Agent_Zero and Generative Social Science

JOSHUA M. EPSTEIN

JOHNS HOPKINS UNIVERSITY

SANTA FE INSTITUTE

SOCIAL AND BEHAVIORAL SCIENCES SUMMIT

NATIONAL ACADEMY OF SCIENCES

OCTOBER 4-5 2016

Mainly…

Want to introduce you to a new theoretical entity by the name of

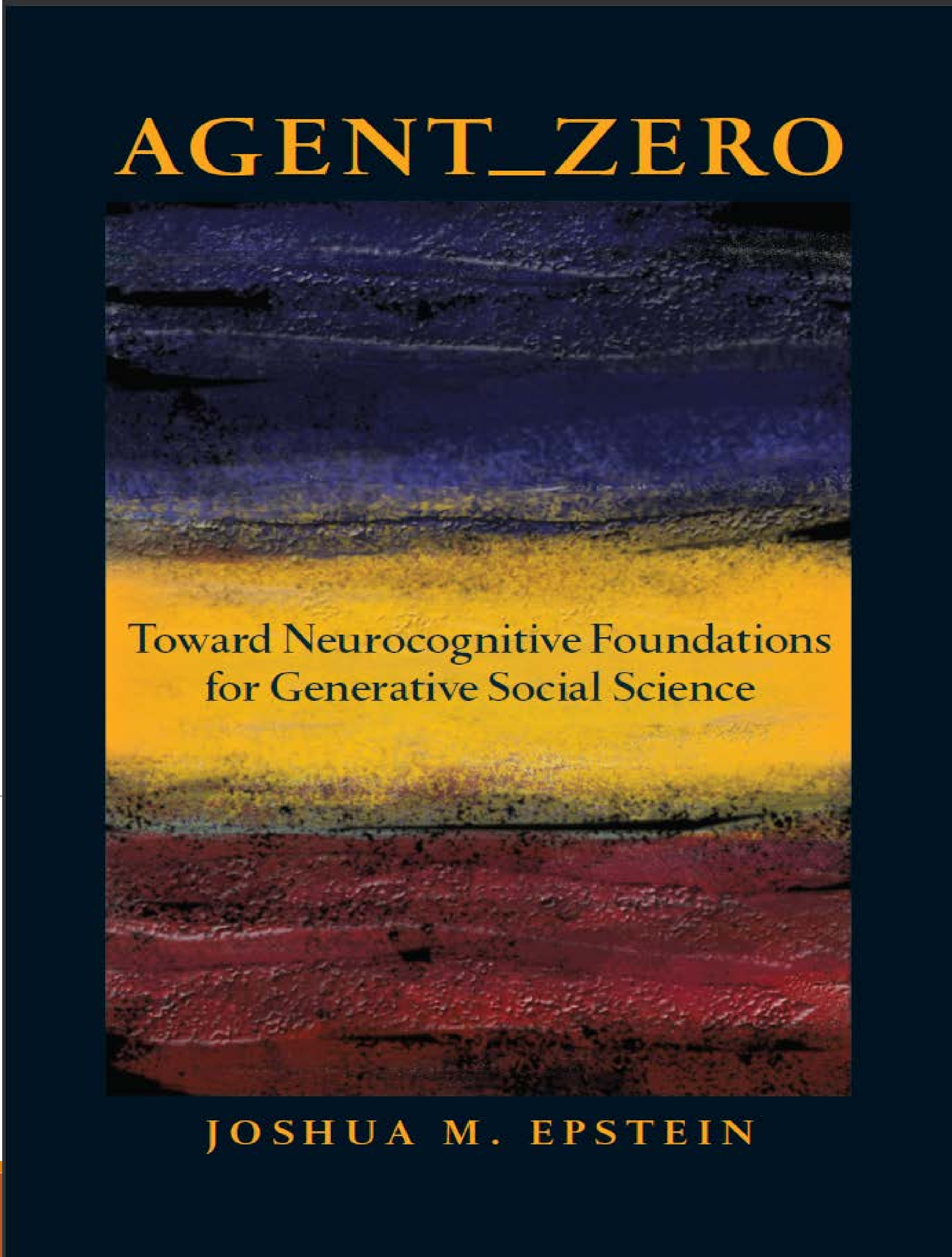
Agent_Zero

Recently published by Princeton:

Princeton Press

2013

NIH Director's Pioneer Award



AGENT_ZERO

Toward Neurocognitive Foundations for Generative Social Science

JOSHUA M. EPSTEIN

Agent_Zero

A neuro-cognitively grounded agent capable of generating a wide range of important social phenomena including collective violence, financial panic, endogenous dynamic networks;

A mathematically explicit functioning alternative to the rational actor, dominant since Nash;

Foundation for Generative Social Science.

