INTRODUCTION

MOTIVATION

In his *Treatise of Human Nature*, David Hume (1739; 2000 ed.) famously wrote, "Reason is . . . the slave of the passions."[2] In using the term *slave*, however, Hume's point is not that the passions always prevail over reason in a tug of war,[3] but that the two are, in some sense, incommensurable. Distinguishing the passions from factual (true/false) claims, he writes, "Tis impossible [that] they [the passions] can be pronounced either true or false, and be either contrary or conformable to reason" (p. 458). I take this "nonconformability" to mean that, as a modeling proposition, passion and reason—emotion and ratiocination, if you prefer—belong on different axes. They might be seen as basis elements of a dispositional vector space. If we adopt Hume's geometry, then beyond passion and reason there is a third basis element, or axis, inspired by Spinoza, who wrote that "Man is a social animal."[4] In this view, then, our actions may be motivated by the passions, but are influenced by reason—and by society.

Needless to say, Hume was not the last, nor Spinoza the first, to philosophize on the roles of emotion, deliberation, and social influence in human behavior. Indeed, innumerable contrasting perspectives on the balance among them have been articulated over the centuries. I do not purport to end this dialogue. I do claim, however, to inject into that discussion a new approach.

Specifically, my central objective here is to develop a simple explicit model of individual *behavior in groups* that includes some representation of "the passions," of (imperfect) reason, and of social influence. In other words, I will offer an exploratory synthesis of three (partially understood and obviously intertwined) processes:

[2] The entire passage is, "Reason is, and ought only to be, the slave of the passions." His use of the term *ought* is, of course, not an admonition, but carries the meaning that it is in the nature of things that reason be the slave of the passions, a point that struck me as incidental to the main topic at hand. Hence the ellipses.

[3] On the contrary, he writes, "We speak not strictly and philosophically when we talk of the combat of passion and reason." That is, we speak incorrectly when we do (Vol. 2, p. ix).

[4] Baruch Spinoza (1632–1677). *Ethics*, Pt. IV, prop. 35: note. The addage is often attributed to Aristotle. But, technically, he wrote that, "Man is by nature a political animal." *Politics 1.* One can argue that "social" is a defensible translation of the original Greek, *politikos,* in that the latter means "of the polis" and for Aristotle, the polis *was* society. But with Spinoza there is no doubt.

- The emotional
- The cognitive
- The social

To this end, I introduce a new theoretical entity, *Agent_Zero*, endowed with emotional/affective, cognitive/deliberative, and social modules whose—often nonconscious—interactions determine his or her observed behavior.

Generate Social Dynamics

The *aim* is to *generate* recognizable dynamics of social importance. Specifically, we humans often do things in groups that we would not do alone (Le Bon, 1895; Canetti, 1984; Mackay, 1841, Browning, 1998). We do things for which we have no basis in evidence. Indeed, we do them *knowing* we have no evidence; and sometimes, despite this, we are even the *first* in the group to do them!

Dispositional Contagion, Not Behavioral Imitation

Notice that in this latter case, the *imitation of behavior* cannot be the mechanism. Clearly, if I am the first actor, I cannot be imitating anyone's behavior, because no one before me has acted! But, more to the point, in this model my successors are not imitating my behavior either. No agent is imitating any other's *behavior* as defined in this model. Here, behavior is binary—agents act or do not—and, by design, this Boolean variable does not appear as an independent (right-hand side) variable in any agent's algorithm (discussed more fully later). Of course, agents may end up *doing* identical things, but the mechanism introduced here is *dispositional contagion*, not "monkey see, monkey do" imitation of observable binary action, which, as I say, is barred mathematically. Without positing "behavioral imitation" in the usual sense,[5] the *Agent_Zero* model will be shown to generate a wide range of social dynamics.

A Core Target

As a core target for the model, I want to "grow" the person who *feels* no aversion to black people, who has never had any direct *evidence* or experience of black wrongdoing (his empirical estimate of this probability is and always has been literally zero), and who yet initiates the lynching. Though,

[5] Behavioral imitation can be a productive modeling assumption. I have used it myself. See, for example, Chapters 7 and 10 of J. M. Epstein (2006). It is not used here.

as first mover, he does not copy the *behavior* of others, a type of contagion is at play, but it is dispositional. He is, of course, an ideal type, but an important one.[6]

Mathematical social science has largely attempted to characterize the rational; I am interested in generating behavior that is far from rational. And I believe there is more at play than mere bounds on information or computing resources. While cases of collective cruelty abound, there are also examples of collective resistance to it. A general theory should generate both. I will attempt to do so here.

Specifically, *Agent_Zero*'s observable behavior will result from the interplay of (a) emotional/affective, (b) cognitive/deliberative, and (c) network/contagion components. As defined here, the agent's total *disposition* to act will be an explicit mathematical function of these. This disposition will be a real number. If it exceeds some threshold, *Agent_Zero* acts. Otherwise she does not.

