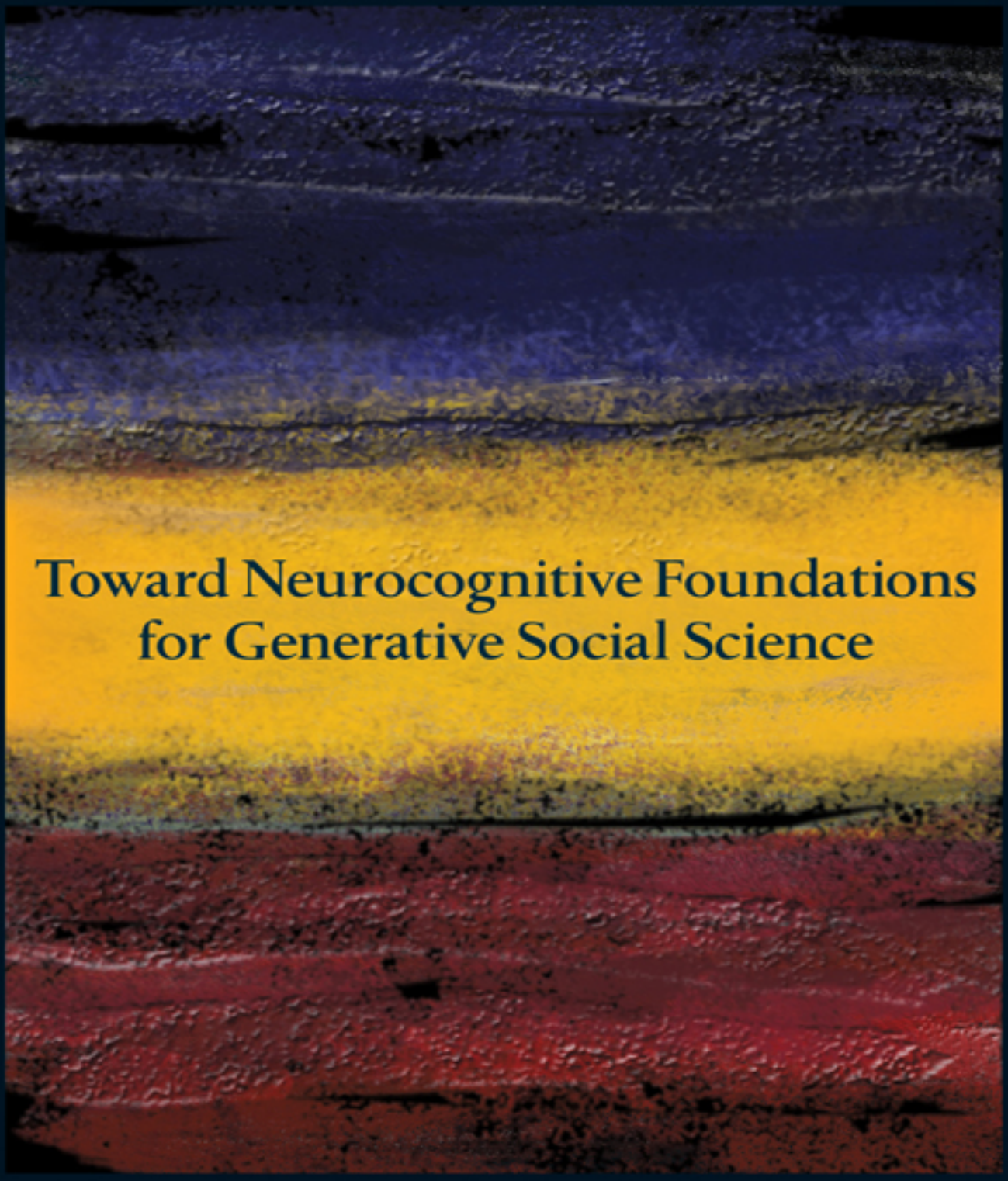
AGENT_ZERO



Toward Neurocognitive Foundations for Generative Social Science

JOSHUA M. EPSTEIN

PART I

Mathematical Model

IN THIS PART, we specify explicit mathematical models for the emotional, deliberative, and social components of the *Agent_Zero* framework. These choices are not cast in stone, and different components should certainly be explored, as discussed in the Future Research section. First, however, we review some underlying neuroscience of fear and its throne: the amygdala.[39]

This review is worthwhile because the Rescorla-Wagner equations (used for the affective model component) do not presuppose that fear acquisition is largely unconscious, while this is a crucially important fact from a social science standpoint, and the amygdala discussion demonstrates that it is a neuroscientifically sound modeling assumption. Also, important evidence of emotional contagion comes from fMRI studies of the amygdala, and if we didn't know anything about the amygdala, these images would mean very little.

Understanding, then, that unconscious fear acquisition is what we have in mind, we now discuss the *elementary* neuroscience of fear as prelude to the famous Rescorla-Wagner equations of conditioning, all en route to our more general model of behavior in groups.

I.1. THE PASSIONS: FEAR CONDITIONING

Humans are born with a variety of innate endowments or capacities. One of these is the capacity to acquire fear (and other) associations through a process of synaptic change in which, as Donald Hebb (1949) presciently put it, "neurons that fire together wire together." That is, after *certain*[40] pairings of an initially neutral stimulus (e.g., a tone) and a stimulus that is innately aversive (e.g., a shock), the

initially neutral stimulus will evoke the same response as the innately aversive stimulus. This associative process—often termed *conditioning*[41]—is generated by synaptic change, or "plasticity." For a lucid nontechnical exposition, see LeDoux (2002). We, of course, cannot cut open a human and observe her fear, but we can intelligently speak of a fear circuit—a distributed neurochemical computational architecture[42]—whose proper functioning is of obvious evolutionary value and whose activation is strongly correlated with physical, autonomic, and other observable symptoms of fear (e.g., freezing). Indeed, LeDoux and others have mapped the fear circuit's operation in considerable detail and have made huge strides in explaining the observed capacity for associative fear acquisition, retention, and extinction by Hebbian plasticity and long-term potentiation at the cellular-synaptic level (LeDoux, 2002, pp. 79–80).

The same Hebbian picture is mirrored in the higher-level Rescorla-Wagner (RW) equations, which we shall employ in the affective component of the model. These operate not at the neuronal level but at the level of the person, or subject, where certain conditioning stimuli (the bell) become associated with specific unconditioned stimuli (the shock) through repeated pairings. There is certainly an underlying mathematical theory of neuronal function (action potentiation and firing), of which the cornerstones are the famous Hodgkin-Huxley model (Hodgkin and Huxley, 1952) and its relatives, notably the Fitzhugh-Nagamo (Fitzhugh, 1961) model. As suggested earlier, one can imagine filling in the gap between the cellular-synaptic account and the high-level RW equations with such intermediate models.[43] This is an important scientific challenge. Here, we attempt only a crude plausible synthesis of simple emotional, cognitive, and social components. But to begin at the beginning, let us examine some basic features of fear.

Fear Circuitry and the Perils of Fitness

A snake is suddenly thrown in your path. You automatically freeze. Why? From an evolutionary perspective, a reasonable hypothesis is

that we freeze (are "scared stiff") because the predators faced by our evolutionary ancestors used motion detection to home in on prey, and animals (i.e., species) that didn't freeze were wiped out.[44] Animals hard-wired to freeze enjoyed a selective advantage, in other words, and have passed the relevant wiring down as part of our genetic endowment.

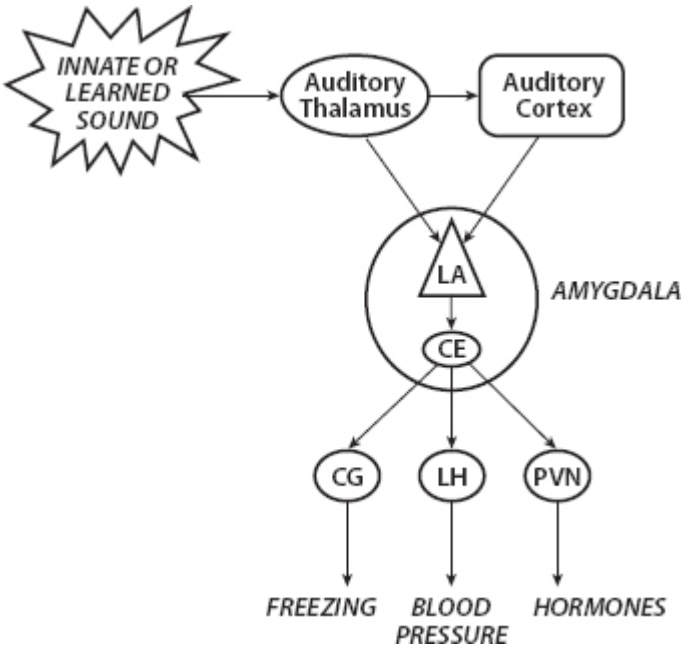


FIGURE 1. Auditory Amygdala Stimuli and Defense Responses. Source: LeDoux (2002, Figure 5.6)

Wiring: The Amygdala in a Nutshell[45]

As LeDoux writes, "The basic wiring plan is simple: it involves the synaptic delivery of information about the outside world to the amygdala, and the control of responses that act back on the world by synaptic outputs of the amygdala. If the amygdala detects something dangerous by its inputs (discussed further below) then its outputs are engaged. The result is freezing, changes in blood pressure and heart rate, release of hormones, and lots of other responses that are either preprogrammed ways of dealing with danger or are aspects of body physiology that support defensive

behaviors." (LeDoux, 2002, pp. 8–9). A simple depiction is given in Figure 1 for an auditory threat stimulus.

Having classified an auditory stimulus as threatening (innately or through conditioning), the auditory thalamus projects (emits an action potential) to the lateral amygdala (LA) and auditory cortex, which also projects a more refined signal to the LA. The central amygdala (CE) then activates various systems to produce responses, such as those shown: freezing, increases in blood pressure, and the release of various hormones. (Further responses are discussed later.)

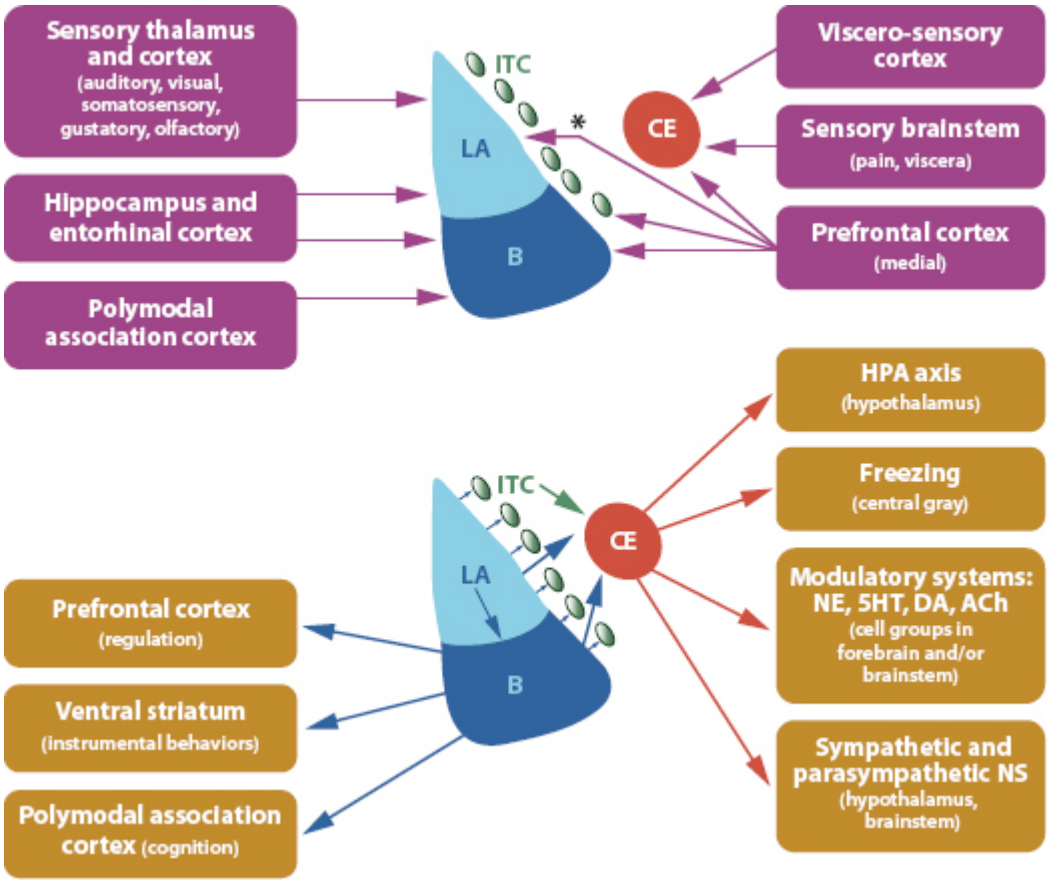


FIGURE 2. Amygdala Inputs and Outputs. Inputs to some specific amygdala nuclei. Asterisk (*) denotes species difference in connectivity. (Bottom) Outputs of some specific amygdala nuclei. 5HT, serotonin; Ach, acetylcholine; B, basal nucleus; CE, central nucleus; DA, dopamine; ITC, intercalated cells; LA, lateral nucleus; NE, norepinephrine; NS, nervous system. Source: Rodrigues, LeDoux, and Sapolsky (2009)

In somewhat greater detail, the neural mechanism of amygdala inputs and activation, and amygdala output, are conveyed in the diagrams of Figure 2. Inputs are depicted in the top, and outputs are shown in the bottom diagram (Figure 2).

The blue almond-shaped structure here corresponds to stunning micrographs of stained brain slices like the one shown below in Figure 3 (LeDoux, 2008).

One essential point is that this architecture supports a critical delay between unconscious and conscious responses to stimuli.

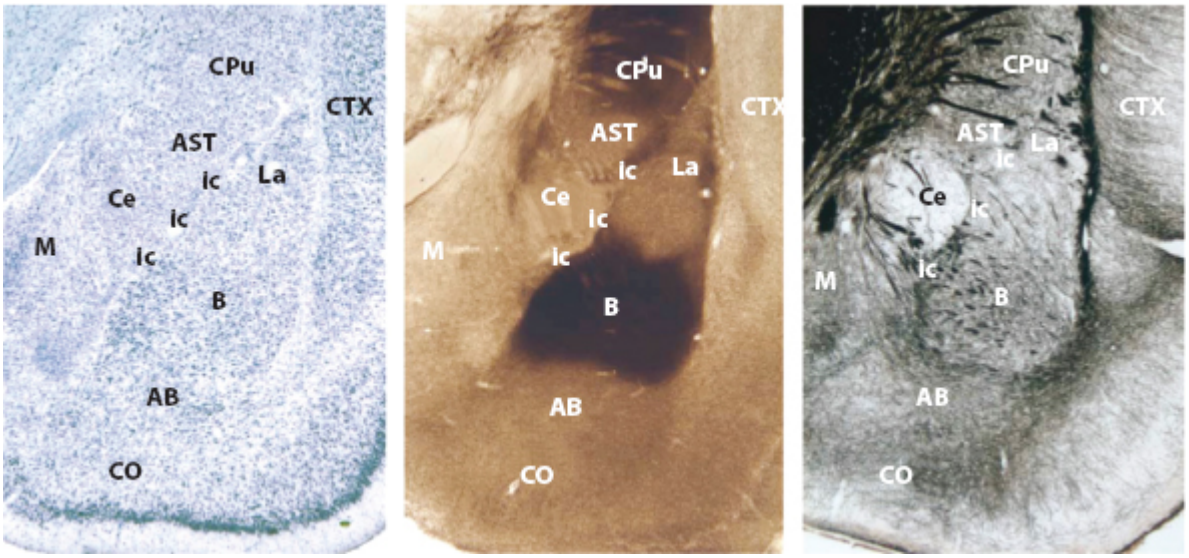


FIGURE 3. Key Areas of the Amygdala. Key areas of the amygdala, as shown in the rat brain. The same nuclei are present in primates, including humans. Different staining methods show amygdala nuclei from different perspectives. Left panel: Nissl cell body stain. Middle panel: acetylcholinesterase stain. Right panel: silver fiber stain. Abbreviations of amygdala areas: AB, accessory basal; B, basal nucleus; Ce, central nucleus; itc, intercalated cells; La, lateral nucleus; M, medial nucleus; CO, cortical nucleus. Non-amygdala areas: AST, amygdalo-striatal transition area; CPu, caudate putamen; CTX, cortex. Source: LeDoux (2008, p. 2698); reprinted courtesy of Joseph E. LeDoux

Inputs: High Road and Low Road

For example, "auditory inputs reach the lateral amygdala[46] from the auditory thalamus and auditory cortex … These provide a *rapid but imprecise* auditory signal to the amygdala. Cortical inputs from the auditory and other sensory systems … provide the amygdala with a more elaborate representation than could come from the thalamic inputs. However, because additional synaptic connections are involved, *transmission is slower*" (Ledoux, 2007). Hence, LeDoux (2002) calls these "the low road and the high road," as depicted in Figure 4.

