16

Advancing Agent_Zero

Joshua M. Epstein and Julia Chelen

Abstract

Agent_Zero is a mathematical and computational individual that can generate important, but insufficiently understood, social dynamics from the bottom up. First published by Epstein (2013), this new theoretical entity possesses emotional, deliberative, and social modules, each grounded in contemporary neuroscience. *Agent_Zero's* observable behavior results from the interaction of these internal modules. When multiple *Agent_Zeros* interact with one another, a wide range of important, even disturbing, collective dynamics emerge. These dynamics are not straightforwardly generated using the canonical rational actor which has dominated mathematical social science since the 1940s. Following a concise exposition of the *Agent_Zero* model, this chapter offers a range of fertile research directions, including the use of realistic geographies and population levels, the exploration of new internal modules and new interactions among them, the development of formal axioms for modular agents, empirical testing, the replication of historical episodes, and practical applications. These may all serve to advance the *Agent_Zero* research program.

Introduction

This chapter proposes selected research directions stemming from the *Agent_Zero* framework introduced by Epstein (2013). *Agent_Zero* is a new theoretical entity, an actor endowed with distinct emotional/affective, cognitive/deliberative,[1] and social modules. Grounded in contemporary neuroscience, these internal modules interact (and may conflict) to produce individual behaviors. When multiple agents of this new type influence one another, they collectively generate a wide range of important social and economic dynamics, some of which are very far from desirable.

As an alternative to *Homo economicus, Agent_Zero* is well suited to this Forum's aim of contrasting the evolutionary and "economistic" approaches. Rooted in the neuroscience of fear, conformity, and cognitive bias, *Agent_Zero*

[1] We recognize that not all cognition is deliberative and that terms like "emotion" are fraught. The mathematical operationalization given in Equations 16.4–16.7, however, is unambiguous.

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

highlights the double-edged nature of these evolved capacities—outputs of "survival circuits"[2] in Ledoux's (2012) terminology—offering a new generative mechanism for such important social dynamics as financial panic and genocide. As such, *Agent_Zero* connects strongly with other work in this volume, on mismatches between human welfare and the evolved preferences driving individual behavior. For example, some instances of Type 2 diabetes manifest the mismatch between our taste for sugar (evolved in the Pleistocene when it was scarce) and its modern overabundance due to agricultural and refining advances.[3]

The chapter is intended to be provocative and broad. Some of the research topics proposed require longer-term initiatives, while others—such as *Agent_Zero* "twin studies"—might be achieved in the relatively near term.[4]

Background to Agent_Zero

By way of background, the rational actor model and its refinements have dominated formal social theory since the work of Von Neumann and Morgenstern (1944, 1947) and Nash (1951).[5] Over the same period, psychology, behavioral economics, and neuroscience have accumulated systematic departures from canonical rationality. These departures come in (at least) three varieties.

First, *emotions* such as fear (often acquired unconsciously) drive human behavior (Barrett 2006; Frijda 1986; LeDoux 2012). Second, although conscious deliberations play a role, they are *boundedly* rational: information is partial and noisy, and our processing of it is subject to a variety of constraints and systematic errors (Gilovich et al. 2002; Simon 1955, 1978). Individuals driven unawares by powerful emotions, doing erroneous statistics on poor data, can exhibit irrational behavior even when acting alone. But, third, amplifying these effects, we humans are also unwittingly influenced by the

[2] LeDoux argues that the minimal set of survival circuits includes those involved in defense, maintenance of energy and nutritional supplies, fluid balance, thermoregulation, and reproduction. He emphasizes that "the survival circuits listed do not align well with human basic emotions. However, my goal is not to align survival circuits with basic emotion categories. It is instead to break free from basic emotion categories based on human emotional feelings (introspectively labeled subjective states) and instead let conserved circuits do the heavy lifting…Survival circuits are not posited to have any direct relation (causal role) in feelings. They indirectly influence feelings, as described later, but their function is to negotiate behavioral interactions in situations in which challenges and opportunities exist, not to create feelings." (LeDoux 2012:655).

[3] Among the earliest researchers to suggest this mismatch was Neel (1962), in the now-revised thrifty-gene hypothesis. For a thorough treatment, see Ayub et al. (2014).

[4] Epstein (2013) offers several topics for future research. The directions proposed here extend that discussion.

[5] The theory was presented in Nash's 1950 Doctoral dissertation, and published in the Annals of Mathematics in 1951.

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

emotions and erroneous deliberations of others (Aronson 1972, 2011). The result is that collective behavior can be alarmingly dysfunctional. People, when in groups, may find themselves taking actions diametrically contrary to those they would take in isolation. Indeed, they may even be the first in the group to take them!

Despite its incompatibility with these human social realities, the rational actor model remains the dominant approach, due in part to a dearth of *formal,* mathematical or computational, alternatives. *Agent_Zero* is one (Epstein 2013).

Cognitive Plausibility and Generative Explanation

The *Agent_Zero* model can play a central role in *generative* social science. The notion of generative social science and the generative explanatory standard that distinguishes it from other approaches are elaborated fully in Epstein (2006, 2013). The essential idea is this: To *explain* a social pattern, it does not suffice to furnish a Game for which the pattern is a Nash equilibrium.[6] Nor does it suffice to merely furnish a Functional (as in the Calculus of Variations) with respect to which the observed intertemporal pattern (such as a consumption trajectory) is extremal. Rather, one must *show how the pattern could emerge on timescales of interest to humans in a population of cognitively plausible agents.*

Cognitively plausible agents have emotions, they have bounded deliberative capacity, they have social connection, and all of these can interact to shape behavior. Accordingly, *Agent_Zero* is equipped with interacting emotional, deliberative, and social modules based in neuroscience.[7] In demonstrating that a population of plausible agents (a micro-specification) can in fact generate a social *explanandum*, the agent-based computational model is proving to be a central scientific instrument.