Each model component (affective, deliberative), of course, is an entire discipline in its own right. And the crude models set forth will doubtless merit criticism from domain experts. But I am less interested in the accuracy of the components than in the generative capacity of the synthesis. The components can—indeed must—be refined. But, as far as I know, this particular synthesis has not been attempted. So, to me, *it is less important to get the components finished than to get the synthesis started*, as simply and understandably as possible.

While it impinges on a number of enduring philosophical issues, this synthesis aims to fill an important gap in contemporary social science. In statistical social science, relationships among aggregate macroscopic variables are assessed econometrically. The individual, as such, is not represented. In mathematical economics and game theory, the individual *is* represented, but the representation is not grounded in cognitive neuroscience.[7] But laboratory neuroscience—transformed by advances in imaging—is focused on single individuals (e.g., fMRI subjects in controlled laboratory

[6] I do not study lynching empirically. For "black" and "lynching," one should feel free to substitute "Jew" and "pogrom," "Congolese civilian" and "slaughter," or any number of other bleak equivalents. Indeed, the endless supply of bleak equivalents is among the main things I wish to understand. The most recent confirmed lynching in the United States occurred in 1981, when ". . . an African-American teenager named Michael Donald was murdered by two members of the Ku Klux Klan, who slit his throat and hung his body from a tree in Mobile, Alabama. "*The 'Last Lynching': How Far Have We Come?*" Ted Koppel. Transcript available at http://www.npr.org/templates/story/story.php?storyId=95672737, National Public Radio (2008).

[7] The exciting new field of neuroeconomics (Glimcher et al., 2008) is commendably attentive to brain science. But it has been focused largely on individual economic decisions, not broader social behaviors, like violence, discrimination, or contagious fear, which are explored here.

experiments), not on *networks of* individuals.[8] So, what happens in *networks of cognitively plausible individuals? Indeed, by what mechanisms do cognitively plausible individuals generate networks?* This book explores these questions mathematically, and computationally—in the latter case, by the method of agent-based modeling.

Generative Minimalism

This effort is exploratory and unapologetically theoretical. It is an exercise in generative minimalism. I do not "fit" the basic model or its extensions to numerical data, though this is an obvious candidate for future research, nor do I claim to predict events.[9] Rather, I hope to demonstrate that the *Agent_Zero* model is *sufficient to generate* core qualitative social dynamics across a wide range of domains. These demonstrations establish the model's generative explanatory candidacy, in the language of J. M. Epstein (2006). For a full discussion of the generative explanatory epistemology, generative minimalism, and examples of empirically calibrated agent models, see J. M. Epstein (2006).[10]

Not Modeling Brain Regions

There is another, extremely important, qualification. We will, of course, have occasion to discuss a number of brain regions: the amygdala, hippocampus, and prefrontal cortex, for example. But I am emphatically *not modeling brain regions*. To clinicians, brain regions are of interest because functional localization facilitates therapy (e.g., neurosurgery). But to social theorists brain tissue is interesting exactly and only in so far as it licenses model design or model interpretation.

What is plausible for humans and what is not? The neuroscience will both illuminate and explain those limits, limits that belong in the model and that have been largely ignored by social science. The brain discussions that follow are meant only to support that modeling aim. If my account of the tissue science is wanting (as it surely is) or if the state of tissue science advances (as it surely will), I do *not* believe that per se endangers the mathematical model. One wants to develop a mathematical/computational framework that can exploit and accommodate the evolving

[8] Emerging efforts in hyperscanning take fMRIs of two interacting people simultaneously (Lee, Dai, and Jones, 2012; Montague et al., 2002).

[9] For a large variety of possible modeling goals other than prediction, see J. M. Epstein (2008).

[10] On generative minimalism, see also S. D. Epstein (2000), S. D. Epstein and T. D. Seely (2002, pp. 1–10) and S. D. Epstein and N. Hornstein (1999, pp. ix–xviii).

tissue science but is not hostage to today's snapshot of it, for surely there are few fields developing faster. At the same time, one's model should also be sufficiently binding as to *admit* falsification. I hope to strike this difficult balance.[11]

Thus, I will display the formal models I have in mind and try to undergird them with what little cognitive neuroscience science I know, hoping that the attempt will elicit the empathetic interest and collaboration of domain experts. Now to the model's formal constituents.