I instantly freeze at the snake (low road) but then evaluate it as being a benign garter snake (high road), not a true black mamba, for instance. While the extreme rapidity of the unconscious response is of immense evolutionary value, we will see that, from a social standpoint, the lag between it and conscious appraisal is a decidedly mixed blessing.

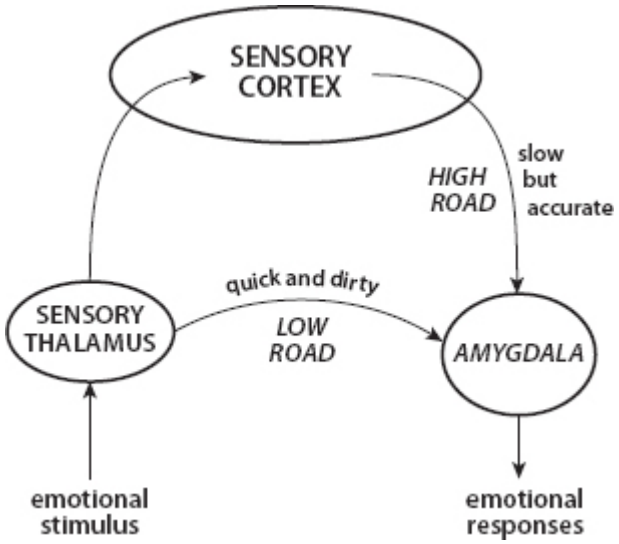


FIGURE 4. Low Road and High Road to Fear. Source: LeDoux (2002, pp. 61–63, Figure 5.7)

Outputs

Continuing, "once the amygdala detects a threat, its outputs lead to the activation of a variety of target areas that control both behavioral and physiological responses designed to address the

threat," (Rodrigues, LeDoux, and Sapolsky, 2009, p. 294). Beyond freezing, amygdala activation induces the release of numerous neurotransmitters (e.g., serotonin and dopamine), increasing arousal and vigilance. Endocrine and autonomic responses are also dramatic, "including increased blood pressure and heart rate, diverting stored energy to exercising muscle, and inhibiting digestion" (Rodrigues, LeDoux, and Sapolsky, 2009, p. 295).

> The pupils dilate to allow more light to enter. The heart rate picks up, and the heart muscle contracts more strongly, driving more blood to the muscles. Contractions of selected vascular channels shift blood away from the skin and intestinal organs toward the muscles and the brain. Motility of the gastrointestinal system decreases, and digestive processes slow down. The muscles along the air passages of the lungs relax, and respiratory rate increases, allowing more air to be moved in and out. Liver and fat cells are activated to furnish more glucose and fatty acids—the body's high-energy fuels—and the pancreas is instructed to release less insulin. The reduction in insulin allows the brain to draw off a sizeable fraction of the glucose entering the bloodstream because, unlike other organs, the brain does not require insulin in order to utilize blood glucose. The neurotransmitter that triggers all these changes is norepinephrine (Bloom, Lazerson and Nelson, 2001, p. 172).

For wonderful discussions, see also Darwin's *The Expression of Emotions in Man and Animals* (1872).[47] Contemporary scientific publications present these input-output (and additional feedback) pathways in various levels of detail. Highly detailed is LeDoux (2007).

On the experience, or "feeling," of fear, Öhman and Wiens (2003, p. 270) paraphrase LeDoux (1996):

> The fear module is primitive in the sense that it was assembled by evolutionary contingencies hundreds of millions of years ago to serve in brains with little cortices.

> However, it now operates in a human brain capable of advanced thought, language, and the conscious experience of emotion. Humans can talk about emotions, and they have emotional experiences. Awareness of an emotion not only depends on the recognition of an emotional stimulus but also originates primarily in feedback from the emotional responses that are elicited by the stimulus. For example, experiencing a racing heart when a shadow appears from a dark alley contributes to the feeling of fear. In fact, in perhaps the most classic of all classical contributions to the psychology of emotion, William James (1884) proposed that such feedback *is* the emotion. To paraphrase, you feel the emotion when you experience its effect on your body. Thus *the feeling of fear is the experience of an activated fear module.*

Current research on the neurophysiology of fear shows James to have been remarkably prescient, despite lacking any modern tools. Very importantly, from a social standpoint, the fear circuit can be activated, and *fear conditioning can occur, unconsciously.*

Unconscious Activation and Conditioning

In humans, "the fear module can be activated, and fear conditioning can occur without our conscious awareness." Indeed we need not *ever* become conscious of it. As LeDoux continues, "… unconscious operation of the brain is the rule rather than the exception throughout the evolutionary history of the animal kingdom. … And this, moreover, confers a selective advantage … if we had to consciously plan every muscle contraction our brain would be so busy we would probably never end up actually taking a step or uttering a sentence" (LeDoux, 2002, p. 11). Among the many demonstrations that amygdala activation per se need not be conscious, the so-called backward masking experiments are particularly elegant.

In backward masking, "an emotionally arousing visual stimulus is flashed on a screen very briefly (a few milliseconds) and is then

followed immediately by some neutral stimulus that stays on the screen for several seconds. The second stimulus blanks out the first, preventing it from entering conscious awareness (by preventing it from entering working memory)" (LeDoux 2002). But the first still elicits the full suite of physiological responses—increased heart rate, blood pressure, sweaty palms, and so forth. "Since the stimulus never reaches awareness (because it is blocked from working memory), the response must be based on the unconscious processing of the stimulus rather than on conscious experience of it. By short-circuiting the stages necessary for the stimulus to reach consciousness, the masking procedure reveals processes that go on outside of consciousness in the human brain" (LeDoux, 2002, p. 208). In short, the stimulus makes it to the amygdala by the quick and dirty "low road," but its arrival in working memory (the high road) *never* occurs. Cacioppo et al. (2007) write, "The amygdala is particularly sensitive to fear faces (Adolphs et al., 1999; Breiter et al., 1996) even when they are presented so rapidly as to not be consciously perceived (Morris, Öhman, and Dolan, 1999; Whalen et al., 1998). For another nice discussion of backward masking, see Penrose, 1999.[48]

As recent evidence of our capacity for unconscious conditioning proper, a very interesting study by Arzi et al. (2012) demonstrates that associative learning can occur even while we are asleep.

Delayed Feelings

If we do become conscious of fear-inducing stimuli, moreover, we may do so only *after* the physiological responses. Only after we have ducked from the darting bat do we notice that our heart is pounding, and we ask, "Whoa, what the heck was that!?" The conscious experience of fear, in other words, is a brain state induced by the *unconscious activation* of neurophysiological precursors driven by the amygdaloid complex (LeDoux, 2002, p. 208). Or, to paraphrase William James (1884), *We don't run because we fear the bear. We fear the bear because we run.*[49]

Adaptive Innate Capacity

A range of stimuli will elicit this unconscious activation—we instinctively crouch protectively at unexpected explosions nearby or when unexpected projectiles dart at our heads. In other words, certain sensory inputs will innately generate the threat response. In rats, for example, cats are in this set of innate threats. In fact, rats bred in colonies completely isolated from cats for many generations will freeze upon first exposure to cat urine (LeDoux, 2002, p. 4).

Notice, however, that animals equipped *only* with a fixed set of specific threats would be vulnerable to novel ones. So, it would be advantageous if the set could be expanded to include novel threats. And it obviously can. Pleistocene man never encountered a BMW, but *we* freeze when a car whips around the corner at us, just as *he* froze when huge animals charged suddenly from the tall brush. *We are harnessing the same innate fear-acquisition capacity—the same innate neurochemical computing architecture.* Miraculously, synaptic plasticity permits us to adapt the evolved machinery to encode novel threats. Detailed neurochemical accounts are given in LeDoux (2002, pp. 89–90).

Retention

There is little point is learning to fear hippos on Monday and then forgetting to on Tuesday. So, the *retention* of acquired fear associations is obviously essential in such cases and is achieved by various forms of long-term potentiation (LTP) at the synaptic level. This is also becoming understood neurochemically and is treated in Bauer, Schafe, and LeDoux (2002). Like unconscious fear acquisition, this fear retention is also significant socially, as we will discuss later in connection with “extinction.”

Observational Acquisition

Finally, it would also be advantageous if one could condition on the aversive experience of others—if you could acquire fear of the red-hot stove by watching me get burned, without having to get burned yourself. As we will review, this too is possible. Indeed, so-called *mirror neurons* may have evolved for this very purpose (see the discussion on p. 62). The result is that fear is, in a defensible sense, contagious (Hatfield, Cacioppo, and Rapson, 1994). This will be further discussed shortly.

All in all, then, as LeDoux observes, "It is a wonderfully efficient way of doing things. …" Rather than create a separate system to encode each new danger, "just enable the [single] system that is already evolutionarily wired to detect danger to be modifiable by experience. The brain can, as a result, deal with novel dangers. … All it has to do is create a synaptic substitution whereby the new stimulus can enter the circuits that the pre-wired ones used" (LeDoux, 2002, pp. 6–7).

Perils of Fitness

It is indeed a most wonderful machinery. But it is also terrible: it makes us deeply vulnerable to the unconscious construction and retention of racial, ethnic, and other fears and biases [on race, LeDoux (2003); Telzer et al. (2012); on racial face masking, Öhman (2005)]. It predisposes us to rash, often violent, overreactions and opens us to all manner of nefarious manipulation. Indeed, fear conditioning has been a fundamental tool in most propaganda since time immemorial.[50] But, equally disturbing, fear can spread in a completely decentralized manner, propelling mass violence, financial crises, and deeply misguided health behaviors for example. See LeDoux (2002, p. 124):

> As Pavlov suspected, defense conditioning plays an important role in the everyday life of people and other animals. It occurs quickly (one pairing of the neutral and aversive stimulus is often sufficient) and endures (possibly for a lifetime.) These features have no doubt become part of the

> brain's circuitry due to the fact that an animal usually does not have the opportunity to learn about predators over the course of many experiences. If an animal is lucky enough to survive one dangerous encounter, its brain should store as much about the experience as possible, and this learning should not decay over time, since a predator will always be a predator. In modern life we sometimes suffer from the exquisite operation of this system, since it is difficult to get rid of this kind of conditioning once it is no longer applicable to our lives, and we sometimes become conditioned to fear things that are in fact harmless. *Evolution's wisdom sometimes comes at a cost."* [Emphasis added.]

In other words, fitness is perilous: the innate fear module is double edged. The self-same rapid-fire, unconscious, nondeliberative fear-association machinery that allowed us to avoid predators on the African savannah leaves us profoundly vulnerable to manipulation, to unreflective acquisition of biases, and to being swept up in mass hysterias from Salem witches to genocides to the run on banks that precipitated the Great Depression.

Know Thyself

Self-knowledge (and self-control) requires that we recognize these powerfully evolved forces. Denying their existence simply increases our vulnerability to them and to their manipulation. That we possess this fear-conditioning apparatus is beyond reasonable dispute. This, however, is emphatically not to say that human behavior is *determined by* conditioning. First, the capacity to extinguish conditioned fear is also part of the innate human endowment, is also backed by overwhelming experimental evidence, and is also being mapped neurochemically. This is discussed later under the topic of *extinction*. But, beyond that, a central point of the present model is that unconscious conditioned fear may be modified both by conscious deliberation and (often unconscious) social influence. This is not to say that the overall outcome is necessarily "better than" the

purely fear-inspired one; only that, as a scientific matter, (a) conditioning can be transitory and (b) much beyond conditioning is going on. Fear conditioning is incontrovertibly a part of the human condition (pun intended), and it is part of my model, along with much else.

Nomenclature of Conditioning

With all this as background, we review some standard nomenclature adopted by Pavlov in his monumental study, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*.[51] This terminology is necessary to present the Rescorla-Wagner model. To begin, the following are now standard definitions.[52]

Definitions
US: unconditioned stimulus [food]
UR: unconditioned response [food-induced salivation]
CS: conditioned stimulus [bell]
CR: conditioned response [bell-induced salivation]

Initialize
CS (bell) alone → 0 (no response)
US (food) alone → S (salivation)

Associative Learning
CS–US pairing trials: bell/food, b/f, b/f, …

When Conditioned:
B alone → S … CR = UR

The US is called *unconditioned* because no conditioning is required for it to elicit the response. For Pavlov's dog, food (US) induces[53] salivation without any conditioning. Hence salivation is termed the unconditioned response (UR). The conditioned stimulus (CS), by contrast, initially elicits nothing. Pavlov (1903) actually repeated his experiments with a number of different conditioning stimuli, including the famous bell. Through repeated pairings with the US,

the CS acquires salience and eventually alone elicits a response, called, naturally, the conditioned response (CR). Because it emerges through repeated associations of the US and the CS, the conditioning process is also called *associative learning*.

Hume's "Association of Ideas"

Although a synaptic account of this process would require another 300 years of research, Hume recognized the general phenomenon of conditioned association, and even saw this as his signal contribution.[54] In *An Enquiry Concerning Human Understanding* (1748; 2008 ed., pp. 106–7), he writes "… after the constant conjunction of two objects … we are determined by *custom* alone to expect the one from the appearance of the other … Having found in many instances, that two kinds of objects—flame and heat, snow and cold—have always been conjoined together; if flame or snow be presented anew to the senses, the mind is carried by *custom* to expect heat or cold." It is not by reasoning, moreover, that we form the connection. "All these operations are a species of natural instinct, which no reasoning or process of the thought and understanding is able either to produce or to prevent" (Section V, Part I).

Hume even recognized the benefit (though perhaps not the cost) of a lag between reflexive response (low road) and deliberation (high road). He writes that, since this innate associative capacity

> is so essential to the subsistence of all human creatures, it is not probable that it could be entrusted to the fallacious deductions of our reason, which is slow in its operations; appears not in any degree in infancy;[55] and at best is, in every age and period of human life, extremely liable to error and mistake. It is more conformable to *the ordinary wisdom of nature* to secure so necessary an act of the mind, by some sort of instinct or mechanical tendency, which may be infallible in its operations, may discover itself at the first appearance of life and thought, and may be *independent of all the laboured*

deductions of the understanding [emphases added] (Section V, Part I).

It is difficult to fathom so modern a perspective from someone born (in 1711) a century before Darwin (b. 1809), who would discover that Hume's "ordinary wisdom of nature" is none other than natural selection. Through the researches of Pavlov, de Cajal, Hodgkins and Huxley, and many others, we now do know something of the "mechanical tendency" Hume intuited. But it was only in the mid-20th century that mathematical models of conditioning emerge.

En route to a very famous one of these, modern nomenclature uses *V* for the *Associative Strength* of the CS and US. It is the extent to which the CS (the Bell) elicits the UR (salivation), or, equivalently, it is the proximity of the CR to the UR.[56] Obviously, *V* changes over time, with repeated pairings, and in a manner usefully captured by the Rescorla-Wagner equations (to be presented shortly).

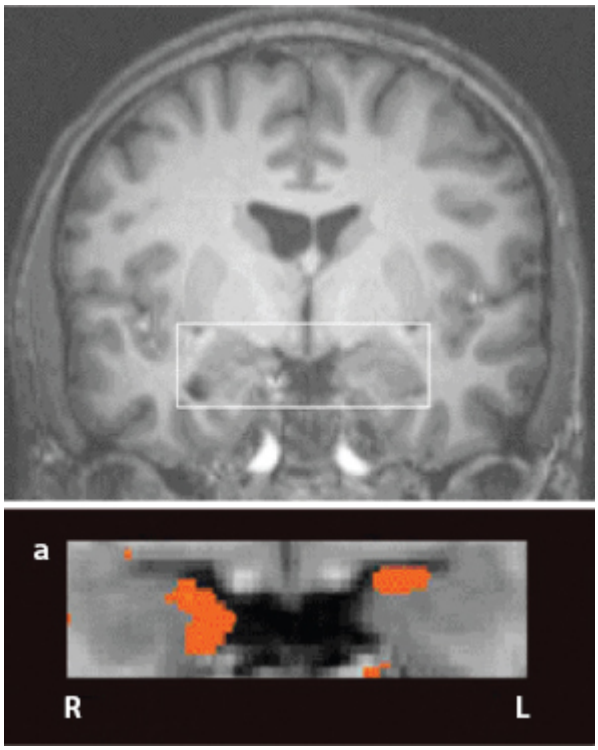


FIGURE 5. Postconditioning Amygdala fMRI. Source: Reprinted by permission from Macmillan Publishers Ltd: *Nature Neuroscience* (Olsson and Phelps 2007, p. 289), copyright 2007

One exemplary experiment (Olsson and Phelps, 2007) uses color as the CS, electric shock as the US, and repeated color-shock conditioning trials. After a number of these color-shock pairings, the associative strength (V) of color and shock is sufficiently great that the color (the CS) alone elicits the anticipatory shock fear, measured by conductive skin response (CSR) and by functional magnetic resonance imaging (fMRI). The fMRI image in Figure 5 shows the (blood-oxygenation-level-dependent) BOLD signal in the subject's amygdala on presentation of the CS after fear conditioning. This will prove to be of central interest below.