Third in a Trilogy Concerning Explanation

Epstein and Axtell, *Growing Artificial Societies: Social Science from the Bottom Up* (MIT Press, 1996).

- Exploratory
- Immature Epistemology

Epstein, *Generative Social Science: Studies in Agent-Based Computational Modeling* (Princeton Press, 2006).

- Explanatory: Artificial Anasazi, Epidemics, Civil Violence, Classes, Retirement, Organizations.
- Mature Epistemology

Epstein, *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science* (Princeton Press, 2013)

- Cognitively plausible agent as foundation for generative explanations

Generative explanation

- To explain a social regularity
- Demonstrate how it could emerge on time scales of interest to humans in a population of cognitively plausible agents
- Does the micro-specification *m* generate the macroscopic *explanandum x*
- If so, *m* is a generative explanatory candidate.
- Motto (Epstein, 1999) is negative : If you didn't grow it, you didn't explain it.

$$\forall x(\neg Gx \supset \neg Ex)$$

- *Not* the converse (any old way of growing it is explanatory).
- *Not* uniqueness (might be many *m's*).
- Generative sufficiency a necessary (but not sufficient) condition for explanation.
- NOT: Furnish a Game in which the pattern is Nash
- NOT: Furnish a Functional with respect to which the trajectory is an extremal

Generative explanation

- To explain a social regularity
- Demonstrate how it could emerge on time scales of interest to humans in a population of cognitively plausible agents
- Does the micro-specification *m* generate the macroscopic *explanandum x*
- If so, *m* is a generative explanatory candidate.
- Motto (Epstein, 1999) is negative : If you didn't grow it, you didn't explain it.

$$\forall x(\neg Gx \supset \neg Ex)$$

- *Not* the converse (any old way of growing it is explanatory).
- *Not* uniqueness (might be many *m's*).
- Generative sufficiency a necessary (but not sufficient) condition for explanation.
  - ¬Furnish a Game in which the pattern is Nash
  - ¬Furnish a Functional with respect to which the trajectory is an extremal

Cognitively Plausible Agents

- Have emotions
- Have bounded deliberative capacity
- Have social connection
- And all of those might matter.

- Hume: “Reason is a slave to the passions.”
- Aristotle/Spinoza: “Man is a social animal.”
- Looking for a simple convolution of:

$$\text{Passion} \oplus \text{Reason} \oplus \text{Social}$$

Accordingly, *Agent_Zero*

Endowed with distinct affective/emotional, cognitive/deliberative, and social modules, grounded in neuroscience.

Internal modules interact to produce individual, often far-from-rational, behavior.

Multiple agents interacting generate wide variety of collective dynamics: health, conflict, network dynamics, economics, social psychology, law.

Get synthesis started.

All provisional….

But Formal

Lots of empirical criticisms of the rational actor.

Gripes do not change scientific practice.

Need explicit formal alternatives.

Agent_Zero is one: mathematical and computational.

Before laying out any equations …

Big Picture…where we're going. Conflict Interpretation.

Agents occupy an landscape of indigenous sites

There's a binary action agents can take: destroy some sites

Experience produces a *disposition* to take the binary action

Some sites are inactive/benign. Some active/fear-inducing

Affect: Agents fear-condition on local stimuli

- Passion

Bounded rationality: Local sample relative frequency

- Reason

Add these. Solo Disposition.