THE MODEL COMPONENTS

For the *affective/emotional* component, we will use the essentially Pavlovian theory of associative learning; specifically, a generalization of the Rescorla-Wagner (1972) model of conditioning. For the *cognitive*—ratiocinative and deliberate—component, agents will form a probability based on local sampling (which can introduce bias). For the *social* component, I will use a network transmission model, but with a crucial wrinkle. Most work on contagion in networks (including some of my own) posits the conscious observation and imitation of *behavior*. Very importantly, however, in the present model it is *not* observable (binary) behavior that is contagious in these networks, but dispositions. Finally, as a starting point, the function combining these affective, deliberative, and social components—the agent's executive function, if you like—will be the simplest imaginable: addition.[12] While numerous refinements are possible, the first issue is the generative capacity of this simplest of all possible syntheses.[13] As Einstein admonished, in modeling, *"everything must be made as simple as possible, but not one bit simpler"* (A Letter from the Publisher, 1962).

[11] Relatedly, it is clear that freezing in primal fear of the wriggling snake is different from cool taxonomic appraisal of the danger it might represent—different in the brain systems engaged and the physiological responses evoked. In a certain literature, these systems are sometimes nicknamed the "hot" and "cold" spheres of cognition. This terminology is used, for example, by the evolutionary psychologists Tooby and Cosmides (2008). Others use the terms *automatic* and *controlled* (Schneider, 2003). Stanovich and West (2000) introduced *System 1* and *System 2*. This terminology is adroitly employed by Kahneman (2011), who emphasizes that System 2 is "effortful," unlike System 1. I do not adopt any of these particular terminologies, which is not to criticize their use by the authors mentioned. I simply have a different objective here.

[12] The neural network literature offers a wealth of transfer functions *of* this sum. Classical backpropagation uses a sigmoid of this sum, for example. We will return to this, but for this initial development, we will simply *superpose* the components and compare to a threshold.

[13] Agents are endowed with memory only in the spatial agent-based model of Part II.

MODEL OVERVIEW

Though full details will be presented shortly, the basic mathematical scaffolding of the model is as follows. First, we imagine a binary action, $A \in \{0, 1\}$. Raid the icebox or don't. Participate in (or even initiate) the lynching or don't. Buy the BMW or don't. Accept a vaccination or don't. Support internment of Japanese Americans or don't. Flee the snake or don't. Wipe out the village or don't. *This is what we mean by "behavior." It is zero or one, a binary matter.* Exactly when do agents act in the model? This requires some steps.

Dispositions and the Action Rule

Agents will be endowed with specific affective and deliberative functions below. For the *i*th agent, these will be denoted $V_i(t)$ and $P_i(t)$. They will be dynamic and will change with experience. But at any time they return nonnegative real numbers between zero and one, inclusive, as values. We define the i^{th} agent's *solo disposition* as simply the sum of these affective (V) and deliberative (P) components. So for Agent *i*.

$$D_i^{\text{solo}}(t) = V_i(t) + P_i(t). \qquad [1]$$

Agents also carry (unconsciously,[14] I assume) a set of numbers—*weights*—from the interval [0, 1] registering the influence of other agents' solo dispositions.[15] So, Agent 1 might be strongly influenced by Agent 2 but oblivious to Agent 3. If we let ω_{ji} denote the weight *of Agent j on Agent i*, then we would give ω_{21} a high value of perhaps 0.9 and ω_{31} a value of 0.

We next define the *total disposition* of Agent *i*, D_i^{tot}, as her own solo disposition plus the sum of the weighted solo dispositions of all other agents. Hence, at any time *t*,

$$D_i^{\text{tot}}(t) = D_i^{\text{solo}}(t) + \sum_{j \neq i} \omega_{ji} D_j^{\text{solo}}(t). \qquad [2]$$

This could, of course, be written as a single global summation with the convention that self-weights (ω_{ii}) are unity.[16]

[14] "Unconsciously" does not mean "while unconscious, as after a concussion," but simply "without awareness." I sometimes use the term nonconscious to avoid confusion.

[15] A single weight is applied to the full solo dispositions of others, not to their separate *V*s and *P*s. Obviously, the distributive law applies, but conceptually, I do not argue that agents have access to others' *V*s and *P*s, and of course the same solo disposition could be formed by an infinitude of *V*, *P* combinations.

[16] This will be our assumption. But agents might discount their own dispositions, setting $\omega_{ii} < 1$. They might also assign negative weights to the solo dispositions of other agents, a possible extension we will not explore here.