Agent_Zero: A Provisional Synthesis

To our knowledge, *Agent_Zero* is the first explicit formal mathematical synthesis of affective, deliberative, and social modules *within* an individual agent. As a first step in this direction, *Agent_Zero* is a new theoretical entity best seen as a *template* (Epstein 2013:183) for a family of models in which individual action is the *resultant* of such modules. In the simplest exposition (below), the individual actions are strictly binary and may be driven by nonconscious processes, some of whose neural mechanisms are quite well known (for a discussion of the neurochemical triggers of choice, see Glimcher, this volume).

[6] We use the term broadly, to include the now standard refinements.

[7] Hence the subtitle of Epstein (2013) is *Toward Neurocognitive Foundations for Generative Social Science.*

These mechanisms underpin human arousal, our rates of learning, our tendencies to conform, and our expectations of reward.

Agent_Zero represents an individual with competing and not necessarily conscious modules, specifically a model of conditioned fear (affective module), a simple relative frequency (deliberative module), and a weighted sum of other agents' affective and deliberative modules where the interagent weights are endogenously produced through affective homophily (social module). By interpreting and arranging a spatial stimulus landscape in different ways, the *Agent_Zero* entity is shown to generate a variety of notable collective phenomena: the slaughter of innocents, financial panics, fear-driven stampedes, and unexpected jury dynamics among them (Epstein 2013).

While the specific equations used to fill these affective, deliberative, and social "slots" in the template are defensible on contemporary neuroscientific grounds (e.g., Pape and Pare 2010), they are explicitly provisional and minimal (for a full discussion of model selection and a review of the supporting literature, see Epstein 2013: Part I). Using three agents of this type, interacting on an abstract landscape, Epstein demonstrates that *Agent_Zero* can exhibit a kind of self-betrayal, joining collective actions he or she would eschew if operating alone. And, by a (nonimitative) process of dispositional contagion, *Agent_Zero* may even *lead* such actions.[8]

All of this said, the published exercise (using a trio of agents on a toy landscape) is really a "proof of principle," and leaves many questions open. Following a compact exposition of the published model, we will discuss several lines extending and applying it, hoping thereby to bring a more advanced *Agent_Zero* research program into view.

The Agent_Zero Framework

We begin by presenting the overall structure of the model, its so-called *skeletal equations*. Then specific affective, deliberative, and social modules are presented, followed by two pictures of the published agent model in action. All source code, interactive applets, and movies (animated model runs) are freely available.[9] This brief exposition will permit the proposed research program to be articulated clearly. For all details, see Epstein (2013).

Agent_Zero Skeletal Equations

Agent_Zero is an autonomous software individual "who" can take actions. For simplicity, his or her possible actions are *binary*: join the mob or don't; buy the yacht or don't; consume heroin or don't. An action *A* is Boolean, a zero or a

[8] A succession of agents can act, oblivious to preceding agents' behaviors.

[9] http://press.princeton.edu/titles/10169.html (accessed Jan. 19, 2016).

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

one. Each agent is endowed with an affective and a deliberative module. These are equations or equivalent pieces of computer code. They change over time, as affected by patterns of stimulus. Formally, they are functions of time (t), denoted $V(t)$ and $P(t)$, respectively, defined on a stimulus space, and can each range from 0 to 1 inclusive. The emotion, $V(t)$, could be fear of indigenous attack, while occupying some foreign land, or of an adverse vaccine reaction, and $P(t)$ the agent's probability estimate that a random site is an imminent attacker or that a random vaccine will harm their child. The agent's *disposition to take action A* (destroy the indigenous village or refuse the feared vaccination) is defined as the sum of these:

$$D_i^{solo}(t) = V_i(t) + P_i(t). \tag{16.1}$$

In fact, V and P are themselves highly *nonlinear* functions of *time*. And, despite its appearance, this "linear" summation is a highly *nonlinear* function of time as well.

Agents are also connected socially. They carry weights (not always consciously), with $\omega_{ji}(t)$ being the weighted influence of agent j on agent i at time t. If there were just two agents, the idea would be that each one's total disposition to act in some way (e.g., fight or flee) is just the disposition they would have alone plus the other's solo disposition, weighted according to influence. Mathematically, in a group, the ith agent's total disposition to act is then:

$$D_i^{tot}(t) = V_i(t) + P_i(t) + \sum_{j \neq i} \omega_{ji}(t)\left[V_j(t) + P_j(t)\right]. \tag{16.2}$$

Finally, each agent has a simple trigger, an action threshold, $\tau_i > 0$. An agent acts (i.e., $A = 1$) if total disposition exceeds its threshold. Otherwise the agent does not act (i.e., $A = 0$). Above the threshold, it dumps the financial asset; otherwise, the agent holds it, for instance.

Dispositional Contagion, Not Behavioral Imitation

Crucially, no agent's observable action (no other agent's binary A) appears in Equation 16.2. Hence, *the mechanism of action cannot be the imitation of others' behavior.* The observable *behavior* of others is not registered in this calculation. Despite suspending an assumption central to the literature on social transmission—behavioral imitation—the *Agent_Zero* model is shown to credibly generate a panoply of collective phenomena, including agents who *join* group actions (e.g., a lynch mob) despite having solo dispositions far below their own action thresholds. They do things in groups they would not do alone. Moreover, they may even lead the group.