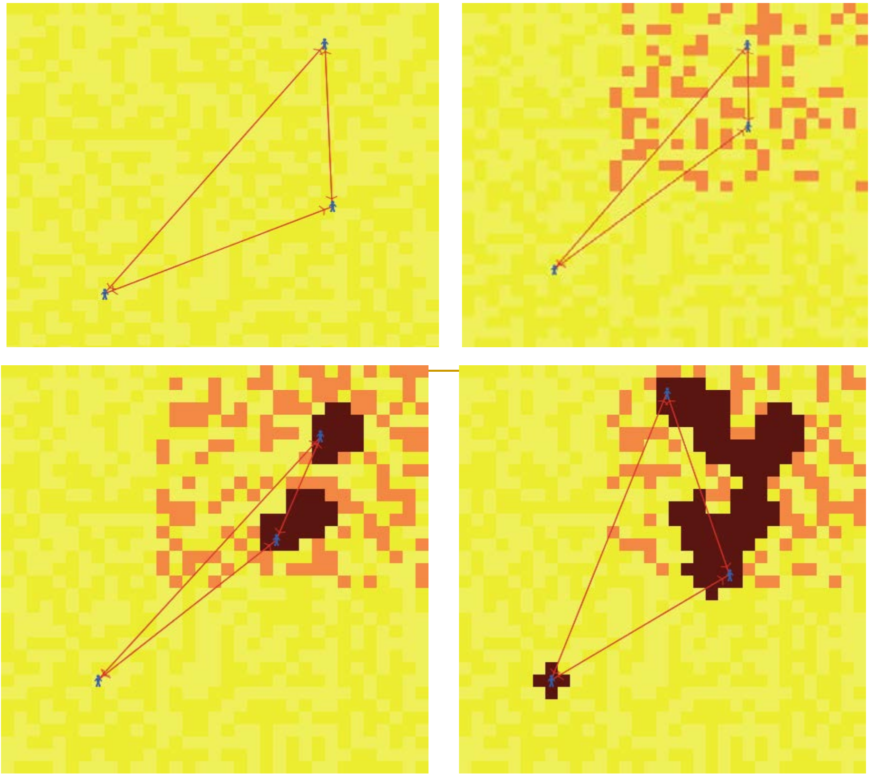
Social animals: Add others' weighted Solo Dispositions.

If Total Disposition exceeds threshold, take the action.

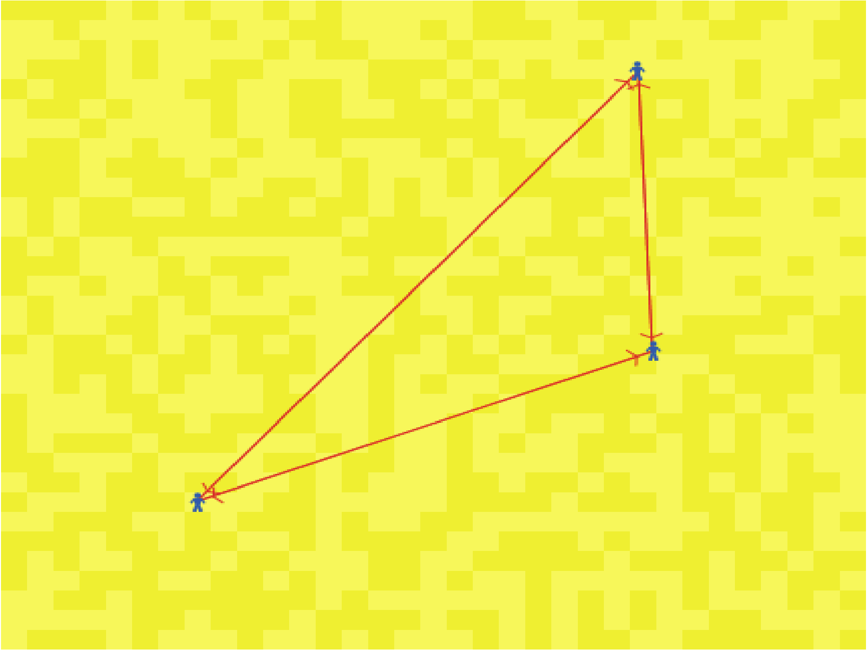
- Destroy

Computational Parable : Slaughter of Innocents

Vision Von Neumann
Agent 0 fixed in SW: zero direct stimulus
Others in NE: stimulus, violent action
By dispositional contagion, Agent 0 acts.



Agent_Zero Joins Without Direct Stimulus



V=P=0, since no stimulus within sensory radius.

Overall Set-Up

Action , Threshold

Binary Action $A \in \{0,1\}$

- Flee snake or don’t
- Raid icebox or don’t
- Join lynching or don’t
- Refuse vaccine or don’t
- Dump stock or don’t
- Wipe out village or don’t
- “Behavior” will mean a binary action.

Nonnegative Real Threshold $\tau \geq 0$

Total Disposition to Act

Agents endowed with Affective V(t) and Deliberative functions P(t) defined on a (stochastic) stimulus space.