Now, each agent carries an *action threshold*: $\tau \geq 0$.[17] If one's total disposition D^{tot} exceeds the threshold, τ, then the action is taken: $A = 1$. Otherwise, it is not, and $A = 0$. In other words, agents act if and only if (denoted *iff*) their total disposition exceeds their threshold:

$$\text{Action Rule: Act iff } D_i^{\text{tot}}(t) > \tau_i. \tag{3}$$

If we further define the ith agent's *net disposition* $D_i^{\text{net}}(t)$ as $D_i^{\text{tot}}(t) - \tau_i$, the total net of threshold, then the agent's action rule can be stated succinctly as follows:

$$\text{Action Rule: Act iff } D_i^{\text{net}}(t) > 0. \tag{4}$$

In terms of our binary action variable, A, this is equivalently

$$\text{Action Rule: } A_i = \left\{ \begin{matrix} 1 \text{ if } D_i^{\text{net}} > 0 \\ 0 \text{ otherwise} \end{matrix} \right\}. \tag{5}$$

In vector terms, if we employ a dot product[18] and the Heaviside unit step function,[19] H, Agent i's binary action rule (equation [5]) can be written compactly as an equality:

$$\text{Action Rule: } A_i = H[\boldsymbol{\omega}_{ji} \cdot (\mathbf{V}_j + \mathbf{P}_j) - \tau_i]. \tag{6}$$

Here (suppressing time for clarity) $\mathbf{V}_j$ and $\mathbf{P}_j$, respectively, denote the vectors of all V and P values and $\boldsymbol{\omega}_{ji}$ denotes the vector of all weights *on* Agent i.[20]

Various formulations of total disposition, described shortly, will prove useful. But one of them is particularly revealing of the core distinction between behavioral imitation and dispositional contagion. We will refer to it as the *skeletal equation*, because it doesn't specify particular V and P functions.

[17] While the model allows heterogeneity, we will for the most part assume a single common threshold here.

[18] One type of vector is simply an n-tuple. Given two vectors, $\mathbf{A} = (a, b, c)$ and $\mathbf{B} = (x, y, z)$, their dot product $\mathbf{A} \cdot \mathbf{B}$ is $ax + by + cd$, the sum of position-wise products. On vector operations in general, see for example, Mardsen and Tromba, *Vector Calculus* (2011).

[19] The unit step function employed here is defined as follows: $H(x - y) = 0$ if $x \leq y$ and 1 if $x > y$. $H(x)$ equals 1 *if x strictly exceeds y* and equals zero otherwise. So, in [6], A_i equals 1 when the disposition dot product exceeds threshold. The Heaviside function is commonly used as a switching function in engineering, and is named after Oliver Heaviside, British mathematician. Some forms assign a third value, such as ½, to the case where $x = y$, and some variants use $H(x - y) = 0$ if $x < y$ and 1 if $x \geq y$. We use the specific definition given.

[20] So, for n agents, $\boldsymbol{\omega}_{ji} = (\omega_{1i}, \omega_{2i}, \ldots, \omega_{ni})$.

Skeletal Equation

Focusing on Agent *i*, this rendering of total disposition[21] is as follows:

$$D_i^{\text{tot}}(t) = V_i(t) + P_i(t) + \sum_{j \neq i} \omega_{ji}(V_j(t) + P_j(t)). \qquad [7]$$

As per [3], if at any time Agent *i*'s total disposition $D_i^{\text{tot}}(t)$ exceeds her threshold, τ_i, then $A_i = 1$ (action is taken). Otherwise, $A_i = 0$ (no action).[22] Others' thresholds do not enter into Agent *i*'s net disposition.[23]

Not Imitation of Behavior

Now, notice that in equation [7], the value of *A*, that is, the agent's *behavior*, is not an input. And the *A*s of the other agents are not inputs either. No one's *A* appears on the right-hand side of this equation. *Hence, the mechanism of action cannot be imitation of behavior, because the binary acts of others are not registered in this calculation. So we are suspending an assumption central to the literature on social transmission*. Let us see how far we can get without it.

In principle, the sum in equation [7] extends over all humans in existence. Those not in one's network have weight zero. Socially isolated agents are, of course those with *all* non-self weights equaling zero. Throughout this book, the skeletal equation [7] alone will prove to be of interest.[24]

[21] The relationship between [6] and [7] is evident if in [6], one deletes τ_i, sets $\omega_{ii} = 1$, and multiplies out as stipulated in the preceding footnotes.

[22] None of this asserts that Agent 1 has access to V_2 and P_2. Indeed, as we will argue at some length, individuals may not even be aware of their own values. The model is flexible, however, and permits the user to explore myriad assumptions regarding any individual's imputation of disposition to others. For completeness' sake, notice again that in [7] a single weight is applied to the sum of V_{j} and P_{j}, not to either individually.

[23] In one particular extension, the Latané-Darley experiment, I have agents impute a threshold to others, because in this case the observability of others' behavior is essential to the result.