Universal Self-Betrayal

Indeed, to our knowledge, *Agent_Zero* is the first formal model in which agents are capable of universal self-betrayal, at least in the sense just described. Specifically, the following configuration can arise. For *every* agent, *i*,

$$D_i^{tot}(t) > \tau_i > D_i^{solo}(t). \tag{16.3}$$

Alone, each agent would acquit the accused. But together, they unanimously convict. Alone, each would be nonviolent. But together they destroy the village. Moreover, the behavioral mechanism is not imitation or even choice, as usually construed. Rather, behavior is generated by powerfully evolved circuitry whose activation is neither accessible to conscious evaluation nor easily overridden by rational suasion. It is, to that extent, social science without choices.

Social Science without Choices

Standard critiques of the rational actor miss this point, focusing on that theory's assumptions of optimal behavior and full information. There is, however, an even more fundamental implicit assumption; namely, that of *conscious choice.*[10] Contemporary neuroscience is learning exactly how human behaviors are driven by nonconscious processes (LeDoux 2003; Glimcher, this volume). The *Agent_Zero* framework differs from existing models in positing an agent whose disposition to act is the result of competing, *not* entirely conscious, modules: emotional, deliberative, and social. The provisional constituents (specific functional forms for *V* and *P*) are as follows.

Affective, Deliberative, and Social Constituents

Epstein (2013) offers differential equation, difference equation, and stochastic spatial agent-based computational implementations of the initial model. Mathematically, affect v[11] updates dynamically by a classical conditioning process (Rescorla and Wagner 1972), the neural underpinnings of which have been extensively studied in the context of *fear.* If α and β are respectively the surprise and salience of an unconditioned stimulus (electric shock), and λ is the maximum strength with which it can be associated with a neutral (conditioning) stimulus (a light), then with successive pairings (light followed by

[10] James March challenges a similar assumption. With reference to several models of calculated rationality, he writes, "All of these kinds of ideas are theories of intelligent individuals making calculations of the consequences of actions for objectives, and acting sensibly to achieve those objectives. Action is presumed to be consequential, to be connected consciously and meaningfully to knowledge about personal goals and future outcomes, to be controlled by personal intention…" (March 1978:592).

[11] Consistent with Epstein (2013), we use lower case variable names for specific functions, and upper case in the skeletal abstract form.

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

shock), the associative strength $v(t)$ increases monotonically, but with diminishing marginal effect, and is bounded above by λ.[12] Specifically, in discrete time, the basic model is:

$$v(t+1) = v(t) + \alpha\beta\left(\lambda - v(t)\right). \tag{16.4}$$

Associative strength on "Tuesday" (i.e., trial $t + 1$) is the strength on "Monday" plus the scaled gap between maximum associative strength and Monday's value. If trials are ceased, the association dies out, toward the minimum associative strength of $\lambda = 0$. Accordingly, this "extinction" phase is modeled as:

$$v(t+1) = v(t) - \alpha\beta v(t). \tag{16.5}$$

A representative trajectory with learning and then extinction is shown in Figure 16.1.

Epstein (2013) generalizes the Rescorla-Wagner model to permit *S*-curve learning and other variants. The differential equation form[13] is as follows:

$$\frac{dv}{dt} = \alpha\beta v^{\delta}\left(\lambda - v\right). \tag{16.6}$$

With $\delta = 0$, we have the original model, whereas $0 < \delta \leq 1$ produces *S*-curves as in skill learning (Fitts and Posner 1967). Many alternative learning models present themselves, of course. This equation was the simplest initial choice, and one which enjoys basic experimental credibility.

In addition to acquiring affect (fear), agents also acquire data about their world and estimate the probability of an aversive event at each time, t, as the

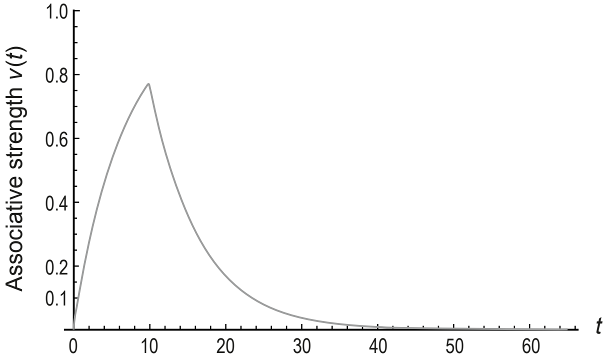

Figure 16.1 Acquisition and extinction (from Epstein 2013; used with permission from the Princeton University Press).

[12] The difference, $\lambda - v(t)$, is sometimes termed a prediction error. There are also settings in which the same basic equations apply to the maximization of reward or minimization of effort (see Skvortsova et al. 2014).