Solo disposition is their sum: $D_i^{solo}(t) = V_i(t) + P_i(t)$

But Agents also carry weights (unconsciously I presume): $\omega_{ji}(t)$

We therefore define the Total Disposition to Act as:

$$D_i^{tot}(t) = D_i^{solo}(t) + \sum_{j \neq i} \omega_{ji}(t) D_j^{solo}(t)$$

Total Disposition to Act

$$D_i^{tot}(t) = V_i(t) + P_i(t) + \sum_{j \neq i} \omega_{ji}(t)[V_i(t) + P_i(t)] > \tau_i$$

If at any time, Agent i’s total Disposition exceeds her threshold τ_i then A=1 (binary action is taken). Otherwise, A=0 (no action is taken). More compactly:

$$\text{Act iff } D_i^{net}(t) > 0, \text{where } D_i^{net}(t) \equiv D_i^{tot}(t) - \tau_i .$$

Dispositional Contagion, Not Imitation of Behavior

No one’s binary *A* appears in 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.

Obvious problem with imitation of observable action: no mechanism for the first actor. Nobody to imitate.

(Noise is cheating…not a mechanism)

Specific Equations

Differential Equations and Agent- Based Computational Model.
Replicable Dialogue

$$D_i^{tot}(t) = V_i(t) + P_i(t) + \sum_{j \neq i} \omega_{ji}(V_i(t) + P_i(t))$$

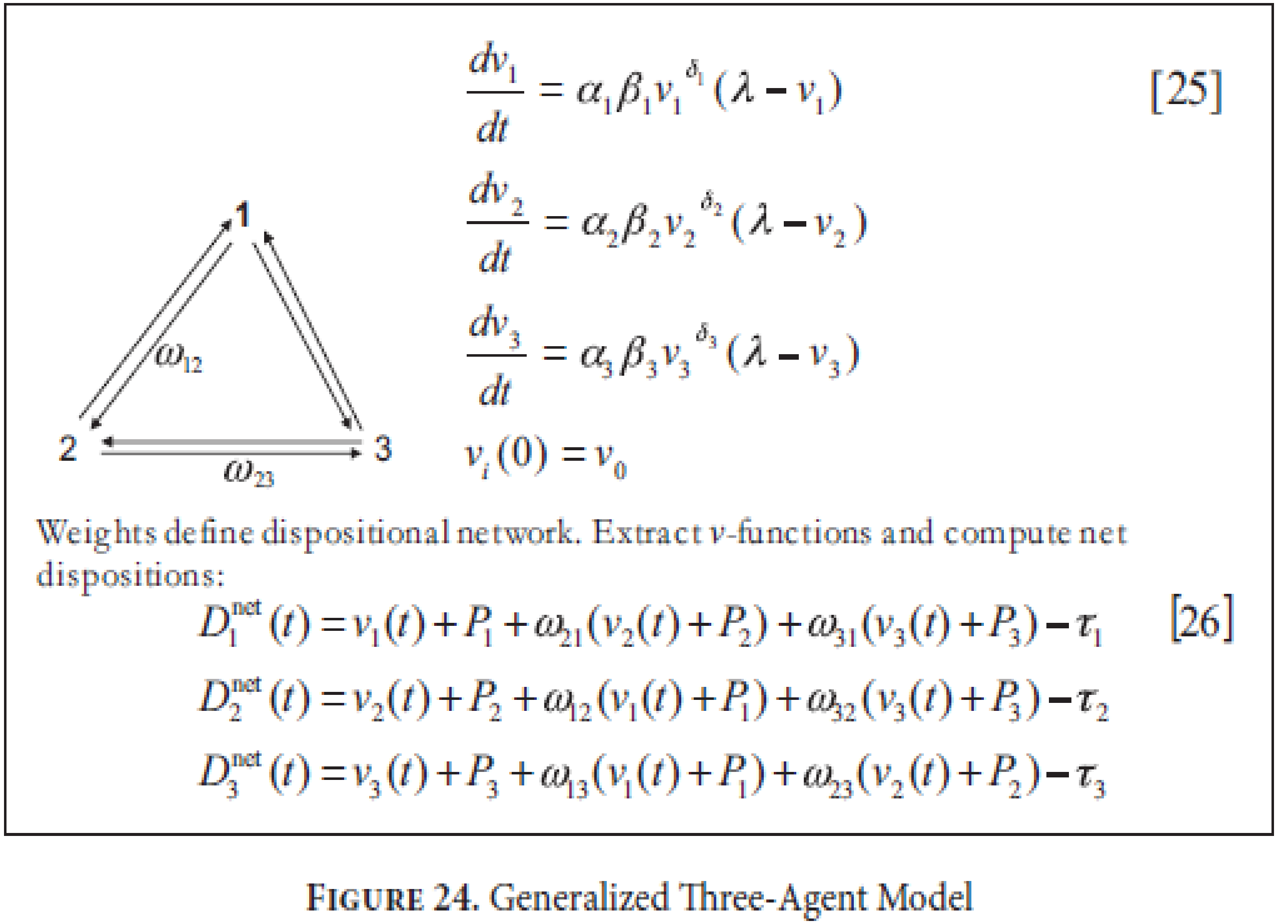
$$v_i(t) \text{ solves } \frac{dv_i}{dt} = \alpha_i \beta_i v_i^{\delta}(\lambda - v_i) \text{ Nonlinear Rescorla - Wagner}$$