[24] Denoting the social term $[\sum_{j \neq i} \omega_{ji}(V_j(t) + P_j(t))]$ as *S* and suppressing time (*t*) for notational convenience, the skeletal disposition formula [7] is simply $D = V + P + S$: passion plus reason plus social. Mathematically punctilious readers will notice that this is different from considering *V*, *P*, and *S* as orthogonal bases of a dispositional vector space, a possible conceptualization mentioned earlier. This is because my model requires disposition to be a single real number comparable to a threshold, and the sum of vectors is not a real number, but an *n*-tuple. However, if the reader is wedded to the vector space concept—which I think is worth pursuing—one could define the basis vectors as $(V, 0, 0)$, $(0, P, 0)$, and $(0, 0, S)$. Their vector sum is (V, P, S), and my sum (*D*) is then simply the dot product of this with the radial vector (1,1,1). So, there is a way to reconcile the basis picture with my superposition. But I have directly adopted the latter for present purposes. The vector picture invites an intriguing geometrical development, however. Staying with two dimensions, the radial vector is (1, 1). Plotted in the (*V*, *P*)-plane, the radial vector represents points of equipoise

Notational Distinction

In case it is not clear, every agent in the model will be of the *Agent_Zero* type, but particular agents (instances of that agent class) will be denoted Agent 1, Agent 2, and so forth. This will be the convention throughout.

Specific Components

Now what are $V(t)$ and $P(t)$? The (emotional) functions $V_i(t)$ will be solutions of the famous Rescorla-Wagner (1972) model of conditioning.[25] The (evidentiary) functions $P_i(t)$—the bounded rationality component—will be probability estimates based on local sampling of a dynamic spatial landscape of events (stimuli).[26] And the social network is encoded in the

between passion and reason. If we imagine our passion-reason coordinates to be (V, P), we have from vector analysis that

$$(V,P)\bullet(1,1) = \|(V,P)\|\|(1,1)\|\cos\theta,$$

where θ is the angle between the passion-reason vector (V, P) and the radial vector (1,1). It follows that

$$\theta = \cos^{-1}\left(\frac{V+P}{\sqrt{V^2+P^2}\sqrt{2}}\right).$$

But, since the numerator is just disposition, D, we may express this as

$$\theta = \cos^{-1}\left(\frac{D}{\sqrt{V^2+P^2}\sqrt{2}}\right),$$

the angle between the nonzero (V, P)-vector and equipoise. If $\theta > 0$ then $V > P$, and so forth.

[25] I will actually introduce a *nonlinear* generalization of the original model, allowing some flexibilities of interest.

[26] Net of the threshold, the skeletal equation [7] is

$$\forall i, D_i^{\text{tot}}(t) - \tau_i = V_i(t) + P_i(t) + \sum_{j\neq i}\omega_{ji}(V_j(t) + P_j(t)) - \tau_i$$

The symbol ∀ is the universal quantifier meaning "for every." Imagining just two agents, one could collect V and P terms, expressing the right-hand side as

$$(V_1 + \omega_{21}V_2) + (P_1 + \omega_{21}P_2) - \tau_1.$$

The second of these terms may, in fact, exceed unity. This might be an issue if one insisted on interpreting this term as Agent 1's probability estimate. I do not. His probability estimate is P_1, pure and simple. Other peoples' probability estimates do affect his overall disposition, but they do not affect his probability estimate proper. Now, I say, "might be a problem" because, while the probability axioms of course preclude probabilities exceeding one, innumerable psychology experiments establish directed weights, the ω's. Again, it is *not* behavior that is contagious in these networks, but (solo) dispositions. This is just a sketch, and I will put more meat on these bones in Part I.

In the purely equation-based version of Part I, there is no spatial component, conditioning events arrive exogenously, and each agent's probability estimate is a fixed exogenous constant. Agents also do not have memory. Then, in the full agent-based version of Part II, agents roam a landscape of stochastic events, which form the basis of their immediate emotional and evidentiary states, to which network effects, memory, and further cognitive apparatus is added to produce overall action dispositions.

Coupled Trajectories

Coupled dispositional trajectories generate patterns of observable behavior in groups. Even on the emotional side alone, this is a very different picture than experimenter-subject "instructed" fear conditioning, for example. In the model developed here—and in the real world—*every agent is at once the experimenter and the subject. Each delivers stimulus to, and receives stimulus from, others, and learning is distributed, concurrent, and, often, unconscious.*

ORGANIZATION

The book is comprised of four main parts.

Part I: Mathematical Model

Part I develops the basic mathematical, explicitly equation-based, model. Relevant cognitive science is discussed as model components are developed. The most extensive discussion surrounds the neuroscience of fear

that humans make fundamental errors here. So, while I do not interpret the term of interest as a probability estimate, such an assumption could, in fact, be defensible as a model of human risk judgment. For example, suppose you ask people, What is the probability of rain today? They answer *x*. Now (knowing that no clouds implies no rain) you ask them, What is the probability that there are no clouds today? They give some answer *y*. It may well be that, from *x* and *y*, an overall probability exceeding unity is deducible. All this means is that human appraisals of likelihood violate the formal axioms of probability theory. But this is hardly controversial (See Kahneman and Tversky, 1972, 1996; Tversky & Kahneman, 1974; Dawes and Corrigan, 1974; Dawes, 1999; Fischhoff, Slovic, and Lichtenstein, 1979; and Lichtenstein, Fischhoff, and Phillips, 1982. I am not sure the term in question requires a name, since I feel no compulsion to group terms in this manner. But, forced to produce one, I might call it simply the model's "propositional output." It is the entire deliberative component, formed from the individual's *P* proper and a weighted sum of others' *P*s proper. In any event, I do not find this to be problematic.