[13] Time is suppressed to reduce notational clutter.

moving average of its local (within agent vision) relative frequencies, *RF*, over a memory of *m* periods:

$$p(t)=\frac{1}{m}\sum_{t-m}^{t}RF(t). \tag{16.7}$$

So, if the memory is zero, agents live only in the moment and the probability estimate, *p*(*t*), is simply the ratio of "bad actors" (aversive stimuli) to total actors[14] within present vision (a sample that may change as agents move about or the environment changes around them). By employing a local relative frequency, *Agent_Zero's* deliberative component exhibits two prominent departures from canonical rationality: base rate neglect and sample selection bias (i.e., reliance on the representativeness heuristic).[15] This is the provisional deliberative module of *Agent_Zero*. Clearly, if the stimulus pattern (the *RF*) drops permanently to zero, then memory is eventually overwritten with zeros and the probability decays to zero. In a stimulus-free world, that is, both components (affect and probability) would be zero.[16]

Social Animals

While their affective and deliberative components are updating through external stimulus, agents are being influenced by the dispositions of others. These social influences are captured by interagent weights, which Epstein (2013) generates endogenously through affective homophily (Golub and Jackson 2012; McPherson et al. 2001). In this way, social connection is stronger among individuals who are emotionally aligned, an assumption which applies in many settings, but can be relaxed at will in the model (Lord et al. 1979; Miller et al. 1993; Taber and Lodge 2006). As shown in Equation 16.8, interagent weights are given by affective strength (the sum in brackets) multiplied by affective homophily (the difference in parentheses):

$$\omega_{ji}(t)=\left[v_i(t)+v_j(t)\right]\left(1-\left|v_i(t)-v_j(t)\right|\right). \tag{16.8}$$

This produces a mechanism of network formation and change based on strength-scaled affective homophily, and is an alternative to so-called preferential attachment (Barabasi and Albert 1999).

[14] These "bad actors" are not of the *Agent_Zero* type, but are sites of the stimulus landscape, the orange sites explained in connection with Figure 16.2.

[15] This applies even if the relative frequency is itself executed without error, an assumption discussed further below.

[16] As discussed on the book's website (http://press.princeton.edu/titles/10169.html), if one wishes to arrange that *P* or *V* can depress disposition absolutely, one can introduce subthresholds, such as using *P* – 1/2 rather than *P* alone. Mathematically, these are equivalent to translations of the overall threshold and will not be pursued here.

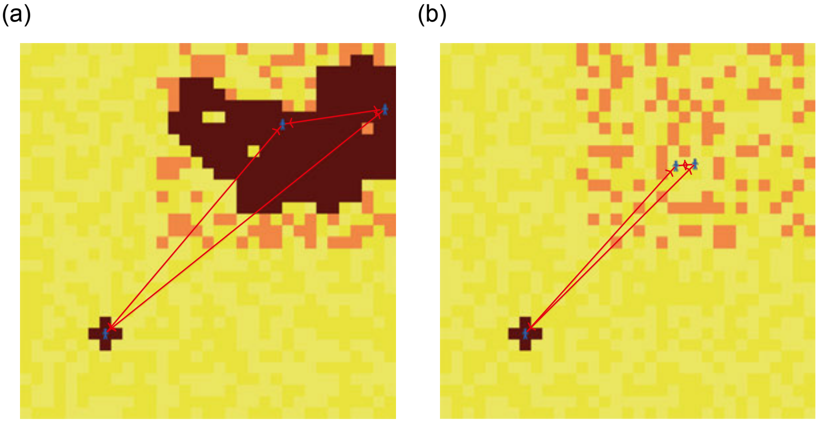

Figure 16.2 *Agent_Zero* in action: (a) *Agent_Zero* joins; (b) *Agent_Zero* initiates (from Epstein 2013; used with permission from the Princeton University Press).

Simple Model Example

Figure 16.2 illustrates how the model works. Three blue *Agent_Zero* individuals are situated on a landscape (a yellow indigenous population) of sites that stochastically activate, turning orange. These activations (ambushes) are fear-inducing stimulus trials. When dispositions exceed thresholds through direct adverse experience and/or network effects, the blue agents retaliate, destroying all (von Neumann) neighbor sites, innocent or not (destroyed sites are colored dark red). One agent (lower left-hand corner in both panels) is stationary and never receives direct stimulus (no orange activations); this agent's *solo* destructive disposition is therefore zero, far below the common threshold for action. Alone, this agent would never act. However, through interagent weights, the other two mobile agents influence him and, through dispositional contagion, he wipes out his innocent village (left panel). Under other settings, this stationary agent, despite lacking any aversive stimulus, can be the *first* to violence (right panel). Is this "leadership" or just extreme susceptibility to dispositional contagion? For a discussion, drawing heavily on Tolstoy (1869), see Epstein (2013).

Interpretations

One can interpret the space in Figure 16.2 as a set of pharmaceuticals. Orange bursts would be adverse drug reactions, with red squares the set of drugs refused through fear, imperfect statistics, and peer effects. Alternatively, one can interpret the space as financial assets, with orange outbursts being sudden losses of value and red patches assets dumped through contagions. The self-amplifying spirals generated by the model may afford a novel approach to the study of systemic risk in financial markets and epidemics. By interpreting and arranging the idealized landscape in different ways, the *Agent_Zero* entity is able to generate a wide variety of important collective phenomena: mob violence, financial crashes, fear-driven flight, and extra-evidentiary jury dynamics (Epstein 2013).

These are examples of what Epstein (2006) terms *generative minimalism*, the aim being to develop the simplest model capable of generating a range of core social phenomena, and replicating a number of small-*n* psychology experiments, among them the seminal Latané and Darley (1968) experiment on bystander effects. Here, subjects situated alone in a waiting room exit soon after smoke begins coming under the door. But if two others (confederates of the experimenter) behave heedlessly toward the same smoke, the erstwhile subject takes roughly three times as long to leave.

The model also produces recognizable market dynamics, like annual oscillations in price, driven by seasonal changes in production costs or regular fluctuations in demand, as in seasonal holiday purchases (Epstein 2013:168–176).