All brain imagery needs to be interpreted with great care (Vul et al., 2009a, b). A black-lung X-ray does not depict *the feeling of* respiratory distress and this fMRI does not depict *the feeling of* fear. But someone with a black-lung X-ray will almost certainly have trouble breathing. In medicine, feelings are symptoms. The instrumental readouts are signs. But signs are often correlated with symptoms, and that is my basic presumption here. There are many and varied correlates of fear, as reviewed earlier. Amygdala activation is a central neural one.[57]

Theory of Conditioning

There is a basic mathematical theory of the conditioning process that we shall adopt, recognizing that numerous refinements and extensions are possible. These are high-level—low-dimensional—equations developed in 1972 by Rescorla and Wagner. They do not *represent* the neural level at all. Their fidelity is *explained by* the contemporary neuroscience. They are analogous to the Kermack-McKendrick disease-transmission model (Kermack and McKendrick, 1927), which gives a very useful account of disease transmission through well-mixed populations without *representing* the microbiological interaction of pathogen and host immune systems,

which, of course, *explains* transmission. From a neuroscience perspective, the Rescorla-Wagner model is a highly aggregate relationship describing an associative process generated by Hebbian plasticity at the synaptic level. It is a summary relationship explained by physicochemical synaptic interactions, which are quite well understood. So, here we are representing gross fear-learning dynamics that are explained by a lower neural level. For our social science objectives, this is a suitable modeling resolution.

The Rescorla-Wagner Model

The Rescorla-Wagner model (1972) is a cornerstone in the mathematical theory of conditioning and will form the basis for *Agent_Zero*'s emotional module.[58] Though I will generalize it slightly to accommodate a broader range of conditioning trajectories, for present purposes, Rescorla-Wagner works nicely. It is

$$v_{t+1} - v_t = \alpha\beta(\lambda - v_t). \qquad [8]$$

Exposition

Imagining a fear-conditioning exercise of the sort discussed before, let t index the paired presentations of the US and CS. So $t = 1$ is the time of first pairing; $t = 2$ is the time of second, and so forth. (One could assume that trials are a Poisson arrivals process, for example. Here, we assume these are equally spaced and identical in duration and all other respects). At each pairing there is some associative strength between the CS and the US—some degree to which the CS alone elicits (i.e., the CR approximates) the UR—for instance, the extent to which the bell alone (CS) elicits the salivation (in milliliters) normally elicited with no conditioning by food (the US). This is the value of v_t, the associative strength at trial t.[59] The pretraining association is v_0 and could be positive but will here be

zero. Before any training, in other words, the bell alone elicits no salivation.

The Rescorla-Wagner model concerns the *change in* associative strength as trials proceed. The left-hand side is the *difference between* v_{t+1} and v_t. If we, advisedly, use the loaded term *learning* to denote this difference, we can coherently ask the question, When does learning stop? It stops when the left-hand side equals zero—when there is no change between v_t and v_{t+1}. The right-hand side must also equal zero, which occurs when v_t reaches the value λ, since α and β are constants (to be discussed shortly). Hence, λ represents the maximum associative strength attainable in the training process of interest and might also have been denoted $v_{\max}$. So, if the association is already at capacity, no further gain in association is possible. Hence, the difference between λ and v_t is interpreted as *surprise* (This is sometimes referred to as the subject's prediction error). Once the associative strength of chocolate and sweetness is unity, we are not surprised when chocolate is, in fact, sweet. But our first taste of chocolate is pleasantly surprising. Finally α and β are nonnegative constants representing the salience of the CS and the salience of the US, respectively. They are often termed *learning rates*. High surprise and salience can produce very rapid conditioning. For Little Albert—recalling James Watson's infamous experiment of the 1920s—the clang of a hammer on a metal bar (the US) is salient and highly aversive, while the little furry white mouse (the CS) is initially salient and snuggly.[60] It is shocking to the infant Albert that the two would be associated so—with Watson's repeated pairings of the clang and the mouse—little Albert "learns" quickly to fear the mouse. Albert, in fact, generalized this to fear all furry animals. If either the mouse or clang had lacked salience, he might have had a pet rabbit (no aversive association would have been formed).

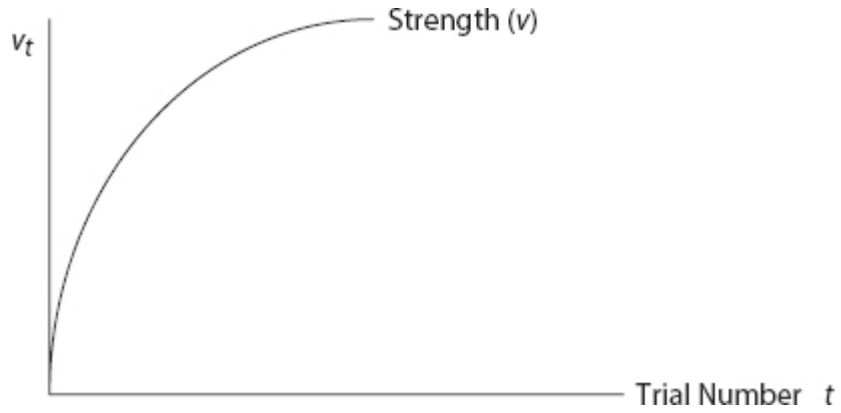


FIGURE 6. Rescorla-Wagner Associative Strength Trajectory

Observe that with, α and β positive and $\lambda \geq v > 0$, v increases with each trial, but at a decreasing rate, approaching λ asymptotically. Learning—the *change in* association—is greatest at the outset, declining as the maximum association is approached,[61] as shown in Figure 6.

Unless otherwise stipulated, the variable, t, will denote trials. The Rescorla-Wagner model is a first-order initial value nonhomogeneous linear difference equation and is solvable analytically. To wit,

$$\begin{aligned} v_{t+1} - v_t &= \alpha\beta(\lambda - v_t); v_0 = 0. \\ v_{t+1} &= (1 - \alpha\beta)v_t + \alpha\beta\lambda, \end{aligned} \qquad [9]$$

whose solution is:

$$v_t = \lambda[1 - (1 - \alpha\beta)^t]. \qquad [10]$$

The asymptotic value of v is λ, as it should be (that is, $v_\infty = \lambda$). Now, analytical solution in hand, we can begin to ask a number of interesting questions about our disposition model.

Tipping Point

For example, assuming the preceding affective trajectory, at what trial will the individual "tip" into taking action? In the model

(without deliberative or social modules yet), this occurs when $v_t > \tau$, the action threshold, as shown in Figure 7.

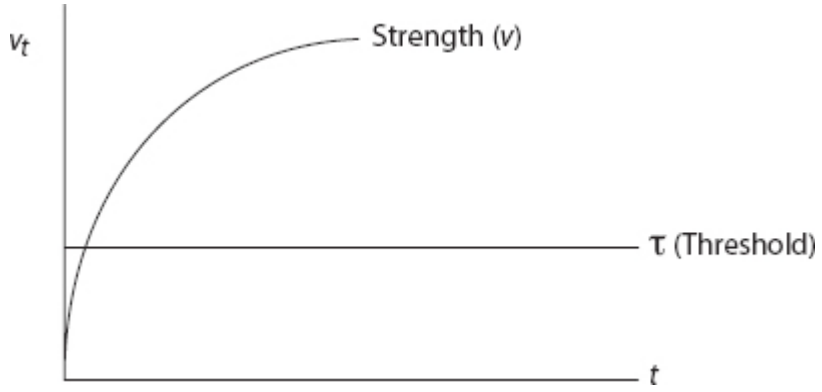


FIGURE 7. Rescorla-Wagner Trajectory and Threshold

Using the general solution, this is equivalently

$$\lambda\left[1-(1-\alpha\beta)^t\right]>\tau.$$

We solve for the threshold time:[62]

$$t^*>\frac{\ln\left(1-\frac{\tau}{\lambda}\right)}{\ln(1-\alpha\beta)}. \qquad [11]$$

This makes good sense. Increasing the threshold delays the tipping. If either α or β is zero, tipping never occurs since

$$\lim_{(\alpha\beta)\to 0} t^*=\infty,$$

meaning that if both stimuli (CS and US) lack all salience, the action is never tripped.

For completeness sake, and because we will employ it below, the differential equation (as against the difference equation) version of the Rescorla-Wagner model is

$$\frac{dv}{dt}=\alpha\beta(\lambda-v). \qquad [12]$$

With $v(0) = 0$, the solution is

$$v(t) = \lambda (1 - e^{-\alpha\beta t}). \qquad [13]$$

The central point of the Rescorla-Wagner model, in either form, is that *high surprise combined with high salience produces strong associative conditioning.* To anticipate slightly, another distinctive feature is that conditioning depends on the aggregate stimulus, the sum of the v_i's—the associative strengths taken over all stimuli—a point to which we will return.

Once again, the model does not assume that people are *aware of* these affective dynamics. Indeed, a central point of the preceding discussion is that, typically, they are not (Phelps in Lewis, Haviland-Jones, and Barrett, 2008, p. 236; LeDoux, 2002). We may not be conscious of our conditioning (i.e., the reduction in prediction error $\lambda - v$) even if we are conscious of the CS-US pairings. And, in backward masking, even these are not registered.

Social Examples

We are well acquainted with conditioning trials in which the CS is a bell, the US is food, and the CR is salivation or where the CS is a light, the US is a shock, and the CR is fear. But, there is every reason to postulate analogous patterns of profound social consequence, as suggested in Table 1. It is surprising and profoundly salient when an unfamiliar social group attacks others, or when a trust is betrayed. Associations born of shocking and salient social events can elicit extremely strong reactions, as suggested in Table 1.[63]

Recognizing the model's generality, let us focus on one of our social examples. The September 11, 2001, World Trade Center attacks were surprising and salient. One could argue that there were four primary conditioning trials, one for each tower plus one for the Pentagon attack and one for Pennsylvania flight 93. In fact, there were countless further exposures in the form of video replays of the aircraft impacts and subsequent collapse of towers, people leaping from buildings, terrified flight, and other images. The unconditioned

response (UR) was fear and intense anger toward the perpetrators. The conditioned stimulus (CS) was the face of Osama Bin Laden or Mohammed Atta—the "Arab face," as it were.

Table 1. Surprise, Salience, and Conditioning

CS	US	UR/CR
Light	Shock	Fear
Japanese face	Pearl Harbor	Rage/internment
Arab face	9/11	Rage/internment
Vaccine	Report of adverse reaction	Fear/vaccine refusal
Doctor	Tuskegee	Enduring distrust
Asset	Collapse	Dumping

The resulting associative strength was extremely high, as expected on good Rescorla-Wagner (RW) grounds.[64] Most Americans had little prior exposure to Muslims and certainly had never heard the phrase "Al Qaeda" before, so there was no damping of the association by prior conditioning. And, as expected, we saw rapid "learning," in the RW sense. After the four direct trials and countless reexposures in all media, the average CR to symbols of the Muslim world (CSs) was very high up the learning curve.

"A comprehensive LexisNexis database survey of U.S. newspaper reports between September 1 and October 11, 2001, found an increase in hate crimes toward persons believed to be of Middle Eastern descent (from 1 to 100 events involving 128 victims and 171 perpetrators) across 26 states" (Swahn et al., 2003; Marshall et al., 2007, p. 311). Fourteen of these were murders. "Most [attacks] occurred within the period 10 days after the 9/11 attacks" (p. 311).

Remarkably, "only 42% of the victims were of Middle Eastern descent," the remaining attacks being "against persons of color who are perceived to be vaguely reminiscent of the 9/11 terrorists" (Marshall et al., p. 311). Even a very broad and vague attribute (general skin hue) can serve as a CS. This is an example of stimulus overgeneralization, where subjects conditioned on a particular CS—

an 800-hz tone—will respond to a very rough approximation of it (e.g., a 1000-hz tone).

Olsson and Öhman (2009, p. 736) write, "For example, *there are now numerous demonstrations that unknown racial outgroup members, that is, individuals not belonging to one's own racial group*, can elicit a rapid threat response associated with the amygdala" (Cunningham et al., 2004; Phelps et al., 2000). They even speak of "the possibility of a hard-wired disposition to develop xenophobic responses" (p. 736). This is not to say that xenophobia itself is inevitable. Indeed, they also note that out-group dating experiences can nullify the effect. Why this in-group bias might have evolved is modeled in Hammond and Axelrod (2006a, b). The first fMRI study of prejudice was Hart et al. (2000; see also Cunningham & Van Bavel, 2009, p. 978).

Blocking and Selective Discrimination

Relatedly, I find it very revealing that the Japanese were the only ethnic group interned on a mass scale by the United States during WWII, even after 1944, when the United States was fully at war with Germany and Italy. In 1939, The German-American Bund had staged a 20,000-person pro-Nazi rally in New York's Madison Square Garden. The Bund ran a dozen youth camps in various states and published eight newspapers. Once the United States entered the war, the Bund was banned, but, unlike the Japanese (who had never organized anything comparable), few German-Americans were interned. Even fewer Italian Americans were interned, despite fascist Italy's alliance with Hitler. Beyond the unique scale of their internment, the Japanese were distinctive in that more than 60% of those interned were, in fact, American citizens.[65]

This, too, is entirely consistent with the general version of the Rescorla-Wagner model, in which the total associative strength v^{TOT} is distributed over the *sum of* all conditioning stimuli of relevance. If we let v_t^J stand for the associative load on the Japanese at time *t*, and v_t^E. the load on European axis powers, the total strength is given by[66]

$$v_{t+1}^{TOT} - v_t^{TOT} = \alpha\beta\,[\lambda - (v_t^E - v_t^I)]. \qquad [14]$$

If the associative strength of v_t^I is already close to λ, there is very little associative capacity left for v_t^E. In these terms, the associative load on the Japanese face was so large after Pearl Harbor (shocking and salient) and the ensuing war in the Pacific as to "block" a comparable association on Aryan features or Italian accents.

Interestingly, very few Japanese *in Hawaii* were interned.[67] It could be that they were grudgingly tolerated as essential to American naval base operations. But one could also explain this by Rescorla-Wagner: Japanese people were part of the fabric of Hawaiian society, composing a third of the population. Non–Japanese Americans living in Hawaii had accumulated sufficient positive experience (prior exposure) as to "block" the level of fear that continental Americans associated with the (far less familiar) Japanese after Pearl Harbor. Analogously, on 9/11, most Americans had no such prior exposure to Muslims, and no blocking of the maximal associative load occurred. The phenomenon of blocking has been studied extensively, beginning with the classic paper of Kamin (1969).

Betrayals Real and Imagined

Betrayals of trust are often very surprising and salient. The Rescorla-Wagner model may explain why they generally loom so large in human memory. The betrayal of trusting black Americans by the medical establishment at Tuskegee is a stark example. It was very surprising and highly salient. It actually continued until 1972 and so is a deep trauma well within the memory of black Americans today. Judas betrays Christ; Brutus betrays Caesar; Greek mythology is rife with betrayals (Clytemnestra betrays Agamemnon); "Uncle" Joe Stalin (ally in WWII) betrays the war allies by occupying Eastern Europe; the Jews allegedly "stabbed Germany in the back" after WWI.[68] Benedict Arnold betrays the colonies. They are instances of highly salient surprise. The "revelation" of betrayal by conspiracies is a trusted tactic among fear mongers to this day. See Richard

Hofstadter's (1964) wonderful essay, "The Paranoid Style in American Politics."[69]

By the same token, some salient surprises are reserved for occasions meant to elicit a burst of strong and happy associative strength—like marriage proposals. Many expectant parents choose not to learn the sex of their children till birth, preserving a highly salient surprise.

So, here we have our first building block of *Agent_Zero*—Rescorla and Wagner's very elegant model of conditioning. We understand that this is an aggregate relation that is ultimately explained by the neuroscience, which *licenses us to interpret the model* as unconscious fear acquisition harnessing the same Pleistocene apparatus that got us here,[70] a neurophysical apparatus that was reviewed in some detail.

Now, as noted, a number of factors can militate against our blind submission to unconscious fear impulses. Counterevidence is one; peer pressure is another. These will both be further building blocks added to the model. But, even within the Rescorla-Wagner framework, there is allowance for the fading of fear.