$$P_i(t, x; m) = \frac{1}{m} \sum_{t-m}^{t} RF(x) \text{ Moving average of local relative frequency}$$

$$\omega_{ji}(t) = [v_i(t) + v_j(t)](1 - \left| v_i(t) - v_j(t) \right|) \text{ Strength - scaled affective homophily}$$

Where t meters trials for mobile agents on a spatial stimulus landscape

ODE version



$$\frac{dv_1}{dt} = \alpha_1 \beta_1 v_1^{\delta_1} (\lambda - v_1) \qquad [25]$$

$$\frac{dv_2}{dt} = \alpha_2 \beta_2 v_2^{\delta_2} (\lambda - v_2)$$

$$\frac{dv_3}{dt} = \alpha_3 \beta_3 v_3^{\delta_3} (\lambda - v_3)$$

$$v_i(0) = v_0$$

Weights define dispositional network. Extract v-functions and compute net dispositions:

$$D_1^{net}(t) = v_1(t) + P_1 + \omega_{21}(v_2(t) + P_2) + \omega_{31}(v_3(t) + P_3) - \tau_1 \qquad [26]$$

$$D_2^{net}(t) = v_2(t) + P_2 + \omega_{12}(v_1(t) + P_1) + \omega_{32}(v_3(t) + P_3) - \tau_2$$

$$D_3^{net}(t) = v_3(t) + P_3 + \omega_{13}(v_1(t) + P_1) + \omega_{23}(v_2(t) + P_2) - \tau_3$$

FIGURE 24. Generalized Three-Agent Model

The Subtitle of *Agent_Zero*

Toward Neurocognitive Foundations for Generative Social Science

Talked about GSS

What’s this neurocognitive business?

Fear Instantiation

Build Up Model in Fear Context (But More General)

Centrally implicated in many cases of

- Collective violence
- Mass flight
- Vaccine refusal
- Financial panic
- Salem (and other) witch hunts
- Stampedes

- Here I will butcher the very fine neuroscience of my NYU colleagues…

Amygdala Circuit

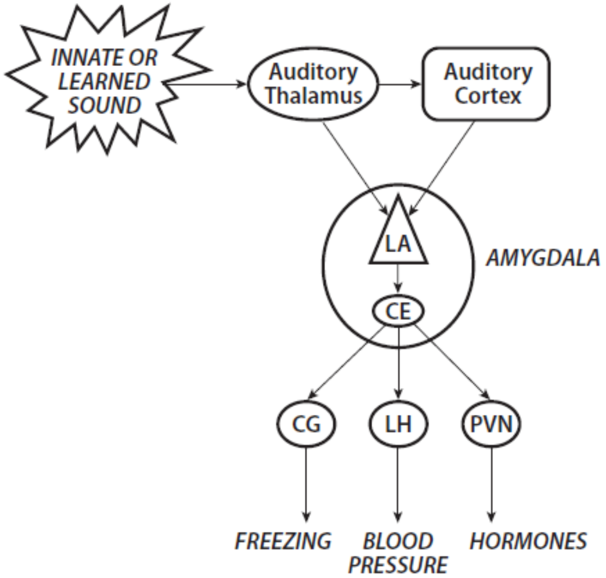


FIGURE 1. Auditory Amygdala Stimuli and Defense Responses. Source: LeDoux (2002, Figure 5.6)