Agent_Zero is Purposive

While *Agent_Zero* is not canonically rational, this agent is arguably purposive. One can think of *Agent_Zero* as taking actions that seek to reduce aversive stimulus: wiping out attacking sites, or fleeing contaminated ones. In acting to minimize aversive stimulus, *Agent_Zero* could be interpreted as a disposition minimizer.[17] Were one to introduce a cost of action, one could frame *Agent_Zero* as a constrained optimizer, subsuming a more orthodox boundedly rational actor.[18] This could be an interesting extension in its own right, but there are many more. Hence, we next offer several projects, with our thoughts in various stages of maturity, which together might constitute an *Agent_Zero* research agenda.

Elements of a Research Agenda

Realistic Geography and Movement to Grow Spatial Patterns

Epstein's (2013) published landscape is, as illustrated in Figure 16.2, a small uniform grid of equally accessible patches. Except for the example of flight—where Epstein assigns agents a destination—movement is simply a local random walk. Of course, normal people do not typically execute local random walks. They might, for instance, move according to a search algorithm, ascending a gradient of some kind. By contrast, they could have detailed spatial itineraries, in whose execution they may face physical barriers and transportation

[17] We thank Erez Hatna for this insight.

[18] This would be particularly so if the agent's deliberative component were Bayesian, as discussed below.

constraints. Such movement rules and constraints shape the agents' direct social interactions and with them the collective patterns that emerge in society, such as clustering.

Many social phenomena—from urban segregation to voter turnout—exhibit spatial clustering. One way to test *Agent_Zero* on realistic geographies would be to see if the model generates the same spatial clustering statistics as the real world. Principal among such statistics is Moran's index *I* of spatial association (Getis and Ord 1992) applied to binary data (Griffith 2010; Lee 2001). A value of Moran's *I* close to zero represents a random pattern of action whereas a value close to 1 represents complete "segregation" of decisions (Anselin 1995).

Moran's Index is well suited to test a binary choice model such as *Agent_ Zero* who, recall, takes action if, and only if, total disposition exceeds threshold, producing binary spatial patterns of agent behavior. Moran's *I* offers a global measure of patterns but can be refined to detect local patches, "hotspots" of similar behavior (e.g., a hotspot of vaccine refusal, crime, political violence, or heroin addiction). For instance, the summer of 2015 saw a resurgence of measles in California due to within-state clusters of vaccine refusal (Majumder et al. 2015). In the case of the 2014–2015 West African Ebola epidemic, burial ceremonies were local hotspots crucial to spread. For this purpose, one can use the *local* Moran's *I* (Anselin 1995; Zhang and Linb 2007).

Increasing the number of agents (from Epstein's trio) goes hand in hand with this increase in spatial realism, but raises issues of its own, not all of which are computational.

Scale the Model up to Larger Population Phenomena

There was a time when the main hurdle to the scaling up of agent models was computational. These obstacles have largely been overcome, in part by Moore's Law and in part by innovations in parallel programming itself. Both are demonstrated by the 6.5 billion agent planetary-scale infectious disease model (Parker and Epstein 2011). In the present context, then, the outstanding issues in scaling up are behavioral, not technical.

For example, at the level of three agents, Epstein (2013) could employ simple addition (weighted linear superposition) of others' dispositions in arriving at the individual's total disposition to act in a group (as in Equation 16.2). While a convenient and defensible default assumption, this linearity could certainly break down at higher numbers.

Behavior at Scale

One departure from linearity is an important kind of affective diminishing marginal returns. The *first* person you encounter with symptoms of smallpox may have a huge emotional (fear) impact on you. But the 101st person may have little impact, relative to the 100th. As another example, millions can be riveted by

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

the drama of a single child trapped in a well, yet unmoved by mass famine, malaria, or genocide. Paul Slovic (2007) has led the empirical study of this "psychic numbing" as populations grow, an important effect to capture as we scale up to more realistic populations. Slovic and his colleagues have also pioneered the empirical study of what they term "the social amplification of risk," noted earlier in connection with financial panic (Pidgeon et al. 2003). *Agent_Zero* appears well suited to replicate larger empirical studies of this phenomenon, in which a seemingly small and localized stimulus can—through contagion and amplification—have system-wide, even system-destroying, effects.

As discussed earlier, generativists wish to "grow" such macroscopic regularities from the bottom up, in populations of plausible actors, of which *Agent_Zero* is one. But many variations are worthwhile.

Explore Intra-Modular Variations

On the deliberative side, for example, agents could update probabilities using Bayes' Rule (rather than simply a relative frequency). The memory apparatus now in place would also support agents executing "best reply to recent sample evidence" (Young 1993) in which the stored distribution of recent binary stimuli (enemy, friendly) could be used to project the next encounter, or encounters even farther into the future. Such extensions would include modules where a homeostatic *target*, π (a goal), is compared to a true outcome, η, and a behavioral adjustment is made based on the difference $\xi = \pi - \eta$, or some function thereof.

Replacing the Rescorla-Wagner equations with temporal difference learning (Sutton and Barto 1998) would be one such update.[19] Indeed, Glimcher's work (2011b) demonstrates that the dopamine system encodes a reward-prediction/error-correction algorithm of just this sort. This finding does not mean that we need to model the dopamine system proper at the molecular level. It simply identifies the neurobiological systems underlying this *regularity in human performance*. It thus gives us some confidence in using temporal difference, or more general functions like the probabilistic *SoftMax* algorithm (Beeler et al. 2010) as a model of the regular performance.

As another variation, both the affective and deliberative modules could be set up as so-called PID controllers.[20] These update their actions (e.g., a thermostat's setting, a car's cruise control, or ship's steering angle) using a term proportional to the current error (the *P*), the integral of past errors (the *I*), and the derivative (rate of change) of the error (the *D*).[21] The inclusion of memory (the integral term) and an immediate assessment of rates (the derivative term) can stabilize systems where adaptations proportional merely to current *actual*

[19] For a review of learning theory developments since Rescorla-Wagner, see Niv (2009) and Clark (2013).