Fear Extinction

Our synapses are plastic, and so are we—we can learn, and we can unlearn (Rescorla and Wagner, 1972). Conditioned associations are not necessarily permanent and often decay if pairing trails cease. This is called *extinction,*[71] a term introduced by Pavlov. In the Rescorla-Wagner (RW) model, the extinction phase is handled very elegantly simply by setting λ, the limiting value of v, to zero (since all association is to die out), and the initial value of v to whatever value it had attained immediately when conditioning trials are terminated; let us denote this latter value v_{max}. In differential equation form (moving freely between discrete and continuous time), associative strength thus diminishes according to

$$\frac{dv}{dt} = \alpha\beta\,(0 - v), \text{with } v(0) = v_{max}. \qquad [15]$$

The solution is the well-known formula for exponential decay,

$$v(t) = v_0 e^{-\alpha\beta t} = v_{\max} e^{-\alpha\beta t}. \qquad [16]$$

Overall, then, the conditioning and extinction phases of an RW process are *not* symmetrical, and most likely involve different brain regions [as discussed shortly]. Conditioning is increasing and concave down, with an upper asymptote of λ. Extinction is decreasing and concave up, with an asymptote of zero. When conjoined the acquisition and extinction phases have a distinctive shape, with an abrupt change in concavity at the (nondifferentiable) acquisition-extinction transition point, as shown in Figure 8.[72]

General Solution of the Two-Phased Model

Typically, the two phases (acquisition and extinction) are solved and discussed separately. I have not seen it observed that the entire two-phased model can be expressed using Heaviside unit step functions.[73] With t^* the time at which trials cease, the full solution is then

$$v(t) = \lambda\,(H(t^* - t)(1 - e^{-\alpha\beta t}) + (1 - e^{-\alpha\beta t^*})H(t - t^*)e^{-\alpha\beta(t - t^*)}). \qquad [17]$$

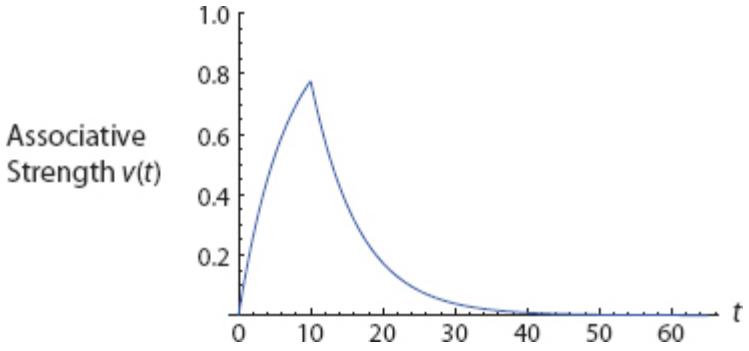


FIGURE 8. Acquisition and Extinction

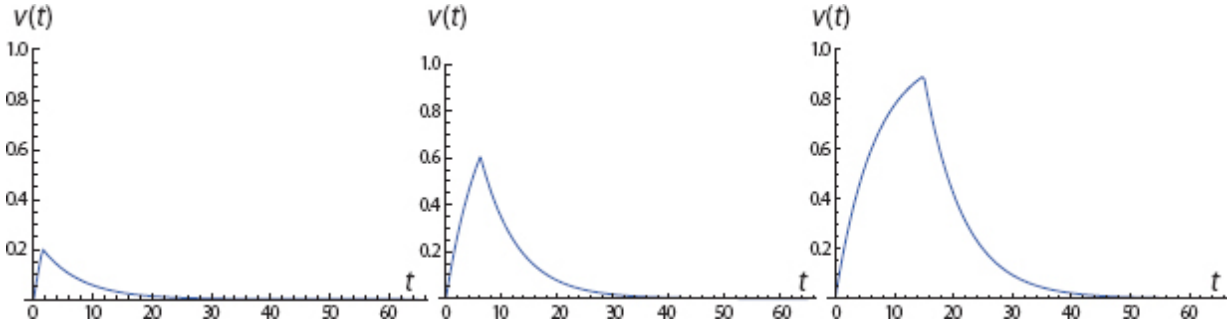


FIGURE 9. Final Acquisition Level and Initial Extinction Rate

This gives the entire learning and extinction trajectory. A movie of the entire acquisition and extinction history, as parameters are varied, is given as **Animation 0** on the book's Princeton University Press Website.[74]

An impressive property of the two-phased model is that the (negative) extinction *slope* increases in magnitude with the terminal phase 1 conditioning *level*, as noted in the panels of Figure 9. The greater the acquisition *level*, the greater (i.e., steeper) the initial extinction *rate*.[75]

The curious asymmetry of the model makes its extensive corroboration all the more impressive. Countless experiments with animals, including humans, have conformed to this basic relationship. For example, conditioning trials with humans exhibit the same qualitative profile as the following rat trials (Figure 10).

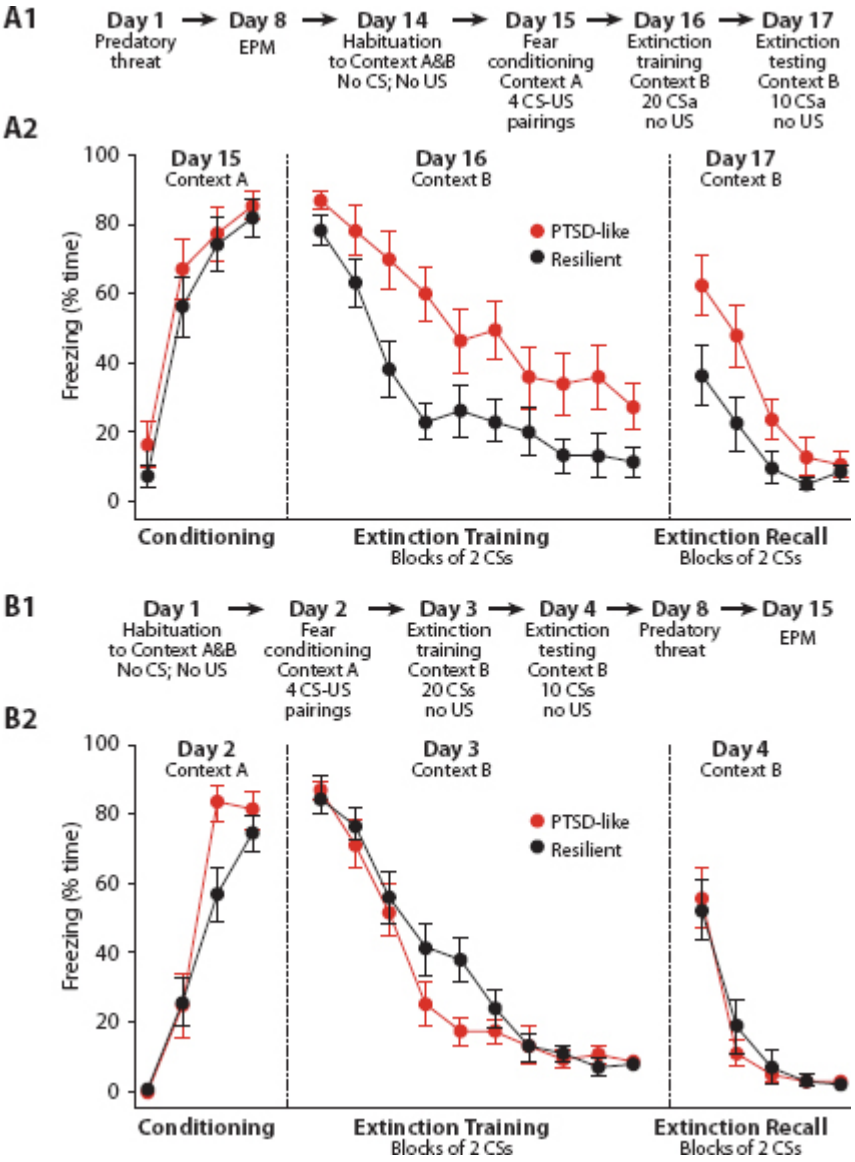


FIGURE 10. Acquisition and Extinction of Conditioned Fear Response for Predatory Threat. Source: Goswami et al (2010, p. 496)

Moreover, there are strong arguments as to why the animal models are reasonable predictors of human fear conditioning behavior (Bloom, Lazerson, and Nelson, 2001). Obviously, we do not fear *what* the rat fears, but we fear *how* the rat fears.[76] The same neuroanatomy and cellular-synaptic mechanisms have been preserved across vertebrate evolution. LeDoux calls these conserved structures "survival circuits." In his most recent terminology, the fear circuitry we've reviewed is cast as an instantiation of these. For

a full account see LeDoux (2012). A compendious review of the human conditioning literature is Sehlmeyer et al. (2009).

Although the Rescorla-Wagner model has been refined in Pearce and Hall (1980), Sutton and Barto (1998), and other descendants, the classic model deserves the status of other canonical models—the Kermack-McKendrick model of infectious disease; the Lotka-Volterra model of predatorprey cycles, the Richardson arms-race model, and so forth. Like fundamental models in many fields, it elegantly offers important insight and explains a wide range of observed phenomena.[77] It is a revealing simple model.

Affective Persistence: The Half-Life of Hatred

Because extinction is seldom immediate, affect (positive or negative) can remain above the action threshold long after the stimulus has stopped. Cycling back to our social examples, then, anti-Japanese sentiment generally continued beyond the war. The informal Jewish boycott of German goods persisted (indeed persists) long after WWII. As public health examples, the scars of Tuskegee still affect minority trust of the U.S. public health establishment (Corbie-Smith, Thomas, and St. George, 2002; Freimuth et al., 2001). Such distrust is evident in The Pittsburgh Barbershop study (Using Social Norms to Attack Prostate Cancer among African Americans, National Center on Minority Health and Health Disparities), in survey results on smallpox vaccine refusal (Lasker, 2004), and in the Washington, D.C. postal workers' cipro (ciprofloxacin) refusal after the anthrax attacks of 2001(Quinn, Thomas, and Kumar, 2008; Quinn, Thomas, and McAllister, 2005). The last example is stark in that predominantly white Congressional staff were eager for cipro. The same general pattern occurred with H1N1 (swine flu) vaccine in 2009, despite the swine flu being declared a global pandemic by the World Health Organization.

Fear conditioning and the extinction of fear, in other words, are not symmetrical, a point made nicely by the Rescorla-Wagner model. It is amazing that seemingly remote processes can share the same mathematical description [e.g., the wave equation; see J. M.

Epstein (1997)]. Even in social science, a huge number of social situations have the form of a Prisoners' Dilemma, or a Coordination Game.[78] Here also, the extinction phase of Rescorla-Wagner is formally the same as radioactive decay. So, just as we could compute the tipping point earlier, let us compute the "half-life" of hatred, if you will. The half-life is, by definition, the time at which half the original "substance," v_{max}, is gone. It is the time at which

$$\frac{v_{max}}{2} = v_{max} e^{-\alpha\beta t}. \qquad [18]$$

An interesting property of exponential decay is that the half-life is independent of the initial level, as in this case, where the v_{max}'s cancel out.

Accordingly, logging both sides, we obtain

$$\ln\left(\frac{1}{2}\right) = -\alpha\beta t,$$

which is to say that the half-life is

$$t_{half} = \frac{\ln(2)}{\alpha\beta}. \qquad [19]$$

This makes basic sense in that the smaller the decay rate ($\alpha\beta$), the greater the half-life (i.e., the longer it takes until half the initial stuff is gone).

Posttraumatic Persistence

Of course, the mere cessation of conditioning trials is not always sufficient to "reset lambda to zero" and induce the exponential extinction of fear. As LeDoux and Phelps write, "It is important to note … that the extinction of conditioned fear responses is not a passive forgetting of the CS-US association, but an active process, often involving a new learning" (LeDoux and Phelps, 2008, p. 164). In fact, acquisition and extinction of fear may be controlled by

different regions—acquisition by the amygdala, and extinction by the ante-rior cingulate of the medial prefrontal cortex (mPFC). A 2005 PET study of women with PTSD resulting from childhood sexual abuse revealed "decreased function or failure of activation in mPFC during fear extinction, in women with abuse-related PTSD compared with controls" (Bremner et al., 2005).

Again, we are not modeling brain regions, but in modeling terms, we have license to say that λ doesn't necessarily reset to zero when conditioning trials cease (simply because genocidal violence stops, for example).[79] Below, we exercise this license mathematically (see Figure 33) and show that a single agent's PTSD can retard the recovery of the entire group. Then, in the agent-based model of Part II, we (I believe for the first time) "lesion" an agent—excising her "software amygdala"—and show the group-level effects.

The Rescorla-Wagner model will be generalized in several ways below. But it will form the backbone of *Agent_Zero*'s affective component. We turn now to the cognitive (evidentiary/deliberative/ratiocinative) building block of *Agent_Zero*, having agreed with Hume and countless others, that reason (not only passion) must play a role in any credible model of people.[80]

I.2. REASON: THE COGNITIVE COMPONENT

However, reason is not here assumed to be perfect, but prone to informational limits and associated biases. There is, of course, a vast literature on *bounded rationality* since Herbert Simon coined the phrase (see Simon, 1982). One can imagine endowing agents with innumerable sources of error. Well-established and systematic departures from canonical rationality include: representativeness and availability biases, anchoring and adjustment, recency effects, the conjunction fallacy, confusion between frequency and magnitude, baserate neglect, and outright logical confusions, to name a few (see Gilovich, Griffin, and Kahneman, 2002; Kahneman, 2003). To start, however, I need the agents to estimate a

PART II

Agent-Based Computational Model

IN AGENT MODELING, we essentially build artificial societies of software individuals who can interact directly with one another and with their environment according to simple behavioral rules. On agent-based modeling in general, see, for example, J. M. Epstein and Axtell (1996), Axelrod (1997a), Resnick (1994), J. M. Epstein (2006), Tesfatsion and Judd (2006), Miller and Page (2007), and the large literature cited in these works.[118]

I developed this model in *NetLogo* 5.0. Source Code for the canonical[119] Parable 1 run is given in Appendix IV. A table of parameter values for every run is also provided. As earlier noted, all movies are posted on the book's Princeton University Press Website. Interactive Applets for each movie run are provided there as well. The Applets allow the user to alter various assumptions with "sliders," movable bars on the Interface. These user-adjustable parameters include the attack rate, search radius, extinction rate, memory length, and damage radius, for example. This offers nonprogrammers an extensive basis for experimentation with the model. For programmers, the Applets also include the full source code for every run. Hence, all results are certainly replicable. However, the English-language exposition that follows is meant to be sufficient to permit replication by reasonably adept programmers (who are also good readers).

Replicability

Apropos of this, I am not sure replicability is an attribute of models proper. Leaving aside the case of authors who are literally pretending to have a model, one could always "replicate" model

output by running the same model on the same inputs. So, when a person says a model was not replicable from some article, they are really asserting that the author's *English-language exposition* of the algorithm was insufficient to permit a reimplementation by that particular reader. If so, it would appear to measure the author's facility in English—or the reader's lack thereof—but it has nothing to do with the actual computer program or mathematical equations, which—if provided, as here—are replicable ipso facto. In any event, such ambiguities as may arise can be resolved by reference to the code provided in Appendix III and on the book's Princeton University Press Website.

Present Interpretation

Later, I will offer a number of alternative interpretations of the model in the fields of health behavior, economics, network science, and law. But for expository purposes, we imagine a conflict, indeed a guerilla war like Vietnam, Afghanistan, or Iraq. As discussed in the Introduction, events transpire on a 2-D population of contiguous yellow patches, each of which represents an indigenous agent. Specifically, we imagine that a single stationary indigenous agent occupies each patch. This expository grid is 33 by 33 (the default *NetLogo* dimensions), so there are 1089 Yellow agents.

These indigenous patch-agents do not move. They have two possible states: inactive and active. At any point in time, they occupy only one of these states. Inactive agents are yellow. I have given them slightly different shades of yellow just so they are visually distinguishable squares, as shown in Figure 34. Active agents are orange. These agents activate randomly, at a rate (the attack rate) adjustable by the user. They will be discussed shortly. The three Blue agents represent occupying forces and are of the full *Agent_Zero* type. They are mobile. Every cycle through the (randomized) agent list, the agent adopts a random heading and takes one step in that direction. So, they do not jump to random distant sites but move to random neighboring ones. They execute a 2-D random walk, in short. It is not a perfect mixing, or mass-action

kinetics, process. The space is a finite bounded square lattice.[120] These Blue "rovers" are connected to one another bidirectionally, as indicated by the two-headed arrows shown in Figure 34. (Hence, there are in-degree and out-degree distributions, and so forth). This exactly parallels the mathematical networks developed above. Some rovers give high weight to other rovers; some do not (see the Appendix IV table, or *NetLogo* Code in Appendix III for the values employed).

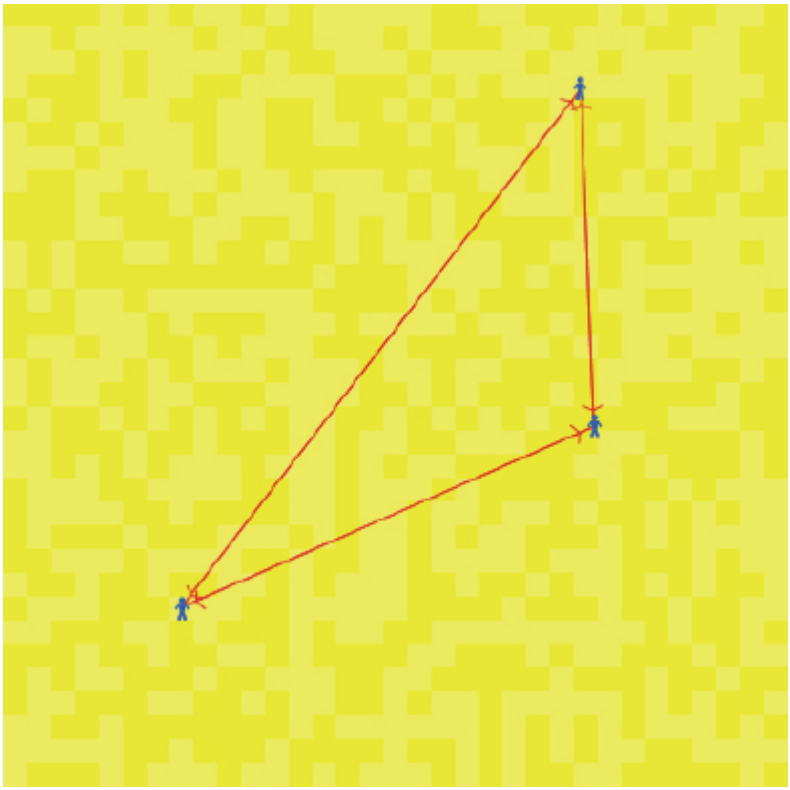


FIGURE 34. Indigenous Population (Stationary Yellow Squares) and Occupying Rovers (Mobile Blue Agents) [Movie 1]

Minimalism

I have hand-coded three agents to ensure complete control over specifics in the smallest possible model that can exhibit majorities. For large *n*, one would of course initialize the agents with random

weights and other parameters drawn from distributions. The agents and their connections are shown in Figure 34.