acquisition, which will dominate this exposition of the model's emotional component. But this is not a model of fear-driven behavior specifically. The model is general, as are the Rescorla-Wagner equations themselves, which govern many sorts of associative learning, a variety of which we shall explore. Important research in other areas, particularly social conformity effects, and cognitive biases are also discussed in the relevant sections. Part I generates the mathematical version of a key phenomenon: *the agent who initiates group action (e.g., violence) despite having no evidence, no adverse feelings, and no orders.*[27] The entire book moves back and forth between mathematical and agent-based models; this dialogue enriches both formulations.[28]

Part II: Agent-Based Model

Part II presents the second of these formulations: it is agent-based and computational. Blue individuals move about an explicit landscape of yellow patches, which we initially imagine as an indigenous population. At a stochastic rate set by the user, these yellow patches turn orange, "attacking" the Blue agents. These attacks are the conditioning trials whereby Blue agents come to associate "the Yellow face" with an aversive stimulus—an attack. This association (as in the mathematical model) is governed by the Rescorla-Wagner model. However, here (unlike the nonspatial continuous deterministic Part I version) conditioning trials occur in discrete time and stochastically as agents encounter adverse events on a spatial landscape. Fear extinction is also included in the model and is explored. To this fear-conditioning/extinction process is added the estimation, by each agent, of the probability that a random patch is an enemy (i.e., will become an orange patch). This estimate is made by *local* sampling of the landscape within a user-specified "spatial sample radius."[29] This denotes a landscape sample radius purely, and enters into the P calculation only. By contrast, agents can influence each other—have dispositional weight—at *any* range, by a large variety of avenues, including auditory, visual, and social media. The spatial sample radius is normally a set of contiguous sites on the landscape proper, such as a von Neumann neighborhood.[30]

[27] I understand an "order" to be an explicit directive from an organizational superior carrying a penalty for disobedience. There is no mechanism for orders in this model. No Eichmann "defense" is available to *Agent_Zero.*

[28] There are, in fact, two different mathematical formulations: the skeletal equations, where V and P functions are not explicitly specified, and the fully fleshed-out model, where they are. Both formulations will prove to be of interest.

[29] Occasionally, we will nickname this "vision," though the search radius must not be confused with literal ocular acuity. In general, it refers to an information set.

[30] The four immediately neighboring lattice sites to the north, south, east, and west of the agent.

This local sample yields a biased estimate of the global probability, quite in keeping with a large literature on biases and heuristics (Gilovich, Griffin, and Kahneman, 2002; Tversky and Kahneman, 1974).[31] These are the agents' *direct V* and *P* values, and their sum is the agent's solo disposition. But through the network, the weighted solo dispositions of others are added, and the sum (the total disposition) is compared to the individual's threshold. When total disposition exceeds that action threshold, (i.e., when $D^{\text{net}} > 0$) the Blue agent wipes out all yellow patches within some destructive radius—colored a dark blood red.[32]

Once constructed, the model is shown to generate a variety of important social parables, including our "leader," who initiates violence against yellow innocents, without evidence, without fear, without orders.

Generality

The focus on violence, like fear, is an expository tactic. I could have presented the full uninterpreted formalism and then assigned interpretations to the variables, the space, the stimuli, responses, and so forth. But this would have demanded much of the reader. So, I chose to present the model in a familiar and obviously important interpretation. However, this is not "a model of violence" per se, but of behavior in groups. And, as discussed further shortly, the general formalism admits many other interpretations. They require simply that the space, the stimulus, and the reaction be reinterpreted.

Interpretations: Vaccine Refusal, Obesity, Economic Contagion

For example, in one public health interpretation, the space could be a landscape of pharmaceuticals, orange "attacks" would be adverse drug reactions, and the destructive radius the set of drugs (e.g., vaccines) refused by fearful individuals. Vaccine refusal is a serious problem in the mitigation of contagious diseases worldwide, and (to say the least) is not uniformly driven by an informed comparison of risks and benefits. Here, the roles of emotion, partial information, and peer effects loom very large.

[31] Later we also introduce memory, so that this computation of ambush probability can be a moving average, a moving median, or some other statistic computed over a memory window. The window and sample method can vary among agents.

[32] Here, flight is precluded, *defensive aggression* (Bloom, Nelson, and Lazerson, 2001, p. 252) being the result. In the extensions flight is explored.