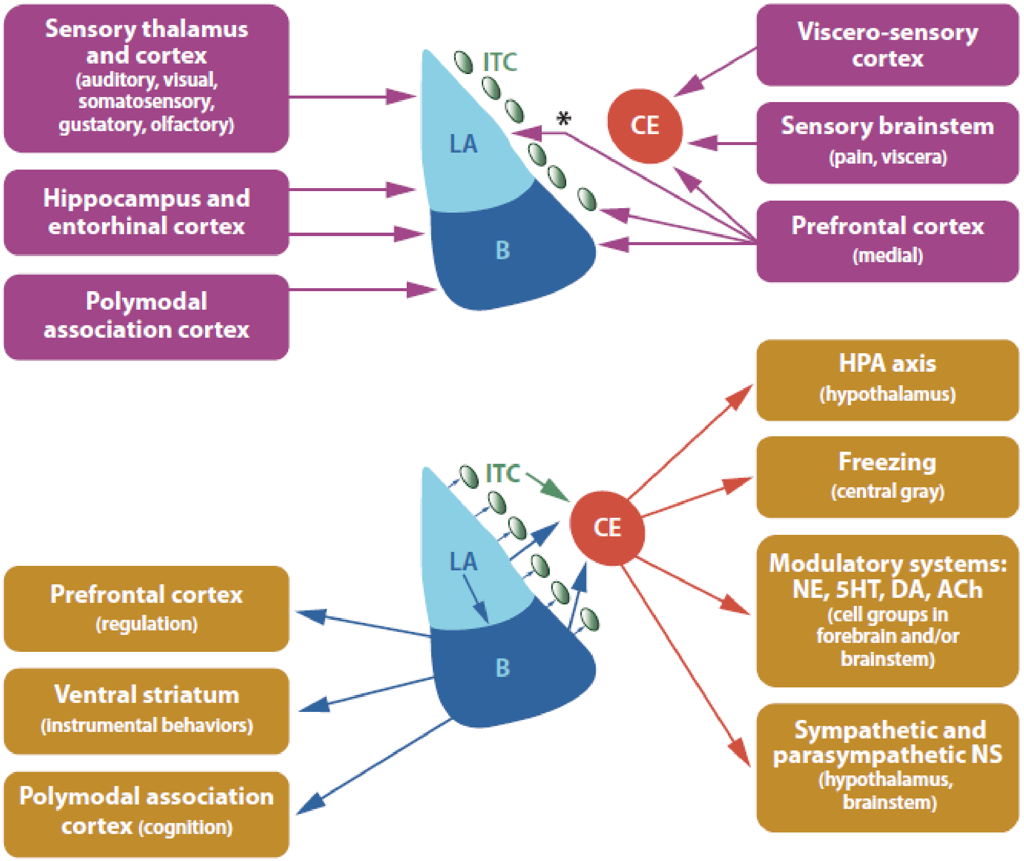

Sensory thalamus and cortex (auditory, visual, somatosensory, gustatory, olfactory)
Hippocampus and entorhinal cortex
Polymodal association cortex
ITC
LA
B
*
CE
Viscero-sensory cortex
Sensory brainstem (pain, viscera)
Prefrontal cortex (medial)
HPA axis (hypothalamus)
Freezing (central gray)
Modulatory systems: NE, 5HT, DA, ACh (cell groups in forebrain and/or brainstem)
Sympathetic and parasympathetic NS (hypothalamus, brainstem)
ITC
LA
B
CE
Prefrontal cortex (regulation)
Ventral striatum (instrumental behaviors)
Polymodal association cortex (cognition)

FIGURE 2. Amygdala Inputs and Outputs. Inputs to some specific amygdala nuclei. Asterisk (*) denotes species difference in connectivity. (Bottom) Outputs of some specific amygdala nuclei. 5HT, serotonin; Ach, acetylcholine; B, basal nucleus; CE, central nucleus; DA, dopamine; ITC, intercalated cells; LA, lateral nucleus; NE, nor-epinephrine; NS, nervous system. Source: Rodrigues, LeDoux, and Sapolsky (2009)