Movie 1 (on the Princeton University Press Website) simply shows the three agents in random motion connected to one another.

Now, let us posit a distinguished region of the space—in this version, the northeast quadrant—where yellow patches "activate" at a user-specified random rate (a global constant, implemented as a user-adjustable slider in the *NetLogo* interface).[121] Think of these agent activations as insurgent attacks. These explosions are shown as orange patches in Figure 35. I make no assumptions whatsoever as to the comparative legitimacy of occupying or insurgent agents. **Movie 2** shows these, with fixed Blue agents.

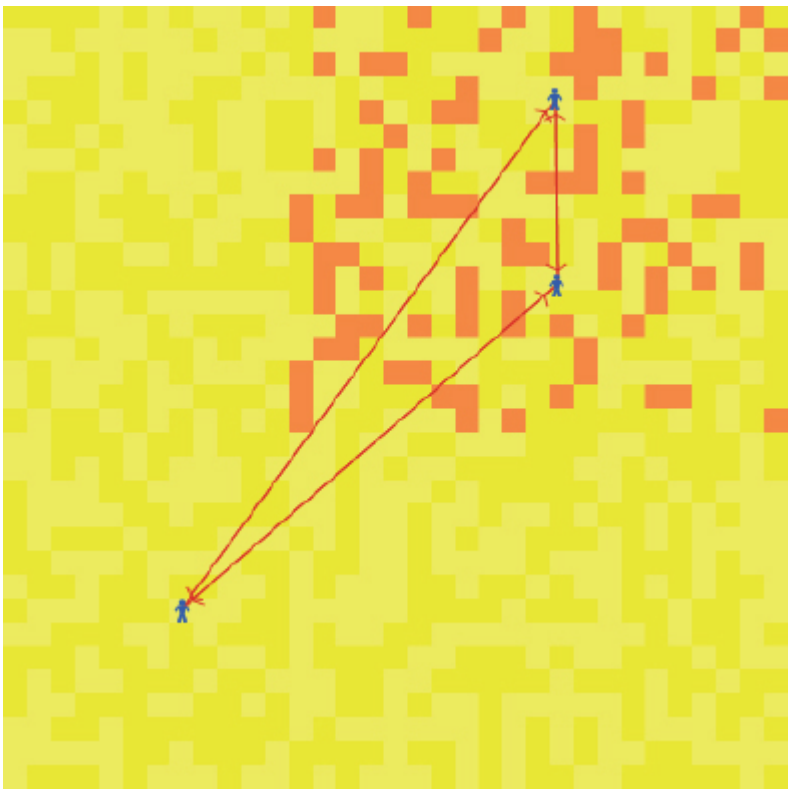


FIGURE 35. Adverse Event Activations [Movie 2]

It is central to distinguish between a Blue agent's separate affective, "rational," and social components and to understand how they are combined to form the agent's overall disposition in the wake of attack. As before, once this disposition is formed, it is compared to the agent's threshold. If the overall disposition exceeds

threshold, action is taken; otherwise it is not. In this interpretation, *action is the destruction of indigenous (Yellow) agents within some user-specified damage radius* (again, a slider in the *NetLogo* Interface). Now each component is described.

Affective Component

These orange activations (explosions) are the conditioning trials for the Blue agents. When a Blue agent "steps on" an orange patch, he updates his affect through the generalized Rescorla-Wagner equations.[122] Learning rate parameters (the α's and β's), limiting values for associative strength (λ), and the exponent (δ) all affect individual learning curves and can vary across agents. An extinction rate is applied at every iteration except those in which an active patch is encountered. So, hostile affect toward the indigenous population evaporates at a user-specified extinction rate in the absence of local attacks.[123] This extinction rate can be zero since extinction, as noted earlier, is by no means assured simply by cessation of trials. This is the *affective* component of the Blue agents' disposition.

A suitable extension would be to include the well-established contextual conditioning that Blue agents would presumably undergo in the course of their spatial movements. They would come to associate the northeast quadrant itself with danger, and this would amplify the estimates made purely from event sampling. The hippocampus is central to this well-established contextual conditioning in space. In animal models, Knierim (2009) and Knierim and McNaughton (2001) have used "multi-electrode arrays to record the extracellular action potentials from scores of well-isolated hippocampal neurons in freely moving rats. These neurons have the fascinating property of being selectively active when the rat occupies restricted locations in its environment. They are termed *place cells*, and it has been suggested that these cells form a cognitive map of the environment (O'Keefe and Nadel, 1978). The animal uses this map to navigate efficiently in its environment and to learn and remember important locations" (from Knierim Research

Page, Johns Hopkins Mind/Brain Institute site). *Agent_Zero* agents do not have a mental map of the area and condition only on the event stream, not also on position, though an Agent 1.0 could certainly have this endowment.

"Rational" Component

Turning to the evidentiary/ratiocinative component, Blue agents have a *spatial sampling radius* (which can be heterogeneous but is also a slider in the *NetLogo* Interface), within which they conduct local sampling of *the landscape*,[124] here interpreted as an indigenous population.[125] As discussed earlier, they estimate the probability that an agent is a hostile agent (e.g., the probability that an agent is a terrorist given that he is Muslim) by comput-ing the relative frequency of orange patches within their sampling radius.[126] Obviously, this probability estimator exhibits *sample selection error*—the local ratio may be a poor estimator of the global one.

Some readers may feel that this simple computation is putting "reason" at an unrealistic disadvantage to "the passions." In fact, while this sample estimate is crude statistically, its computation is remarkably sophisticated cognitively. Indeed, this imputes to the Blue agents more cognitive capacity than untrained humans possess. In *The Mathematical Brain,* Butterworth (1999) makes a powerful argument that among our innate universal endow-ments is a *number module*, giving us the capacity to make crude numerosity judgments; and he provides evidence that the parietal lobes are centrally implicated. So, just to be shamelessly phrenological, while *Agent_Zero* walks into an ambush, his amygdala is activated and so he registers fear, but his number module is also making a very crude frequency judgment: *enemy/total*. Butterworth argues that even this simple relative frequency is very hard for humans to compute, which suggests a neural basis for one of the best-documented biases in all of psychology: base rate neglect (Kahneman and Tversky, 1973; Tversky and Kahneman, 1982). As he puts it, "we ignore base rates because we ignore rates" (Butterworth, 1999). So, simple as it seems, *Agent_Zero*'s computation of a local ratio is far from trivial.

Another factor that would corrupt the Blue agent's estimate of the actual local ratio (itself a biased estimator of the global one) is the specific areal pattern in which the activations present themselves. In experiments, human subjects are quickly shown two spatial arrangements of dots: one has them spread over a wide area, and the other has them tightly packed. We will judge the former pattern to be the more numerous (Krueger, 1972, 1982). One can imagine how this areal bias might have conferred a selective advantage—we are more vulnerable (and so more alert) if surrounded by predators than if they are all clustered within our vision (giving us more escape routes).

It also happens that, even if our dots occupy the same total area, random patterns (as in this model) are typically overcounted as against regular ones, again perhaps because unpredictable predator patterns are harder to anticipate and evade than regular ones.

Related mechanisms may explain why we involuntarily complete patterns like those shown in Figure 36. The seminal example is the Kanizsa triangle (Figure 36A), after Italian psychologist Gaetano Kanizsa (1955).

This "phantom edge phenomenon" (seeing an outline that is not actually there) is due to what neuropsychologists call the "T-effect."

> Groups of neural cells see breaks in lines or shapes, and if given no further input, will assume that there is a figure in front of the lines. Scientists believe that this happens because the brain has been trained to view the break in lines as an object that could pose a potential threat. With lack of additional information, the brain errs on the side of safety and perceives the space as an object. The circle is the most simple and symmetrical object, so the mind usually sees a circle unless active effort is made to see an alternate shape. This illusion is an example of reification or the *constructive* or *generative* aspect of perception, by which the experienced percept contains more explicit spatial information than the sensory stimulus on which it is based (Ehrenstein illusion, n.d.).

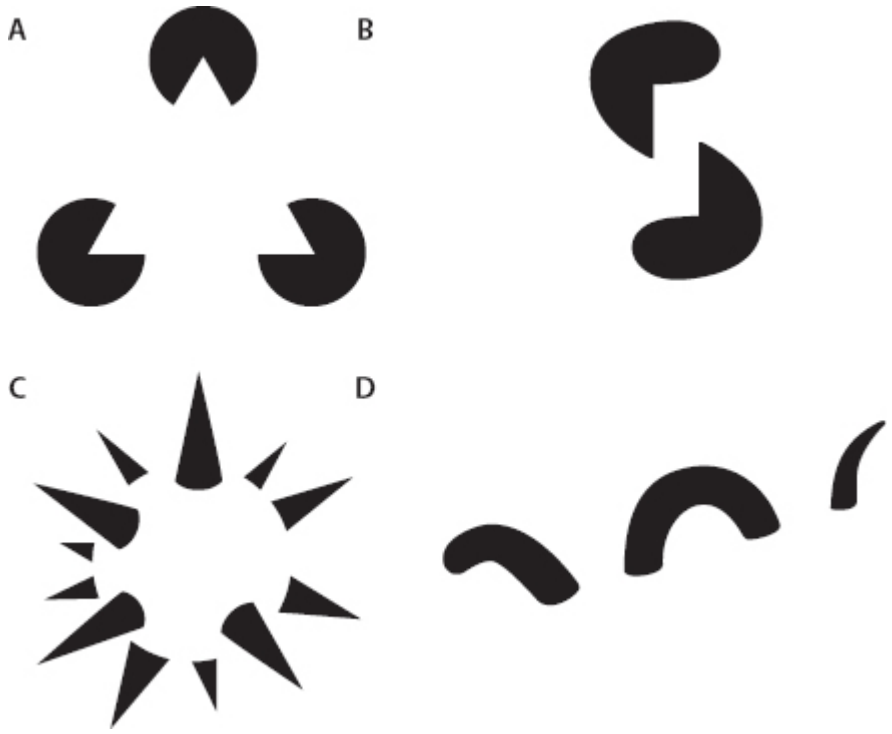


FIGURE 36. Phantom Edges

This is yet another source of potential Blue agent "threat inflation" that we shall ignore. I would say that propaganda generally—in "completing" political patterns that aren't there or inviting their completion—traffics on this same apparatus. Indeed, the entire art of propaganda is to offer as little of the picture as possible, leaving it to the audience to "fill in the blanks" opened by vague outlines of subversive "others."

Finally, the base model treats the affective and statistical estimates as independent when they are almost certainly entangled. There is interesting experimental work on the classification (as hostile or peaceful) of inconclusive data, specifically under circumstances of threat (Baranski & Petrusic, 2010). Affect, in other words, colors one's probability judgments, particularly in settings of the sort we have posited.[127] I introduce this in an extension of Part III. By contrast, in the basic model, these are superposed but decoupled—neither is a mathematical function of the other.

In sum, Blue agents simply compute the relative frequency of orange patches within their spatial sampling radius to estimate the likelihood that patches are immanently violent. This initially

appears to be a very crude algorithm. In fact, it would probably be way beyond most humans, particularly in the stressful circumstances of interest here. But, since we want to give reason a "fighting chance" against passion, we'll start here. So, we now have an elementary type of bounded rationality, in addition to a simple representation of affect.

Social Component

The third *Agent_Zero* ingredient is social. At any time *t,* the total disposition of each agent is the sum of her affect, $V(t)$, and her local probability estimate, now a function of time, $P(t)$, plus the sum of each other agent's weighted solo disposition (each the sum of their own V and P), all minus her threshold.[128] Unless otherwise noted, the term *disposition* will denote *net disposition* in all *NetLogo* graphical output.

Sampling and Dispositional Radii Mathematically Independent

It is important to reiterate that the mechanism of influence in the model is not behavioral imitation, even if agents are within the narrow spatial sampling radius of one another. In Part III, an extension offers a way to introduce this distinction. But we do not use it in the main development. Agents can influence each other (have dispositional weight) at *any* range, by a large variety of avenues (e.g., auditory and textual social media), and the binary actions of other agents can alter the landscape (by destroying sites), which can affect one's frequency calculation. But binary action proper is not registered or, therefore, imitated. The spatial sampling radius is typically a cluster of contiguous sites on the landscape proper, such as a Von Neumann neighborhood. This spatial sampling radius is bounded and landscape specific. The radius of dispositional contagion is neither; the two are mathematically independent.[129]

Action

If overall disposition is greater than the threshold, action is taken: the agent destroys all patches within a user-specified damage radius (another slider).[130] Destroyed patches (indigenous agents) are colored *very* dark (i.e., blood) red and cannot be active (they are dead).

Pseudocode

So, for each Blue agent, the algorithm (pseudocode) is as follows:

Compute own affect (with orange explosions as conditioning trials);
Compute own local probability (relative frequency of orange within spatial sampling radius);
For each other agent in network
    Compute the weighted solo disposition;
Add the above-computed numbers;
Subtract own threshold;
If the result is positive, Act; otherwise don't;
Apply own extinction rate to own affect;
Move;
Repeat.

What Is Time?

Finally, it is worth noting exactly what we mean by "time" in this model. In the agent model (as against the continuous-time differential equations) time is discrete. Here, time advances by one unit with every complete updating cycle of every agent and every patch.[131]

II.1. COMPUTATIONAL PARABLES

Science begins as parable, and ends as probability.[132] As this is a very young science, the runs that follow are closer to parables than to mature scientific claims of any sort. They arguably qualify as explanatory candidates in a broad sense, in that they *generate* certain qualitative behaviors (see J. M. Epstein, 2006). But they are computational parables—fables if you prefer. Of course, some fables endure.

Parable 1: The Slaughter of Innocents through Dispositional Contagion

For the base case run of the agent model, we will immobilize one of the agents. Call him Agent 0. *Netlogo* begins subscripting agents from 0, so this numbering assures consistency with the code provided. But "*Agent_Zero"* is the name of a class, while "Agent 0" is the name of an individual instance of that class. Lest any confusion arise, all the agents are of the general *Agent_Zero* type, just as all the diverse actors in a classical economic model would be of the *homo economicus* type (with different parameters, for instance). Agent 0 will be stationary in the southwestern quadrant of the landscape. The other two agents, Agent 1 and Agent 2, will execute random walks on the landscape but will begin in the hostile northeast quadrant. Agents can sample only the four sites to their immediate north, south, east, and west—their Von Neumann neighborhoods. All agents update their affects and local probability estimates, with dispositions updating over the fully connected network. For the Base Case, there is no affective extinction.

Crucially, Agent 0 is never attacked. As he is subjected to no direct conditioning trials, his immediate direct *affect is zero throughout.* Because he encounters no orange attack events, his estimate of probability (hostile given indigenous) is also *zero throughout.* Yet, he wipes out a "village"! How? As shown, the two rovers are encountering attacks (orange events). They are updating both their affect and their local estimate of the attack probabil-ity. When their total disposition values exceed their thresholds, they retaliate within their destructive radius (here equal to their sample

radius). Destroyed sites are dark red, as shown in the left frame of Figure 37. Their destruction and the escalation of their affects and probabilities continue. At all times, Agent 0 is weighting these (i.e., their solo dispositions) and adding them to his own destructive disposition.