In an obesity interpretation, the space could be a set of foods (e.g., *x*-axis is fat; *y*-axis is carbohydrates). Orange outbursts would be opportunities for unhealthy eating, and the destructive radius is bingeing. The binge radius could be small and include just deep-fried Oreos, or it could be large and include many other foods. Eating behavior—addictive behavior generally—is an excellent candidate for the *Agent_Zero* model, since it indeed does depend on passion (the associative strength between a food's consumption and pleasure), reason (conscious deliberations regarding health), and social influence (e.g., community norms regarding diet and ideal body type).

In an economic interpretation, the space could be a set of assets (e.g., financial instruments or real estate properties), aversive events are sudden collapses in value, and the destructive radius is the set of assets dumped in response. Prices, of course, may be important in many interpretations. As demonstrated later in the book, the introduction of prices is very natural.

So, while the main exposition (particularly of the agent model) is made in terms of violence, myriad interpretations are possible. In Parts I and II, the space, the stimulus, and the action are simply reinterpreted; the core formal model is unchanged.

Part III: Extensions

In Part III the model proper is altered in a number of simple but powerful ways. These are extensions. I present the following fourteen of them:

1. Endogenous destructive radius
2. Age and impulse control
3. Fight vs. flight
4. Replicating the Latané-Darley experiment
5. Introducing memory
6. Couplings: Entanglement of passion and reason
7. Endogenous dynamics of connection strength
8. Growing the 2011 Arab Spring
9. Jury processes
10. Endogenous dynamics of network structure
11. Multiple social levels
12. *The 18th Brumaire of Agent_Zero*
13. Prices and seasonal economic cycles
14. Mutual escalation spirals

They are a heterogeneous lot designed to suggest the fertility of the approach. The fourth of these, for example, is a computational replication of the famous 1968 Latané and Darley experiment from social psychology. Here, the space is a room. The stimulus is smoke entering the room. The reaction is flight from the room.[33] Latané and Darley found that the subject's reaction is strongly influenced by the presence of nonreactive others, a result generated in a surprising way by extending the model and offering a new mechanism for bystander effects, which I term *threshold imputation*.

Entanglement

In Parts I and II, the agents' affective, cognitive, and social components have been decoupled—while disposition depends on all the components, no component is *a function of* any other. They are not entangled. In reality, they are entangled: one's emotions can directly bias one's judgments of probability. I model this in extension (6) by introducing an affective bias into the probability estimation algorithm itself.

Homophily

Also, in the discussions of Parts I and II, the interagent weights are constants. In general, these coupling strengths can vary with similarities in status, expertise, reputation, religion, musical taste, and other attributes. The *Agent_Zero* framework can accommodate any such coupling scheme. In Part III, we extend the network model in various ways. In extension (7), a dynamic affect-dependent weighting is introduced. We model the weight between two agents as a product of affective strength (the sum of their affects) and homophily (one minus the absolute value of their affective difference). This endogenous dynamic weighting then plays a central role in our models of the Arab Spring, jury processes, and the evolution of network structure.

Growing the Arab Spring

Social media enable the process of link strengthening through affective homophily. Inspired by the Arab Spring of 2011, we develop an extension where the aversive orange stimuli (the conditioning trials) are instances of regime corruption. The threshold term represents the potentially rebellious agent's perceived risk of punishment. With no social media, there is no rebellion, despite high aversion. With social media enabling amplification through affective homophily, the threshold is exceeded, and Jasmine

[33] This is why the flight extension (3) precedes Latané-Darley, for example.

Revolutions[34] unfold. This is extension (8).[35] The mutual escalation spirals of extension (14) bring to mind the violent course of events unfolding in Syria at the time of this writing (December 2012), in which indiscriminate government killing of civilians has increased participation in the rebellion, to which the government has responded with yet further civilian attacks, driving further civilians to rebel, and so forth.

Network Dynamics

Now, in classical network theory, there are simply nodes and edges. Nodes are connected to other nodes or they are not. The degree of a node is the number of other nodes to which it is connected, a completely binary matter. Link *strength* is ignored; it is zero or one. However, one could legislate that a social link exists only if the connection strength (the interagent weight) exceeds some threshold. In that case, the very structure of edges—and the degree of each node—will depend upon the underlying affinity dynamics, as well as the threshold. When the affinity curve pops above the threshold, a link is said to exist. It dissolves if the affinity returns to subthreshold level.

The classical theory might then be seen as a kind of binary projection, or embedding—akin to a Poincaré return map (see Guckenheimer and Holmes, 1983; Hale and Kocak, 1991)—of the general continuous dynamics of connection strength. This is extension (10). We compare the network dynamics on affective homophily to those on probability homophily, showing that the evolutions are very different. As in the model, so in life, we not only belong to numerous networks at once, but all of them are changing.