Amygdala Areas: Various Stains

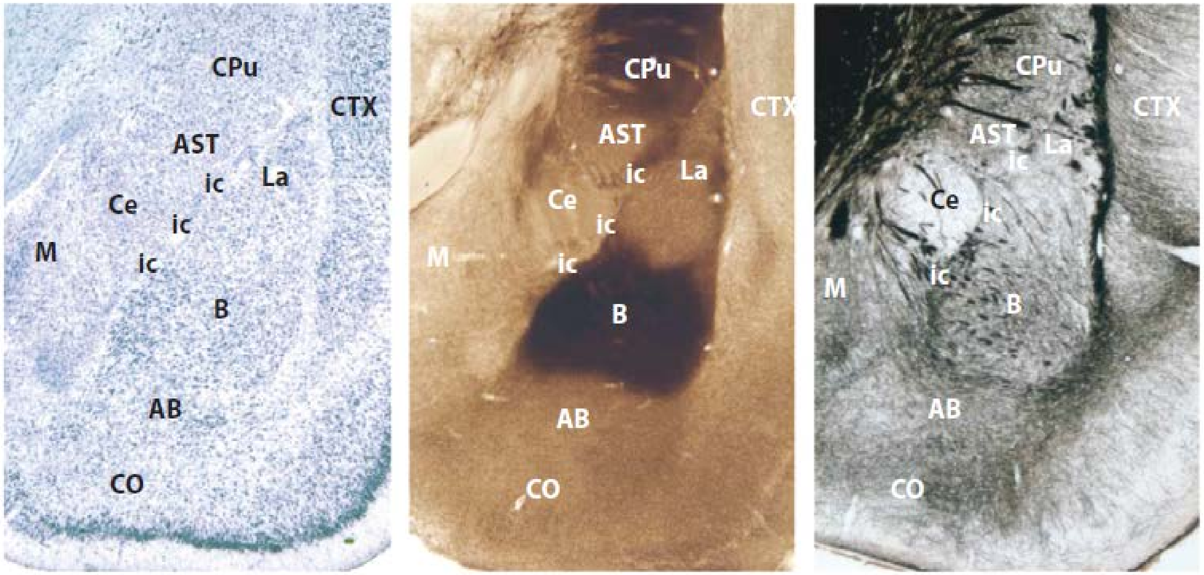


FIGURE 3. Key Areas of the Amygdala. Key areas of the amygdala, as shown in the rat brain. The same nuclei are present in primates, including humans. Different staining methods show amygdala nuclei from different perspectives. Left panel: Nissl cell body stain. Middle panel: acetylcholinesterase stain.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

Innate, Automatic, Fast, Inaccessible to Deliberation

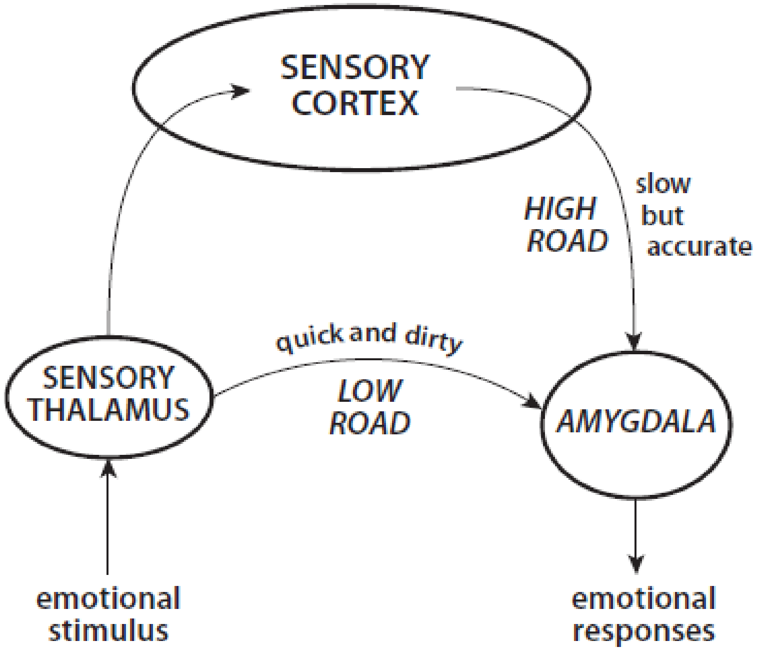


FIGURE 4. Low Road and High Road to Fear. Source: LeDoux (2002, pp. 61–63, Figure 5.7)

Also equipped with an associative machinery. “Neurons that fire together wire together.” Donald Hebb (1949)