Finally, these push him over his own threshold and he wipes out innocents, despite having a sample probability of zero, and no direct emotional grievance against the population, as depicted in the right frame.[133] *Also, Agent 0 is not imitating the destructive behavior of either other agent.*

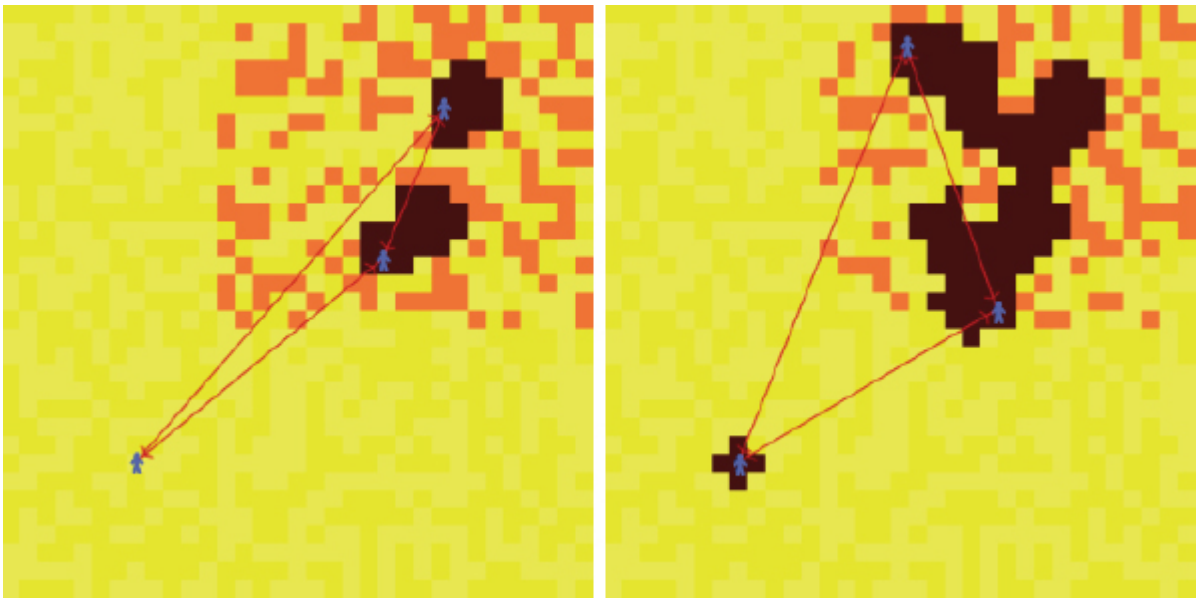


FIGURE 37. Activation by Dispositional Contagion [Movie 3]

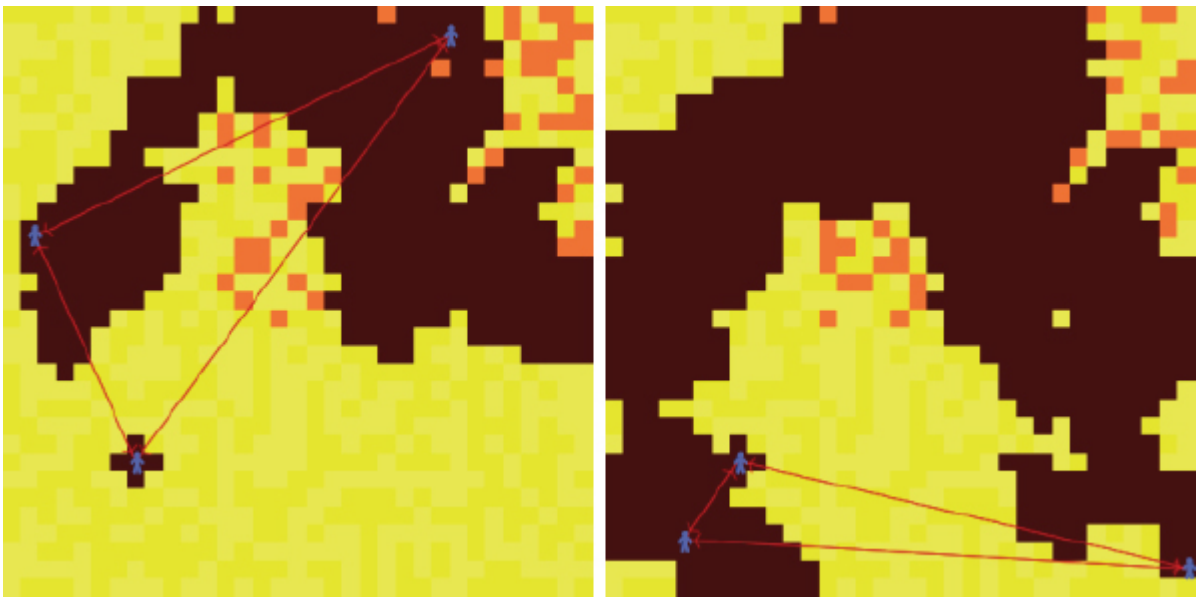


FIGURE 38. Slaughter of Innocents Continues [Movie 3]

In this particular case, he cannot even observe their destruction of the landscape because "vision"—the sample radius—is set to one patch in each direction.[134] With no extinction of affect, the mobile rovers go on to wreak vast destruction in regions that have never done them harm either, as shown in the two frames from **Movie 3** shown in Figure 38.

Specifically, having wiped out many of the insurgents (in the northwest quadrant) and having now drifted out of that quadrant, the rovers are, in fact, encountering mostly yellow (innocent) patches. Accordingly, their estimated probability of a hostile patch (the local relative frequency of orange) falls to zero. Yet, without any evaporation of affect—with no extinction of the conditioned affect—their dispositions remain high, and the killing continues with no direct empirical (observational) basis and no new conditioning trials. This is shown in the time series of net disposition and probability in Figure 39. Disposition and destruction remain high, despite falling probability for Agents 1 and 2. That is, their rampage continues as all empirical basis for it—their probability estimate—evaporates. This behavior is consistent with seminal laboratory psychology work of Zillmann et al. (1975).

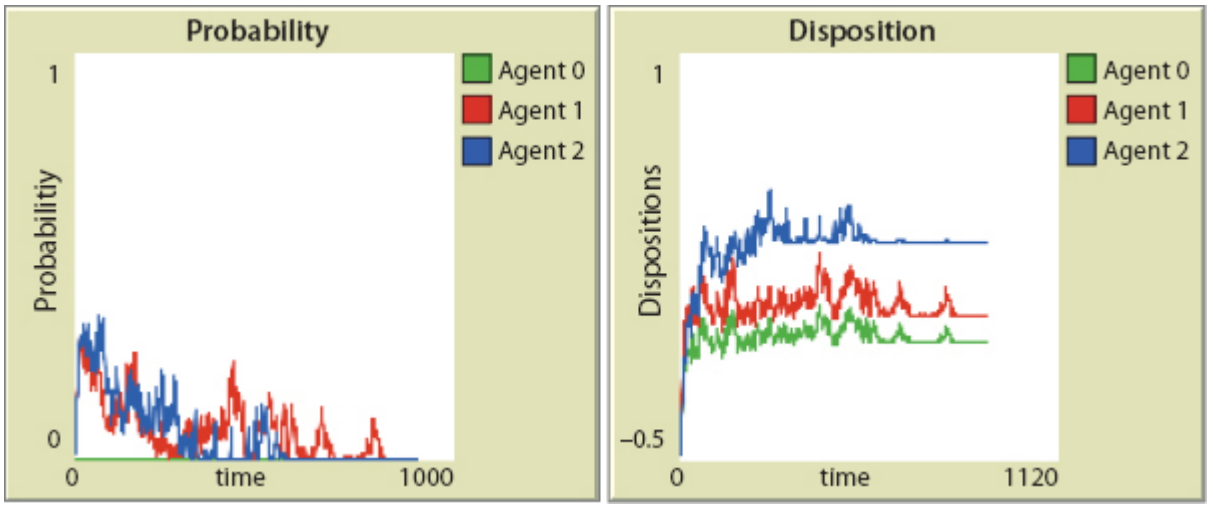


FIGURE 39. Rising Disposition Despite Falling Probability

Zillmann's Experiment

The article describing this experiment in detail is aptly entitled "Irrelevance of Mitigating Circumstances in Retaliatory Behavior at High Levels of Excitation." In sum, Zillmann et al. (1975, p. 282) showed that "Under conditions of moderate arousal, mitigating circumstances were found to reduce retaliation. In contrast, these circumstances failed to exert any appreciable effect on retaliation under conditions of extreme arousal." Specifically, "the cognitively mediated inhibition of retaliatory behavior is impaired at high levels of sympathetic arousal and anger." These conditions of affective arousal are certainly met, and agent behavior is entirely consistent with Zillmann's result.

Again, Agent 0's probability (the green curve of the left panel of Figure 39) is zero throughout. He would never have acted alone. And, he would never have acted even in a model of behavioral imitation, because he literally cannot "see" the others, and he need not, if they are in other forms of communication, such as auditory and social media.[135] The entire *NetLogo* Code for this parable is provided in Appendix III and again on the Princeton University Press *Agent_Zero* Website. This is a disturbing run,[136] but it is not yet our canonical central case, because Agent 0 does not act *first*.

Parable 2: Agent_Zero *Initiates: Leadership as Susceptibility to Dispositional Contagion*

Having developed all this apparatus, we can now generate that case, in which the first agent to act is not the one with the highest affect or the highest empirical estimate of indigenous hostility. Indeed, Agent 0 (again stationary) is subject to no direct aversive stimuli (orange explosions), so his individual (i.e., directly stimulated) affect and probability are both zero throughout, as shown in the corresponding plots of Figure 40. By contrast, the mobile rovers are subject to attacks, are accumulating affect, and are in-creasing their estimates of the probability that an indigenous patch is hostile (that a random patch will turn orange). All thresholds are equal at 0.5,[137] but neither rover's disposition exceeds this, so neither of them acts. Through their weights, however, their dispositions elevate Agent 0's to the highest of levels (see disposition plot), which exceeds the common threshold first. So, he is the first to act, as shown in the Figure 40 screen shot and **Movie 4**.

This is the situation I aimed to generate: *The agent at the front of the lynch mob has no particular grievance V or evidence P, and left to his own devices would never act. Notice that this is not "the banality of evil." Agent 0 is not "just following orders," because none are issued. And he is not imitating the behavior of others, because he is the first to behave!* The deeper point, as emphasized throughout, is that no agent is imitating *the behavior* of others, regardless of the order in which they activate. Thus, the model can generate important group dynamics without recourse to the copying of behavior (which is a binary variable that doesn't enter into the disposition calculus). Action (i.e., behavior) occurs if total disposition exceeds threshold, which occurs first for Agent 0.[138]

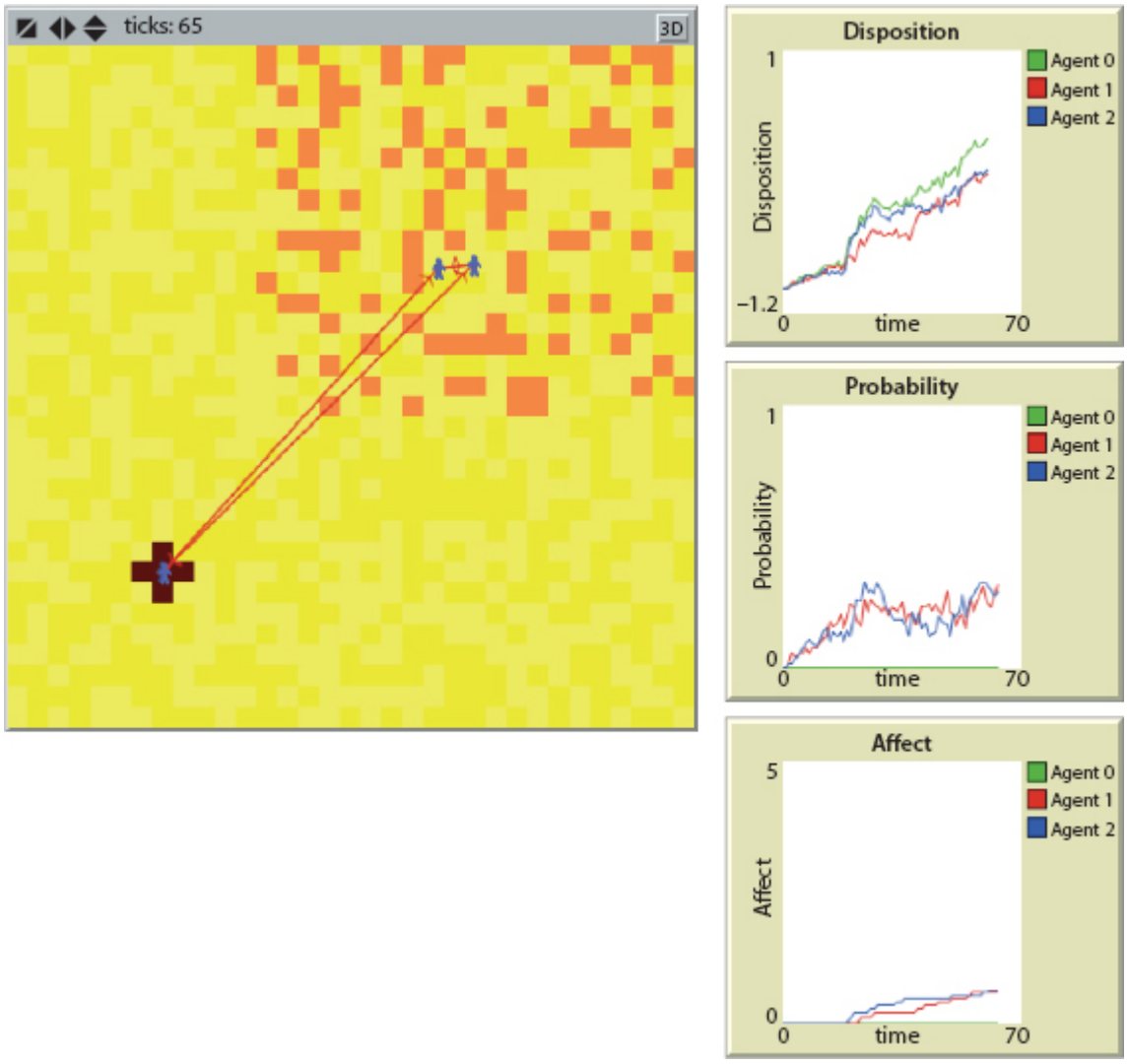


FIGURE 40. Agent with Zero Affect and Probability Acts First [Movie 4]

So, is Agent 0 a "leader," or is he simply the most susceptible to dispositional contagion? Which is the more compelling picture: the "great man" theory, or merely the susceptible one?[139]

This is Tolstoy's *swarm-life of man* in its most virulent form. Speaking of Bonaparte, Tolstoy writes:

> Though Napoleon at that time, in 1812, was more convinced than ever that it depended on him … he had never been so much in the grip of inevitable laws, which compelled him, while thinking he was acting on his own volition, to perform

> for the swarm-life—that is to say for history—whatever had to be performed." (*War and Peace*, p. 648)
> "To perform for the swarm-life. …" What a phrase! And this is the sense in which Tolstoy wrote, "A king is history's slave" (*War and Peace*, p. 647).

Complex Contagion Revisited

It is worth noting that Agent 0 does not act based on either one of the others alone. Here, he requires the swarm, the weighted sum, and multiple dispositional exposures, to go.

Run 3. Information Cuts Both Ways

In the runs thus far, the agents' spatial "vision" (landscape sampling radius) has been limited to a von Neumann (N, S, E, W) neighborhood of radius 4 patches. What is the effect of increasing the agents' vision? Let us begin with Agent 0 in his usual fixed position in the southwest quadrant, which we assume to be peaceful. Now let us give the other two agents fixed positions as well, but in the violent quadrant, as shown in Figure 41. What is the effect of increasing everyone's vision? More peace? More violence? Neither?

Let's consider Agent 0. His vision is his spatial sampling radius. As this extends into the red zone, he is seeing more violence. Hence, his estimate that a random patch is violent grows, as will his violent disposition. The agents in fixed positions in the red zone, however, have the reverse experience. Rather than seeing more violence as their vision grows, they see more yellow—peace! Accordingly, their probability estimate falls. Finally, when vision increases to the point where they can all see the entire landscape, their probability estimates converge, because their samples are now identical. Notice that the sample selection biases were very great at the low-vision outset, with Agent 0 underestimating—and the others overestimating—the global probability. Now they converge on the

correct global probability, as shown in Figure 42, where I simply increased the sampling radius midrun with the program's slider.

