped mazes. But even in these cases the

restrictions are considerable. And the lesion method, a valuable but coarse instrument, has been the mainstay of investigators. But neuroscience is changing fast. Large environments are being used more frequently; experiments with untethered animals are becoming more prevalent; and genetic and chemical manipulations allow investigators to focus on subclasses of cells in specific brain regions, such as a particular population of cells in the basolateral amygdala that has specific chemical properties. Exciting days lie ahead of us.

The overall goal of this chapter is to illustrate how an ostensibly simple behavior—fear extinction—is implemented in the brain. Far from simple, extinction is a complex learning process supported by distributed brain circuitry. Although we barely scratched the surface, we saw that only by acknowledging the simultaneous contributions of several brain areas can we hope to do justice to the intelligence of the behavior. What's more, the interdependence of the regions leads to a process of joint construction of a solution by the brain—a solution with many authors, each of which contributes different materials.

The Entangled Brain

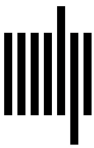
How Perception, Cognition, and Emotion Are Woven Together

By: Luiz Pessoa

Citation:

The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together
By: Luiz Pessoa
DOI: 10.7551/mitpress/14636.001.0001
ISBN (electronic): 9780262372107
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Pessoa, Luiz, author.
Title: The entangled brain : how perception, cognition, and emotion are woven together / Luiz Pessoa.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021061878 (print) | LCCN 2021061879 (ebook) | ISBN 9780262544603 (paperback) | ISBN 9780262372107 (pdf) | ISBN 9780262372114 (epub)
Subjects: LCSH: Perception. | Emotions and cognition. | Brain. | Neuropsychology.
Classification: LCC BF311 .P3767 2022 (print) | LCC BF311 (ebook) | DDC 153—dc23/eng/20220411
LC record available at https://lccn.loc.gov/2021061878
LC ebook record available at https://lccn.loc.gov/2021061879