[20] We thank Robert Axtell for introducing us to this area.

[21] Properly speaking, each of these terms is multiplied by a gain (Åström and Hägglund 2006).

error can self-amplify and explode—like a motorcycle that amplifies a small wobble until it crashes. The novelty would be that a society of PID controllers would be interacting.

A strength of agent-based modeling is the ease with which heterogeneity can be represented. Accordingly, agents might well be heterogeneous by memory and by module. Could a small minority of predictive agents "tip" the population—through network effects—toward more rational appraisals, averting violent episodes borne of unchecked contagious fear?

Formal Axiomatic Foundations

There is certainly also room for purely foundational work. What qualifies as a module at all? What mathematical requirements must each module satisfy? Should one insist that any affective module be bounded above, and monotonically increasing given some standard stimulus pattern? Should it exhibit extinction (possibly including a zero extinction rate) given a cessation of stimulus? Should we attempt to formulate affective, and other modular, axioms analogous to those of utility theory, and prove general (and presumably contrasting) theorems outright? This has been fruitful elsewhere.

Explore Inter-Modular Variations

The connection *between* modules is a separate issue. The affective and deliberative modules may be completely independent and "walled off" from each other. But they may also be inextricably entangled.

An example of the former situation is so-called phantom limbs. I *know* and can literally see that my amputated arm is gone, but I "feel" my arm. The feeling is "walled off" from rational suasion, like certain optical illusions. Are there phantom social biases? I *know* this group to be equal, but I have been conditioned (e.g., by relentless war propaganda) to associate their eye shape or accent with inferiority. Of course, it is very useful to know the difference between a phantom arm and a true one, lest you rely on the former to break your fall, and end up hitting your head. Likewise, it is important to recognize that you have been conditioned, lest you be manipulated into hitting someone *else's* head, or interning them baselessly.

To crudely reflect some of this, the skeletal equations could be modified to put weights on the affective and deliberative components (axes) of the agent's solo disposition. For phantom limbs, set the first weight to 1 and the second to zero. I "feel" the limb knowing full well that it is not there (i.e., that the probability of its presence is zero). A flexible variation on Equation 16.1 would be to let solo disposition be the convex hull of passion and reason.

Even this variation would leave the modules independent in the sense that neither variable V nor P is a mathematical *function of* the other. However, we know that modules can be fundamentally entangled. Glimcher (p. 91, this

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

volume) notes that changes in the evolutionarily ancient hypothalamus, "appear to be able to influence the activity of circuits that control human risk attitudes in a measurable and predictable way (Symmonds et al. 2010). Subjects who are hungry seem to behave differently, even in stock markets, due to changes in the hypothalamic state (Levy et al. 2013). There is a rich and growing literature on how values are influenced by these types of inputs." Hormonal states are also observed to affect decisions. As Glimcher recounts (p. 91–92, this volume):

> A classic example stems from the work of Ernst Fehr and his coworkers (Kosfeld et al. 2005), which examined the effects of the hormone oxytocin (a hormone associated with pregnancy, care for the young, and pair bonding in females) on social decision making. They found that higher levels of oxytocin increase "an individual's willingness to accept social risks arising through interpersonal interactions" (Kosfeld et al. 2005:673; for a more recent extension of our understanding of this phenomenon, cf. De Dreu et al. 2010).….A rich literature is emerging which documents how a number of neurochemicals influence the activity of the brain areas that contribute to valuation, and thus alter valuation and choice (Crockett and Fehr 2013).

Again, none of this means we need to model these biochemical channels. We want elegant models of the *performance enabled by* this circuitry, not models *of* the circuitry, necessarily.

Epstein (2013) proposes candidate functional forms in which P is a direct function of V, so that emotions can directly influence probability judgments.[22] This entanglement of emotion and reason could be taken much farther.

Of particular interest, however, are cases where the frightfulness of an event (e.g., Ebola, a plane crash) inflates our estimate of its likelihood. Epstein (2013) offers one way in which to entangle affect and cognition mathematically. Motivated by Zillmann's experiment,[23] he lets the level of affect V enter explicitly into the agent's probability calculation P, with emotion inflating the probability estimate (in log linear fashion). Letting P_n denote the emotionally neutral value computed as in Equation 16.7, and P_e the emotionally inflated value, Epstein (2013) posits that:[24]

[22] For a comprehensive review of emotions as drivers of decision making, see Lerner et al. (2015); for the influence of particular emotions on judgments of future events, see Lerner and Keltner (2000).

[23] Zillmann et al. (1975) proposed that high levels of sympathetic arousal reduce individuals' abilities to cognitively mediate their behavior, even when they are both given, and correctly process, information mitigating the source of the arousal. The authors examined this hypothesis in the context of provocation, anger, and retaliatory behavior and found 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" (Zillmann et al. 1975:282). Specifically, "the cognitively mediated inhibition of retaliatory behavior is impaired at high levels of sympathetic arousal and anger."

[24] We suppress time for notational economy.

$$P_e = P_n^{1-V}. \tag{16.9}$$

At the social level, this mechanism can turn isolated incidents into self-amplifying behavioral spirals, offering a simple mechanism for social amplification of risk (Pidgeon et al. 2003).