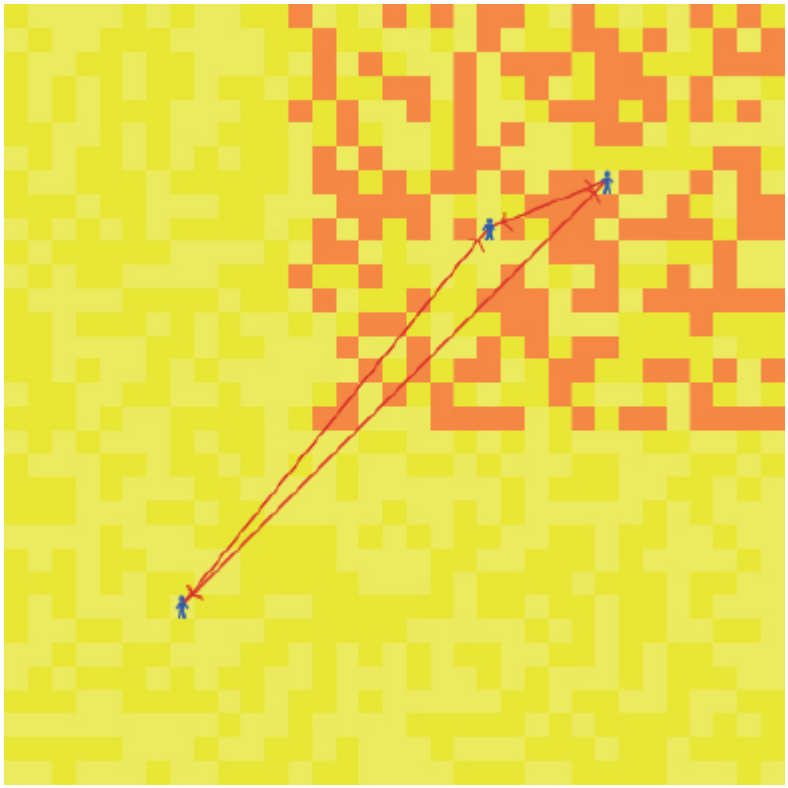


FIGURE 41. Fixed Agent Positions

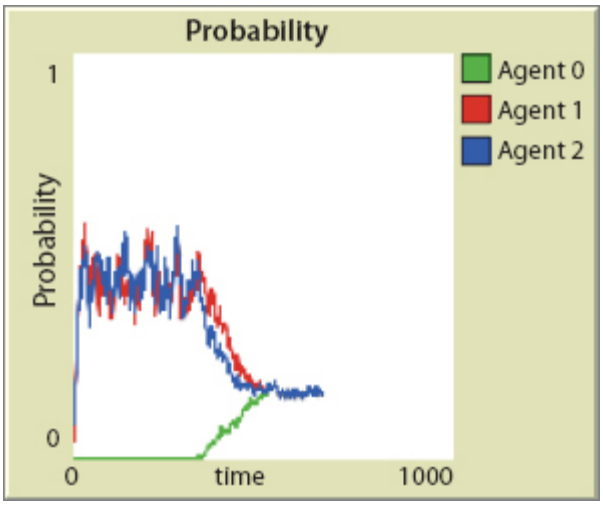


FIGURE 42. Probabilities Converge

This is an example of how sensitivities can be explored "on the fly" (midrun) in *NetLogo*, which readers are invited to do using the interactive Applets posted on the book's Princeton University Press Website.

Heterogeneous Vision

In this experiment (and in this variant of the model), vision was the same for all agents. Again, I am using the term *vision* figuratively, to denote the agent's search space. This could be entirely local, or global, or spread over a network, or confined to an organization. It could be literally ocular, auditory, olfactory, or text based, and so forth. The model permits high heterogeneity of vision. And it would be reasonable to explore this in future research, since people, in fact, differ widely in search spaces, and for a variety of reasons. Some types of information are expensive, for example. Some individuals are simply more inclined to acquire and process information than others. Cacioppo and Petty (1982) dub this "the need for cognition" and present experimental research that could be imported into the agent population. Instead of using a single global value for the sampling radius, one could use an empirically based distribution of vision as a crude analogue of this need for cognition. This could be a nice example of *computational social neuroscience*, where individual agents are based on experimental neuroscience, but then interact with one another in simulated populations.

Run 4. A Day in the Life of Agent_Zero*: How Affect and Probability Can Change on Different Time Scales*

Before we take up the topic of memory in Part III, which also involves time scales, I would like to show how the model can capture three easily recognized spatially explicit examples in which affect and probability change on different time scales. Obviously, many other examples will come readily to mind.

Case 1: Daily Grind

We've all (presumably) had the following experience: we begin the day in a perfectly good mood, go to work, have a lousy day, and come home in a rotten mood. Can we grow this prosaic example? Yes. In Figure 43, Agent 0 starts the day at 7:00 a.m. in his pleasant yellow neighborhood. His affect is zero and his appraisal of the probability of annoying demands is also zero. (Figure 43 also shows the *NetLogo* Interface with its user-adjustable sliders).

Then, as shown in Figure 44, he spends an aggravating and aversive nine-to-five day at the office (located in the upper right quadrant), where he is peppered by annoying demands (orange events). Within hours, his expectation of further annoyance and his aversion (affect) increase until, at quitting time, he is in an absolutely foul humor.

He arrives home, in Figure 45, where he is utterly free of harassment. He knows (since located in the yellow zone) the likelihood of further badgering to be zero, so his *P*-value drops to zero. But unless extinction is very fast, he is still in a foul humor (high *V*) when he arrives home. (Maybe his disposition to have a drink even exceeds his threshold!)

Exactly the same run can be interpreted variously.

Case 2: Emergency Responder

For example, one could interpret this as a story about first responders who enter a burning building—a terrifying experience during which the probability of being burned is high in proportion to the frequency of flames (orange squares). Once out of the building, the responder knows that the probability of burn injury is zero, but this fact does not extinguish the fear, which endures.

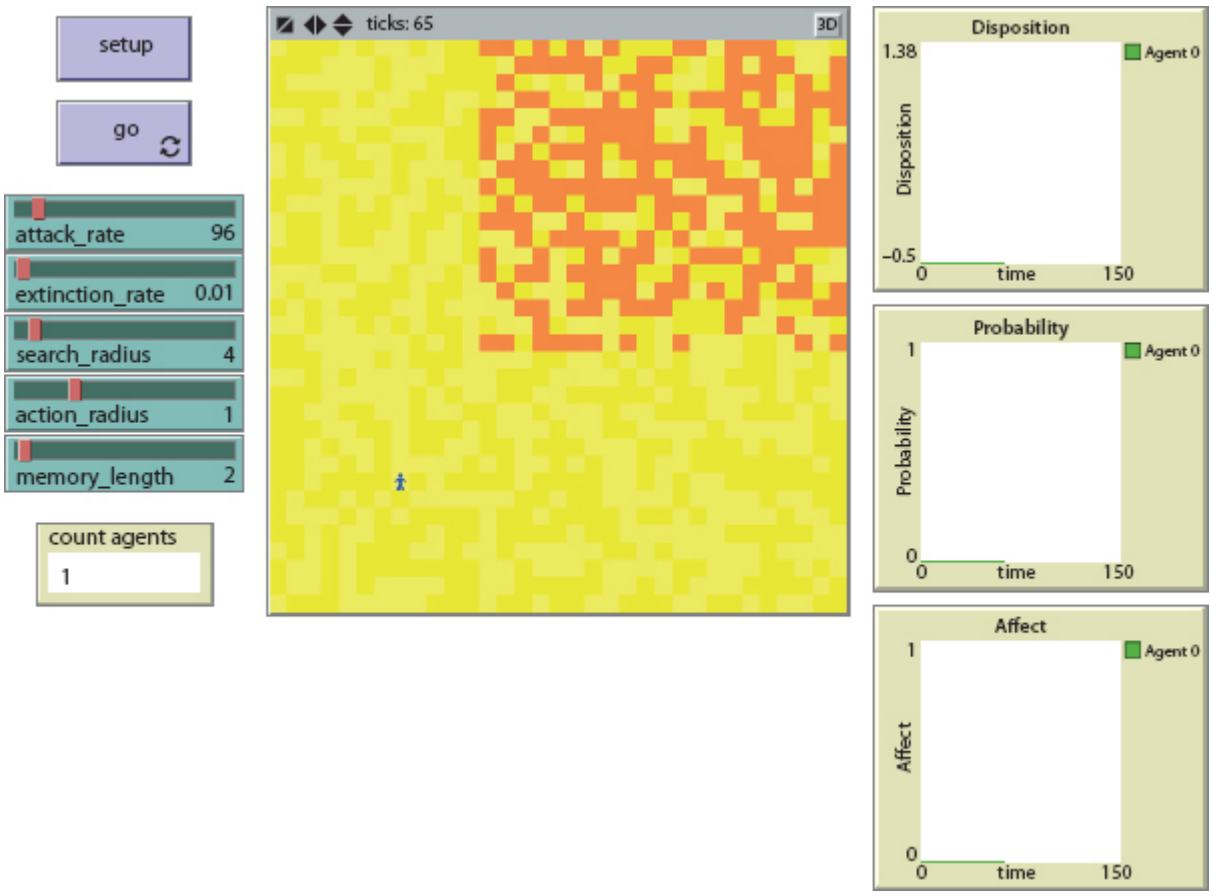


FIGURE 43. 7:00 A.M. Morning Coffee

Case 3: Combat

As the most extreme example, one thinks of entry into a war zone with orange bursts as enemy fire. During battle, fear and the probability of being hit are at their maximum. Upon withdrawal from the field, the probability drops to zero, but the posttraumatic stress can endure.

In all three cases, our agent begins the story in the placid yellow zone and in the affectively neutral state: that is, $v\,(0) = 0$, as shown in Figure 43.

After 150 periods, he ventures to the northeast quadrant—variously interpreted as rife with annoying office demands, flames, or enemy attacks. In this zone, both his affect and his estimate of

the probability of aversive events quickly rise to high levels, as shown in Figure 44.

At period 400—quittin' time—our protagonist departs this zone and heads back to home/base. He recognizes that the probability of further adverse events is zero (the sample probability curve falls to zero). But, this is insufficient to reverse his bad feelings, due to a low extinction rate, yielding the results of interest. As shown in Figure 45, Probability drops to zero, but the aversive affect persists.

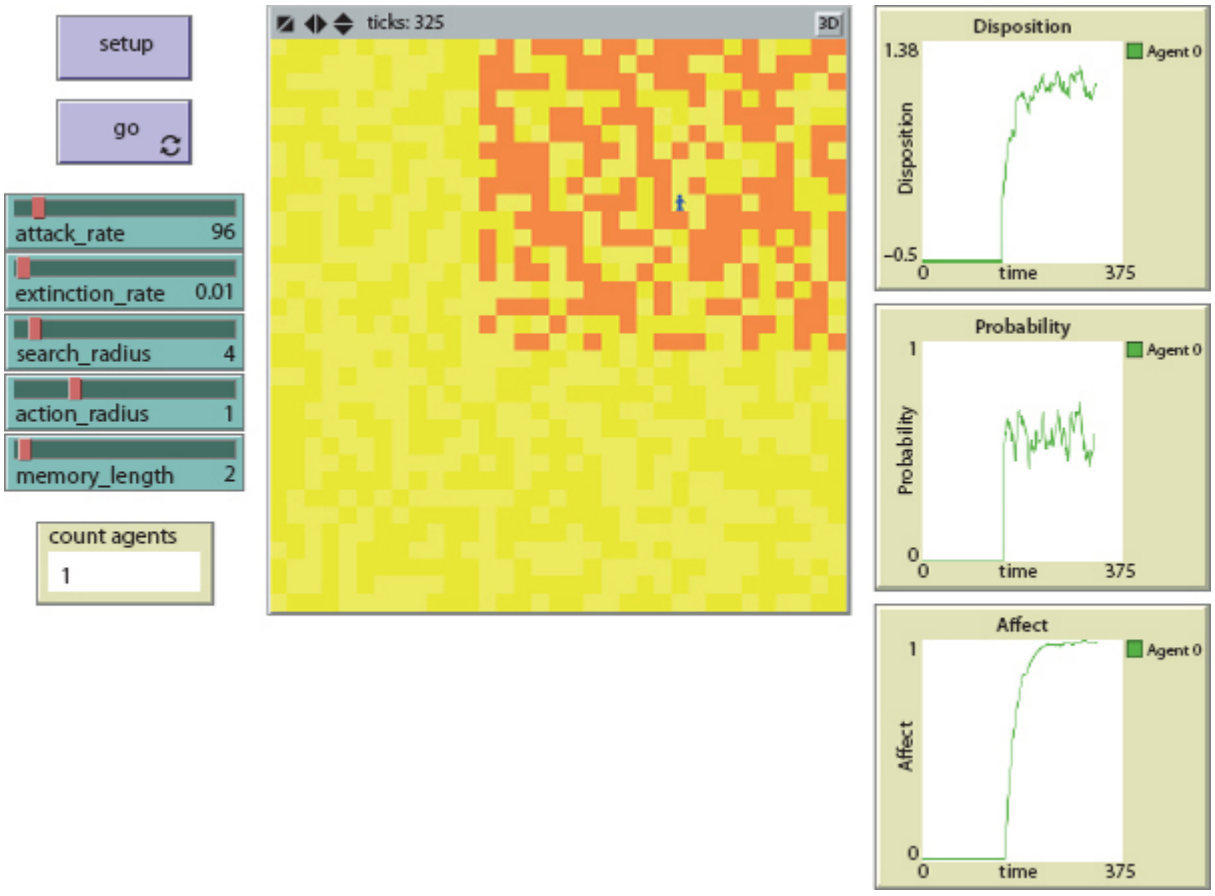


FIGURE 44. Nine to Five: Rising Demands During the Day

Of course, as discussed briefly before, by situational conditioning, the agent will come to associate the workplace itself with aggravation. So, *Agent_Zero* is already aggravated (or afraid, depending on the interpretation) when he walks in the door.

Case 4: A Happy Day

Clearly, the preceding fire and combat interpretations would involve fear and the amygdala, among other regions. But the same general associative learning *model*—though not the same brain regions—could apply to happy days, where one's disposition to break out in song is low at the start (Figure 43). So, suppose Agent 0 leaves home in the southwest for her college reunion somewhere in the northeast. On campus, happy singing breaks out all around (the orange outbursts of Figure 44). Agent 0 is rather shy (has a high sing-along threshold) so would never join in, except that her two best (high-weight) college friends (Agents 1 and 2) join in. Their dispositions to sing have weight, so she joins in. Finally, the party ends, and she heads home. And yet, even as the probability of direct musical encounters is zero, she remains aglow and sings the old college songs all the way home, as in Figure 45.[140]

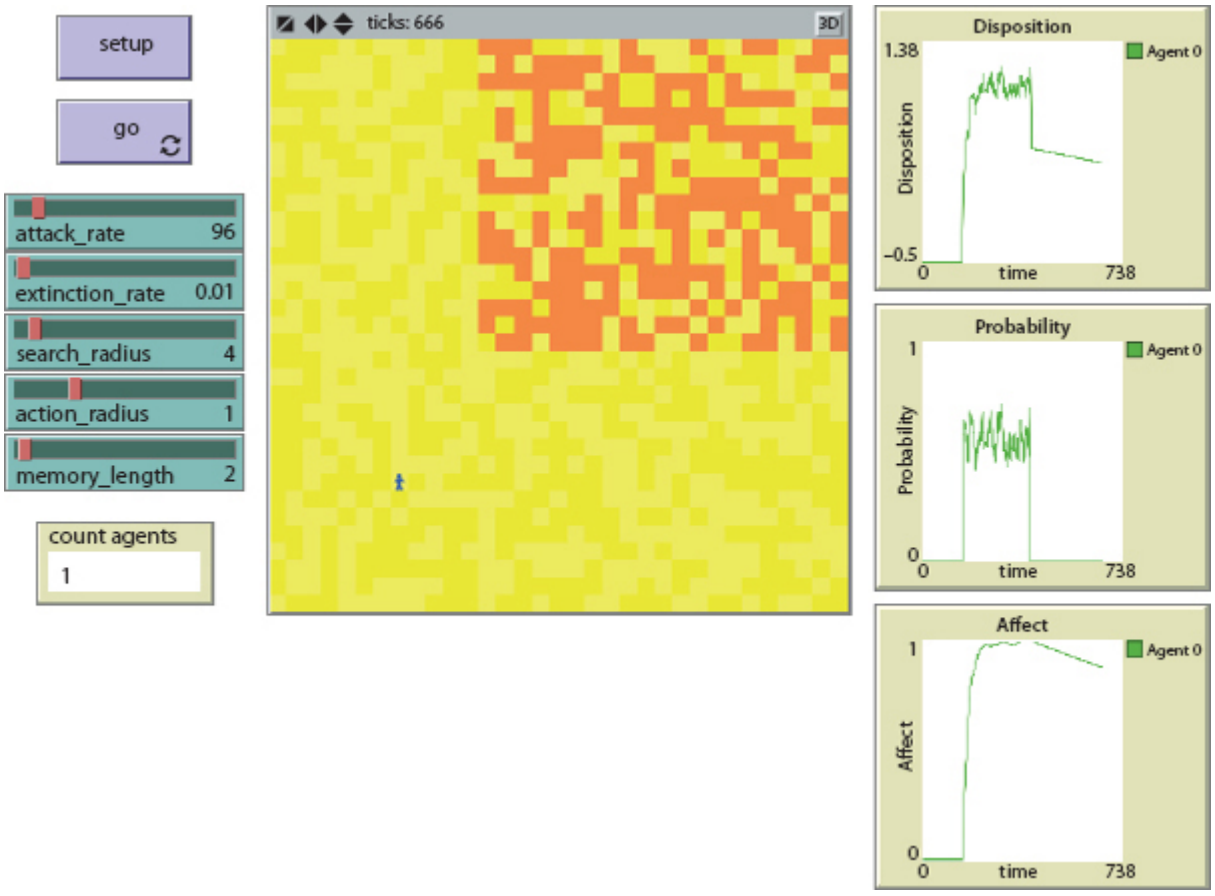


FIGURE 45. Direct Stimulus Stops, but Affect Continues

We have been exploring cases where *Agent_Zero's* sample probability rises and falls abruptly with his or her location in space, but her affect persists long after stimuli (trials) end. It would be interesting to devise cases in which affect evaporates before evidence does. We will do so below, when memory—along with much else—is introduced in Part III.

However, perhaps we have shown that the unadorned basic model—the basic *Agent_Zero*—does generate the intended central parables and much more. Specifically, most modeling focuses on extreme events. But the everyday life of people is equally worth modeling and, like the cases we've just developed, can be seen to "ring true" in the model, which is a good start.