Jury Processes (Twelve Angry *Agent_Zero*s)

As shown in extension (9), jury dynamics offer a very rich application area, one in which affective, cognitive, and social dynamics occur in different phases. In the courtroom phase, the prosecution and defense offer data and emotional stimuli, in a battle for the hearts (V) and minds (P) of the jurors. Intrajury interactions do not (in principle) take place in this phase. So, here the weights (ωs) are zero. But then the jury is sent off to reach a verdict. All the courtroom stimuli cease (they leave the landscape of stimuli) and move to the jury deliberation room. Some jurors may have been inflamed by the evidence (high V), while others were left emotionally cold. Some

[34] So named after the Tunisian uprising, for the Tunisian national flower (Frangeul, 2011).

[35] In 2011, for example, social media facilitated the spontaneous spread of Occupy Wall Street networks and demonstrations in many countries.

found the evidence to be suggestive of a high guilty probability (*P*), while others did not. For some jurors, the *P*-value itself was amplified by passion. Some employed large memory, some very little. In these affective and cognitive states, they enter the jury phase. Behind closed doors, it may be that, as in Yeats's Second Coming, "the best lack all conviction, while the worst are full of a passionate intensity." Here weights are not zero and drive the intrajury social dynamic toward a verdict. The literature demonstrates that conformity and momentum effects play large roles here, of the sort the model generates. See Hastie, Penrod, and Pennington (1983). In the model, networks emerge through affective homophily during the deliberation phase. We model the pretrial, trial, and jury-deliberation phases of a stylized case and show how changes of venue and network effects in the jury deliberations can shape verdicts.

Economics

In extension (13) prices are explicitly introduced, and observed seasonal economic cycles are crudely generated. All in all, the range of economic applications suggested is quite wide, including financial contagion, capital flight, and various marketing strategies. Further extensions are presented below.

Replicability and Research Resources on the Princeton University Press Website

Now, talk is cheap, and so are implicit mental models. All the runs and all the extensions are explicitly implemented in *Mathematica*, in *NetLogo* Code, or both. All *Mathematica* Code is given in Appendix II. The *NetLogo* Code for the Parable 1 run is given in Appendix III. A table of all numerical assumptions used in each of the book's 14 movie runs is given in Appendix IV. All the movies, interactive *NetLogo* Applets, *NetLogo* Source Code, and a complete *Mathematica* notebook, are available at the Princeton University Press Website (http://press.princeton.edu/titles/10169.html). The Website can be used in a number of ways.

Three Ways to Use the Princeton University Press Website

1. *Play Movies.* Every movie discussed in the book is posted on the Website.
2. *Run Applets*. User-friendly *NetLogo* Applets for *every movie run* discussed in the book are also posted on the book's site. These permit nonprogrammers to manipulate sliders on the user interface and to explore different assumptions or even change them on the fly—in the course of a run. The Applets thus have both pedagogical and research value.

3. *Implement and Modify Source Code.* The *NetLogo* Source Code for every movie run is also provided in the corresponding Applet.[36] This ensures that every run is replicable and offers the modeling community a large code base to modify, extend, and use for research. The Table of Appendix IV, "Parameter Settings for Model Runs," includes all the main parameters used. These, and even finer-grained details, are, of course, also available in the *NetLogo* Source Codes themselves.

These appendices and materials on the Website ensure that all results are *replicable.* I hope the English-language exposition of the model also permits its reimplementation. But if not, the *Mathematica* and *NetLogo* programs provided are definitive. Of course, this Code Library also provides a rich basis for further extensions and explorations.

Part IV: Future Research and Conclusions

In Part IV, some ideas for future research are presented, followed by the book's overall Conclusions.

Overarching Claim

My general claim is that wherever emotional, deliberative, and social components combine to generate behavior, the *Agent_Zero* framework can apply.[37] It specializes to purely affective, purely deliberative, and purely conformist individuals. These are the "basis elements" from which a space of recognizable actors can be generated. The versatility of the framework is demonstrated with generative models from the fields of social conflict, public health, economics, law, network theory, and social psychology. I offer *Agent_Zero* as a step in the direction of a unified neurocognitively grounded foundation for generative social science.

Explanation Is Not Justification

Finally, it is worth stating explicitly that the model does not defend or justify any of the behaviors generated: violence, financial panic, or unhealthy eating. Indeed, the entire point is to offer a deeper *explanation*[38] for such dynamics, precisely in order to control them, and to better know and control ourselves.

[36] Simply scroll down from the Applet Interface to find it.

[37] A good model is like a good fugue subject: it "supports" extensive development and interpretation. So, while some may object to *Agent_Zero*, I think Bach would approve.

[38] Again, on the distinctive features of generative explanation in social science, see J. M. Epstein (2006).