But how sensitive are these collective dynamics to the functional intra-agent relationships between affect and cognition? They are simply added in the solo disposition Equation 16.1. A defensible starting point (Epstein 2013:46–48), this is certainly not binding. What if a log-linear combination of the two is used (as in the Cobb-Douglas functional form of microeconomics)?

As noted earlier, emotions (e.g., fear) acquired through associative conditioning can decay when aversive stimulus stops[25] (this, again, is termed "extinction"). Memory acquired through observations can decay as well. But the two may decay at different rates, leading to jumps from action to inaction, as in the abrupt cessation of a destructive behavior.

Explore Other Forms of Homophily for Networks

Many social dynamics are propelled by networks based on shared fears or other emotional ties. Accordingly, the weight between two agents given earlier in Equation 16.8 uses strength-scaled *affective* homophily specifically. However, the code published by Epstein (2013) permits the use of *probability P(t)* in place of affect *V*(*t*) in the network weight formula (Equation 16.8).[26] As noted earlier, a moving average over a memory window is used to estimate the stimulus probability in Equation 16.7. One may also use (by a switch in the code) the moving median rather than the moving average. Another extension easily afforded by the code library is to combine these to make interagent weights a function of *dispositional* homophily.

These all merit exploration and might generate empirically testable hypotheses about social network dynamics. Naturally, one could also explore the effect of expanding, contracting, or *replacing* the memory window itself. In the latter connection, the implantation of a self-serving historical narrative is the very aim of revisionist history in general.

Grow Historical Episodes and Cults

The computational reconstruction of real historical episodes can, of course, be revealing, especially if disciplined by a clear question. In the case of the Artificial Anasazi research (Axtell et al. 2002; Dean et al. 2000; Gumerman

[25] Technically, the conditioned stimulus continues to be presented alone, so it ceases to predict the unconditioned stimulus (US).

[26] All code is available on the book's Princeton University Press site, http://press.princeton.edu/titles/10169.html

et al. 2003), the motivating question was whether purely environmental factors (rather than war or disease) sufficed to generate the observed demographic dynamics of the ancient Anasazi over the period AD 900 to AD 1350. At the latter point this civilization mysteriously abandoned their lands in what is now Arizona. (For the origins of this project, see Epstein 2006:88–89.) More technically, using only such an account, *is the observed history centrally located in the distribution of stochastic model realizations?* Merely reconstructing one episode—one settlement history or single epidemic—on a computer may be no more explanatory than the exact replication of a particular series of coin tosses.

This issue is not lost on leading researchers in the growing field of agent-based archaeology. For interesting work in this area, much of it explicitly based on the *Sugarscape* (Epstein and Axtell 1996) and Artificial Anasazi models, see Wurzer et al. (2015). This computational archaeology work explicitly adopts the generative explanatory standard, and exploits the integration of agent-based modeling and Geographic Information Systems (GIS), which have matured rapidly over the last decade. It includes attempts to model prehistoric mining (Kowarik et al. 2015), the economy of the Iron Age (Danielisová et al. 2015), Patagonian territoriality (Barceló et al. 2015), and other interesting episodes. For a computational reconstruction of ethnic segregation in Jaffa Israel since 1948, see Benenson et al. (2002).

Whitehouse et al. (2012a, b) describe yet other dynamics which might be productively modeled with *Agent_Zero*. Adoption (and abandonment) of religion is often based on emotionally shocking traumatic events, repeated associative trials, network effects with differential weights (affective), and binary behavior, as in switching from one cult to another (Whitehouse et al. 2012a:190). To model the *Kivung*, the authors focus on dysphoric rituals and the traumatic ordeals of initiation cults, which typically involve extreme forms of deprivation, bodily mutilation and flagellation, or participation in shocking acts. Such practices trigger enduring and vivid episodic memories. They write, "Traumatic rituals create strong bonds among those who experience them together" (Whitehouse et al. 2012a:221). The traumatic initiation ritual is a way of *manufacturing affective homophily.*

Experiments, *Gedanken* and Real

In *The Sciences of the Artificial,* Herbert Simon offers a wonderful thought experiment (Simon 1969:64). He imagines the tortuous path taken by an ant traversing a "wind- and wave-molded beach"…"Viewed as a geometric figure," he writes, "the path is irregular, complex, hard to describe." But its complexity reflects "a complexity in the surface of the beach, not a complexity in the ant." Then, he challenges us to consider the following parallel construction:

> [1] An ant, viewed as a behaving system, is quite simple. The apparent complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself.
>
> [2] A man, viewed as a behaving system, is quite simple. The apparent complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself.[27]

While there is now a large literature in the field of so-called *complex* adaptive systems, Simon's parallel above suggests that individuals are best conceptualized as *simple* adaptive systems, whose life course is importantly shaped by their complex dynamic environments.

Along these lines, it would certainly be interesting, and quite easy, to conduct "twin studies" on the effect of environmental variations on the life courses of *identical* software agents: *Agent_Zeros* raised apart, as it were. These would be computational *Gedanken* experiments. But real experiments can and should be replicated as well.

Epstein (2013) reproduces the very real Latané and Darley (1968) experiment ordinally,[28] and the Zillmann et al. (1975) and Asch (1956) experiments more qualitatively. Are there more recent experiments one could replicate using *Agent_Zero*? It would be useful to test whether *Agent_Zeros* would perform like individual humans in controlled laboratory experiments from neuroeconomics and behavioral economics. How would *Agent_Zero* play the Ultimatum Game? Can populations of these agents generate the macroscopic data on psychic numbing and the social amplification of risk accumulated by Slovic et al. (2002)?