Another game one can play with the base model is to explore the effect of one person's deficit on other individuals in her network. Earlier, in connection with posttraumatic stress, we used the mathematical version of the model to explore how one individual's experience affects others. Now, using the agent-based version, we will (I believe for the first time) *lesion* an agent and see the result, not only on her, but on others.

Run 5. Lesion Studies

My limited exposure to the literature suggests the utility of a purely logical dissection of the claims one might make about the amydgala and lesions.

Logic and Lesions

Ever since Klüver and Bucy's (1937) path-breaking work with primates, it had been conjectured that disabling the amygdala virtually eradicates fear. Recalling the rat's apparently hard-wired fear of even cat urine, "Large amygdala lesions dramatically increase the number of contacts a rat will make with a sedated cat.

In fact, some of these lesioned animals crawl all over the cat and even nibble its ear, a behavior never shown by the non-lesioned animals" (Davis and Whalen, 2001). More recent lesion studies—or contemporary studies using animals with genetically engineered deficits, such as "knock-out mice"[141]—establish that disabling or eliminating the amygdala indeed eliminates fear (along with much else). So, recognizing many nuances, just for logical precision, let's write this as[142]

$$\neg A \rightarrow \neg F.$$

If no amygdala, then no fear.[143] It follows logically that where there is fear, there is amygdala activation.[144] That is,

$$F \rightarrow A.$$

I have never understood why nature should ever respect our paltry rules of deduction,[145] but this is an observed regularity also (LeDoux 2003). Neither of these entails that excitation of the amygdala causes fear … that is, that

$$A \rightarrow F.$$

Lesion (or knockout) studies alone show *necessity, not sufficiency,* in other words. However, a history of experiments has shown that, "In humans, electrical stimulation of the amygdala elicits feelings of fear or anxiety as well as autonomic reactions indicative of fear. While other emotional reactions occasionally are produced, the major reaction is one of fear or apprehension." (See Davis and Whalen, 2001, and references cited there.)

While granting, then, that there are experimental grounds for an inference that $A \rightarrow F$, as a general proposition, this is equivalent to

$$\neg F \rightarrow \neg A.$$

which clearly *fails* since fear-inducing stimuli (e.g., snakes) are not the only inputs stimulating the amygdala (*A*). For example, erotic nude pictures and loud music can activate it[146] (Holland and Gallagher, 1990). So, we are not yet at the point where, given a subject's imagery (even accompanied by many other readings), we can infer their emotional state, or self-reported feeling, if there even is one!

In sum, while the amygdala does exhibit high functional specificity for fear (Kanwisher, 2010), the amygdala is *not* the only brain region involved in fear (Lindquist et al., 2012), *nor* is it the case that the only stimuli that activate the amygdala are fear inducing.

Lesioning *Agent_Zero*

Obviously, lesion studies on healthy humans are unethical. But lesion studies on software people are not (at least not yet). We can knock the amygdala out of *Agent_Zero,* as it were, and explore not only how it affects her behavior, but also how it affects the behavior of all others in her social group!

Here is *Agent_Zero*'s *NetLogo* amygdala, speaking very figuratively:[147]

```
[
if pcolor = orange + 1
  [set affect affect + (learning_rate * (affect ^ delta) * (lambda − affect))]
if pcolor != orange + 1
  [set affect affect + (learning_rate * (affect ^ delta) * extinction_rate *(0 − affect))]
]
```

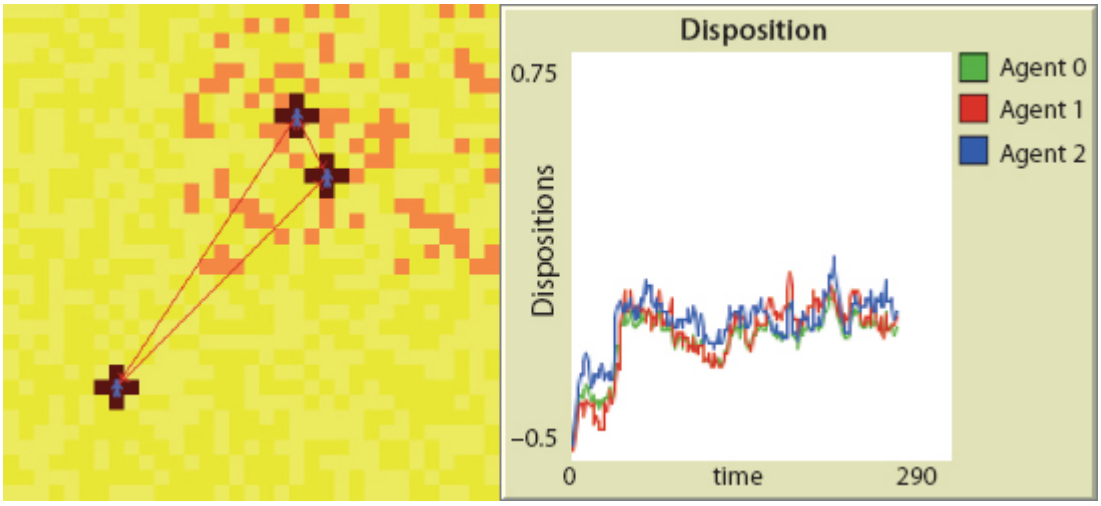


FIGURE 46. All Agents Normal

It is the agent's update-affect routine. In English, it says, "If the patch you're on bears the adverse event color, then set your new affect to your old affect plus the product of: (a) your learning rate, (b) your old affect to the delta power, and (c) the difference between lambda and your old affect. Otherwise (i.e., if the patch does *not* bear the adverse event color), do as before but replace (c) with the extinction rate times the negative of old affect, all of which follows the Rescorla-Wagner scheme.[148]

To knock out an agent's amygdala, we simply knock out this *NetLogo* Code block.[149] We are interested not only in how this lesioned agent behaves, but also in how her neurocognitive deficit affects the whole network. Depending on the agent's weight, the effect can be dramatic. One agent's deficit can have far-reaching ramifications. In Figure 46, we see a Run with all agents functioning normally.[150]

If we now knock out the amygdala of Agent 2 (the upper-right rover), it eliminates her fear (and her violence) and her transmission of fear to both Agents 0 and 1. Agent 1 (upper left) still acquires fear directly from events and transmits this to Agent 0. But lesioned Agent 2 is no longer contributing to Agent 0's fear (either directly or through Agent 1). As a result, the total fear acquired by Agent 0 is now beneath her action threshold, and she never engages in violence. This is shown in Figure 47.

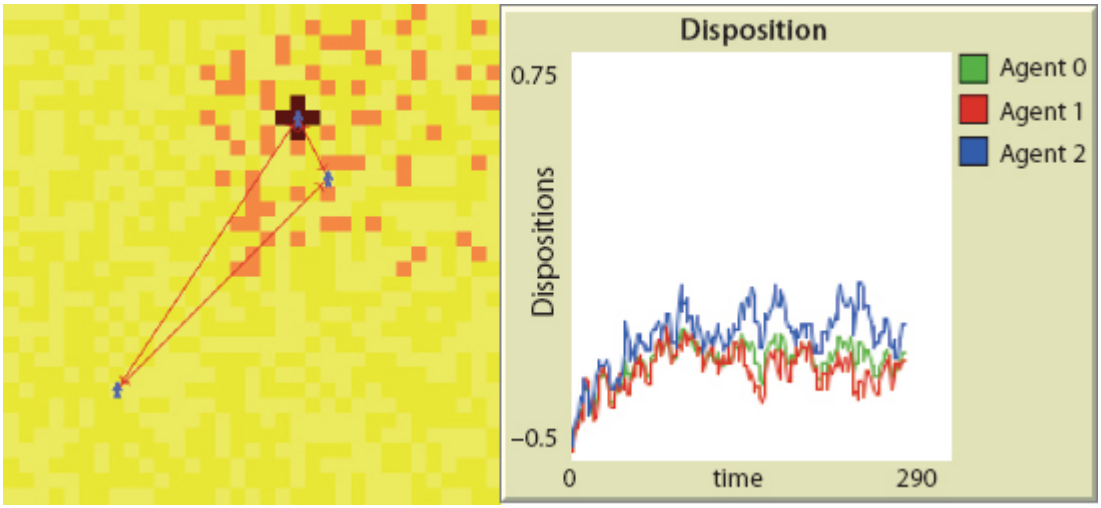


FIGURE 47. Agent 2 (Upper-right) Lesioned

Emotional contagion dynamics are affected if one agent's *direct* fear acquisition is disabled. But, as we now discuss, *observational* fear acquisition may also be impaired by amygdala damage.

Patient S. M.

The famous subject S. M. suffered from Urbach-Wiethe disease. In their classic paper, Adolphs et al. (1994, p. 670) write that her condition "caused an nearly complete bilateral destruction of the amygdala, while sparing hippocampus and all neocortical structures, as revealed by detailed neuroanatomical analyses of her computed tomography (CT) and magnetic resonance imaging (MRI) scans." The result was that S. M. was unable to recognize fear—and emotion generally—in the faces of others. To represent S. M. in the *Agent_Zero* framework, we would add to the disability just discussed the further inability to acquire fear *observationally* (as discussed earlier). Mathematically, this would be arranged by zeroing out the weight S. M. assigns to the affect of others.[151] In the social setting, this will further damp network transmission because she will not pick up emotion; and so she will not pass it on either. So, the damping social effect would be even more pronounced. That, at any rate, would be the hypothesis.

Generative Minimalism

The runs and discussions presented thus far involve no extensions to the basic *Agent_Zero* model. While the agent specification is quite minimal, considerable generative capacity has been demonstrated. While much more exploration of the basic model is warranted (and is easy given the Applets and Source Code posted on the book's Princeton University Press Website), we turn now to 14 significant extensions.

---

[118]A nice "Guide to Newcomers" is available in Axelrod and Tesfatsion, Appendix A of Tesfatsion and Judd, eds., *Handbook of Computational Economics: Agent-Based Computational Economics, Volume 2*. Among the closest things to a textbook on agent modeling is Railsback and Grimm (Princeton 2011). The best hands-on way to get started is to do the three excellent agent-based modeling tutorials that download with *NetLogo* (http://ccl.northwestern.edu/netlogo/).

[119]By canonical, I mean simply the base model for this development.

[120]A torus topology is readily available in *NetLogo* but would have been visually confusing for most of the runs explored here.

[121]*NetLogo* offers a wide variety of distributions from which to draw random numbers. Here, *U*(0, 1), the uniform distribution on the unit interval, is used.

[122]*Update-affect* is the relevant *NetLogo* code block. See Appendix III. My code extends Rescorla-Wagner in allowing extinction rates different than the classical model, which, of course, is an available setting.

[123]Properly speaking, this extinction-rate slider is a multiplier. If it is 0, there is no extinction. If it is 1.0, we obtain classical Rescorla-Wagner extinction curves. Typically, we use a value in the interval (0,1). So, this is a second extension of the original model (beyond *S*-curve learning), permitting yet another type of flexibility.

[124]Hence the adjective "spatial."

[125]Later, we will interpret the set as a space of financial assets, a family of vaccines, or opportunities for unhealthy eating, over which a *local relative frequency* is being computed and updated.

[126]This is the number of orange patches over total patches within the spatial sampling radius.

[127]An anxiety-provoking context without question (Behrens et al., 2007).

[128]Lest there be any replicative or other confusion, the agent source code and *NetLogo* graphical output use the name *disposition* for *net disposition*. This should occasion no confusion. The *NetLogo* code block (see Appendix III, p. 218) governing this calculation is:

```
to update-disposition
  ask turtle 0 [
    set disposition affect + probability + [weight] of red-link 1 0 * ([affect] of
    turtle 1 + [probability]
    of turtle 1) + [weight] of red-link 2 0 * ([affect] of turtle 2 + [probability]
    of turtle 2) − threshold]
  ask turtle 1 [
    set disposition affect + probability + [weight] of red-link 0 1 * ([affect] of
    turtle 0 + [probability]
    of turtle 0) + [weight] of red-link 2 1 * ([affect] of turtle 2 + [probability]
    of turtle 2) − threshold]
  ask turtle 2 [
    set disposition affect + probability + [weight] of red-link 0 2 * ([affect] of
    turtle 0 + [probability]
    of turtle 0) + [weight] of red-link 1 2 * ([affect] of turtle 1 + [probability]
    of turtle 1) − threshold]
end
```

Terms could be collected in a variety of ways, all equivalent computationally but different conceptually. This form seems expeditious for expository purposes. *NetLogo*'s name for a generic agent is "turtle." I choose to imagine that this is in honor of a famous exchange between Bertrand Russell and an audience member who told Russell that the earth was supported on the back of a great turtle. Russell asked, 'And what, pray tell, is supporting *that* turtle?' The answer was immediate. "Oh, another turtle … it's turtles all the way down."

[129]As noted, agents can be in the same dispositional network even if they are not within one another's spatial sample radius. In such cases, communication (and dispositional contagion) could be by voice, by text message, by iPhone, by field radio, or other social media. Below we offer an ex-tension allowing one to change weights step-functionally when others enter (or exit) one's sampling radius. We do not exploit that in the main exposition.

[130]Later, I endogenize this radius as a function of affect.

[131]Technically, time is measured in *ticks*, a reserved word in *NetLogo*. In this model I advance *ticks* by 1 with each cycle through the *NetLogo* "go" routine, which corresponds to *main* in C or C++. In this case, the full "go" code is a follows:

```
to go
  if ticks > = maximum-stopping-time [stop]
  move-turtles
  activate-patches
```

```
    update-event_count
    update-affect
    update-probability
    update-disposition
    take-action
    deactivate-patches
    do-plots1
    do-plots2
    do-plots3
    tick
end
```

[132]For example, the primordial fire god parables have been displaced by the uncertainty principles of quantum mechanics.

[133]One might well say that Agent_Zero betrays himself in that his solo disposition is below his threshold, whereas his total (in the group) disposition exceeds it. In the nomenclature of the Introduction, $D^{\text{tot}} > \tau > D^{\text{solo}}$.

[134]As noted earlier, emotion and disposition can be communicated by numerous routes beyond immediate vision.

[135]It is important not to muddy the distinction between the spatial sampling radius and the distance over which dispositional contagion may occur. The two are completely independent in the model. Weights do not increase with spatial proximity or shrink with distance. An extension allowing this is offered under Future Research.

[136]Even 10% extinction alters this considerably. A little forgiveness, or counter-learning, can go a long way.

[137]Notice that this makes Agent 0's solo net disposition negative in fact, since it is v (here 0) plus P (here 0) minus τ (here 0.5). The others' dispositions begin negative but rise quickly with aversive stimulus. Notice also that we do not arrange the activation order by giving agents different thresholds.

[138]There is an asymmetry in the model as developed to this point. In defining the binary action to be X (equal to 1) rather than not-X, one induces a reference direction. In the cases just described it is positive. One acts when one exceeds the threshold, not when one drops below it. As we see, the solo dispositions of other agents can indeed move one's net disposition in a positive direction (e.g., from negative to positive). But, since solo dispositions are nonnegative, they cannot move net disposition in a negative direction, that is in a direction contrary to the reference direction. To permit this, various mechanisms present themselves. One is threshold imputation, which I intro-duce in Part III, to replicate the Darley-Latane experiment. Another would be to introduce negative weights. A third variation, for which I thank Julia Chelen, would be to have agents assign weight to the *average* of others' Vs and/or Ps. I also thank Jon Parker for discussions of this issue.

[139]See Tolstoy (1869; 1998 ed.)

[140]Less frivolously, the model captures cases where the individual has a good impulse but simply needs the support of others to act on it. I thank Julia Chelen for this observation.

[141]See Mayford et al. (1997).

[142]We employ the logic symbols ¬ (the negation symbol meaning *not*) and → (meaning *implies*).

[143]However, see Cunningham and Brosch (2012).

[144]If p implies q, then not q implies not p. Each implication is the so-called contrapositive of the other, with $\neg A$ playing the role of p.

[145]For example, nature respects Newton's second law, that $F = ma$. But, evidently, it also respects every proposition deducible from this law. But deduction is an entirely human invention. Why should nature select the deducible claims as the ones to which it will physically conform? I find that mysterious.

[146]This indicates that the amygdala is, in fact, not specialized to fear. It is implicated in many kinds of arousal.

[147]Again, I am not modeling brain regions.

[148]My code actually generalizes Rescorla-Wagner extinction slightly by introducing the variable named extinction_rate, which is a user-adjustable slider in the *NetLogo* Interface. If extinction_rate $= 1$, then the scheme is exactly the classical Rescorla-Wagner model. If $0 <$ extinction_rate < 1, slower extinction trajectories can be explored. Typically, extinction_rate $\in [0, 1]$.

[149]This is the sense in which *Agent_Zero*, as a software object, is "modular," which is to make no claim whatever regarding the modularity (however defined) of the human brain.

[150]The slider settings in this case are: Attack Rate 25, Extinction Rate 0, Sampling Radius 4, Action Radius 1, Memory 1, with Seed 2, which are also given in the Table of Appendix IV.

[151]Technically, weight has thus far been assigned only to the sum of V and P.