Experiments might clarify the physical bases of dispositional contagion. Again, we distinguish sharply between imitation of observable action (which logically implies a preceding actor) and transmission of disposition (which does not). One may grant the theoretical fruitfulness of the dispositional contagion postulate, but fairly ask for more on its "mechanisms."

For instance, dispositional contagion can be facilitated by senses: visual, auditory, and olfactory.[29] Visual processes might include recognition of facial expressions or body postures indicative of emotions, such as fear or disgust. Auditory (e.g., by cell phone) and olfactory (e.g., pheromonal) signals may operate at different spatial and temporal scales. Physical space and assorted media may permit certain types of transmission and systematically exclude others.

[27] Simon (1969:64–65) qualifies [2], noting the role of memory. Very provocatively, he then casts memory as part of the "external" environment. "A thinking human being is an adaptive system; man's goals define the interface between his inner and outer environments, including in the latter his memory store" (p. 66).

[28] Departure in the absence of nonreacting confederates is systematically earlier than in their presence. No attempt is made to replicate the absolute timing.

[29] Sensory channels could be vestibular, thermoceptive, proprioceptive, nociceptive, chronoceptive, or interoceptive.

From "Complexity and Evolution: Toward a New Synthesis for Economics,"
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.

For emotion to be communicated through olfactory signals (Haviland-Jones and Wilson 2008), for instance, agents must be within some proximity (and perhaps wind direction) of one another. Multiple perceptual systems may be active simultaneously, and there may be lags suitably modeled in various ways.[30] Michael Chwe's (2013) discussion of *nunchi*, the Korean word meaning "eye reading," is quite interesting in this connection. Observational learning is yet another distinct channel (Mineka and Cook 1993; Olsson et al. 2007).[31]

In short, dispositionally salient signals may be transmitted through an array of deliberate and nondeliberate forms of communication, by various channels, at many ranges and rates. These transmission modes, and their interruption, might also be studied experimentally and introduced more explicitly into future extensions of *Agent_Zero*.

Practical Applications of Agent_Zero

Vaccine refusal was mentioned earlier. For example, measles is a deadly and vaccine-preventable disease, which saw a 2015 resurgence in California, due to parental refusal of childhood MMR vaccine. One project now underway seeks to replicate the California data on parental vaccine refusal. This is a good candidate for *Agent_Zero* since the behavior is driven by high emotion (fear of adverse reactions), partial and misleading information, and strong social network effects.

An *Agent_Zero* model could also inform the communication of risk. In particular, the model (and experimental results on learning) suggests that fear acquisition is fastest when surprise and salience are highest. Policy makers seeking broad adoption of vaccine might preempt excessive fear precisely by announcing the possibility of some adverse reactions. "Entertain failure" sounds like strange policy advice. But the theory suggests it could be quite helpful in sustaining the uptake of beneficial medications despite rare adverse events,[32] or in averting contagious financial panics, despite a few isolated bursting bubbles.

Taking Animal Spirits Seriously

On financial panic, mainstream economics usually dismisses Keynes's "animal

[30] One understudied approach is the use of delay-differential equations.

[31] Mirror neurons may prove to be relevant in this connection as well (Gazzaniga et al. 2014; Heyes 2010). Without recourse to mirror neurons, certain forms of observational learning may be explained by novelty and uncertainty, engaging the amygdala (Moriguchi et al. 2011; Somerville and Whalen 2006; Weierich et al. 2010).

[32] Commendably, the Centers for Disease Control notes that "vaccines are developed with the highest standards of safety. However, as with any medical procedure, vaccination has some risks...Individuals react differently to vaccines, and there is no way to absolutely predict the reaction of a specific individual to a particular vaccine." http://www.cdc.gov/vaccines/vac-gen/safety/ (accessed Jan. 19, 2016).

spirits” as qualitative hand waving, best captured in an exogenous error term (Keynes 1936/1973). Akerlof and Shiller (2009) part company with this orthodoxy and very engagingly advocate that economists take emotional factors seriously. *Agent_Zero* may allow us to endogenize animal spirits (e.g., fear), seriously including them in testable mathematical models of economic behavior.

Another important application area where panic can subvert the best laid plans is evacuation. In particular, fear-driven mass flight can produce congestion, undermining safe egress from theaters, stadiums, and cities. Obesogenic eating may also be a function of affect (deficient neurochemical reward systems), community norms, and imperfect information. Smoking and drug addiction are areas where a habit can take over, even when the subject knows it to reduce the duration and quality of their own lives (not to mention others). All of these are applied spheres to which neurocognitive agents can contribute.

Summary

We have explored a variety of research directions stemming from *Agent_Zero*, inviting work on the following natural topics: increases in the population scale and geographical realism of the model; the use of alternative affective and deliberative modules, and further entanglements among these; exploration of different endogenous mechanisms (homophily) in generating networks; experimental and statistical work unifying the micro and macro scales; formal axiomatic development, and practical applications of these more sophisticated and better-calibrated extensions. All of these are ways of advancing *Agent_Zero* as an integrated scientific framework, leading to deeper understanding of our behavior as individuals, and of the social dynamics we generate.

Acknowledgments

For detailed comments on the manuscript, we thank David Colander, Erez Hatna, John Padgett, and Scott Page. For various discussions and valuable suggestions, we also thank Robert Axtell, Lisa Barrett, Baruch Fischoff, Paul Glimcher, Alan Kirman, and David S. Wilson.

From “Complexity and Evolution: Toward a New Synthesis for Economics,”
David S. Wilson and Alan Kirman, eds. 2016. Strüngmann Forum Reports, vol. 19,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03538-5.