Associative Fear Conditioning: Acquisition Phase

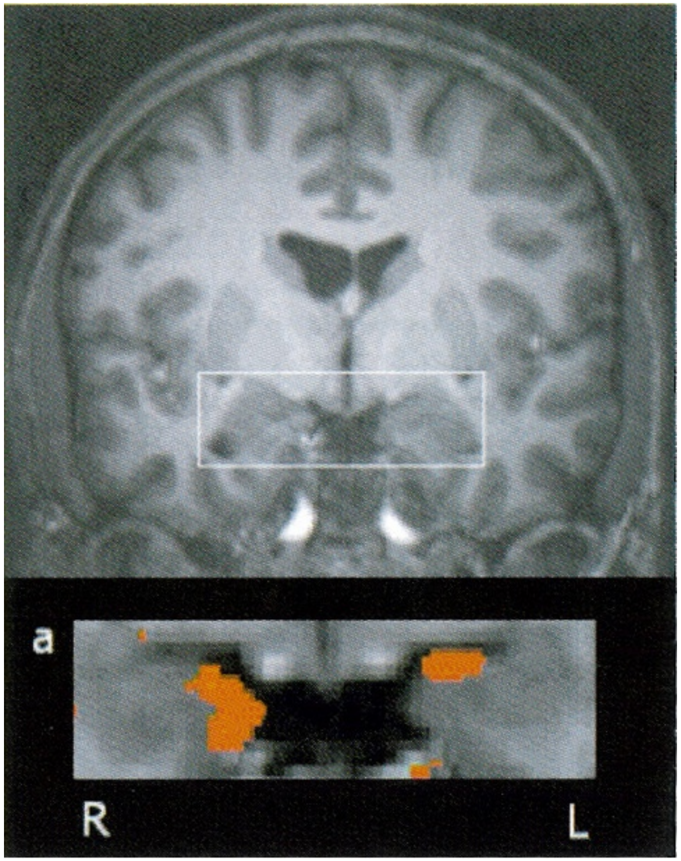
US: Shock cuff

UR: Amygdala activation

CS: Blue Light (neutral)

CS-US Pairing Trials

Light…Shock
Light…Shock
Light…Shock
Light alone …………→



Simple Elegant Model of Associative Learning
Rescorla-Wagner Model (1972)

$$v_{t+1} - v_t = \alpha\beta(\lambda - v_t)$$

Learning rates (α, β) : Surprise and Salience

Associative gain requires Surprise and Salience

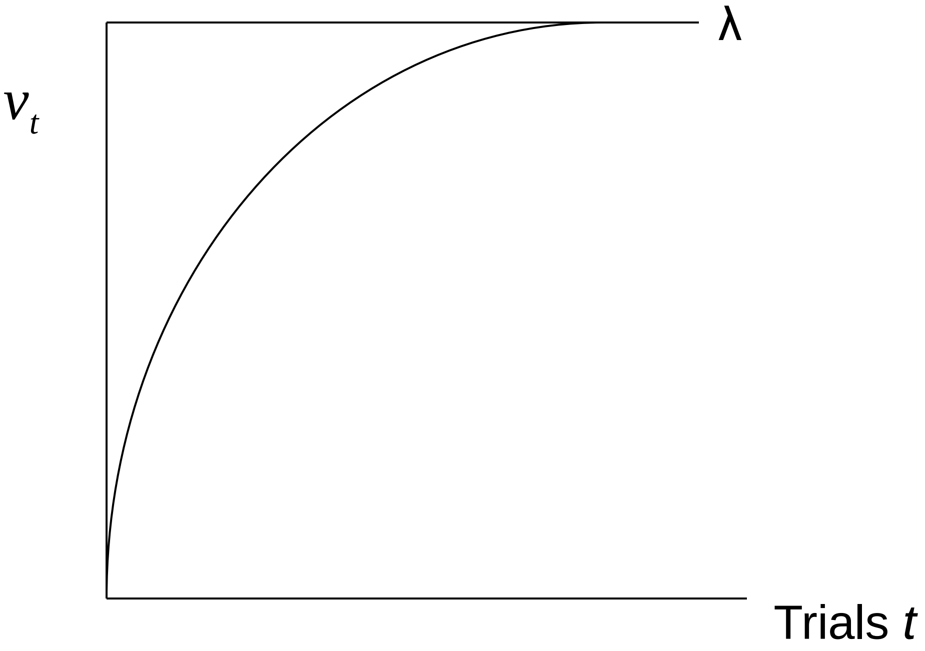
λ (typically 1) is max associative strength, asymptote.

$$v_{monday} = v_{sunday} + \alpha\beta(1 - v_{sunday})$$

$$If\ v_0 = 0, \text{ and } \alpha\beta = .01, \text{ then } v_1 = .01$$

$$Then\ v_2 = .01 + .01(1 - .01) = .0199$$

Acquisition Curve Under Classical RW



Important

NOT modeling brain regions or tissue.

Modeling an innate associative performance conferred by the neural architecture and

Explained by the underlying neuroscience.

The neuroscience ‘licenses’ the modeling and its interpretation.

…can now explain what Hume observed.

Hume’s Association of Ideas

“. . . after the constant conjunction of two objects . . . we are determined by *custom* alone to expect the one from the appearance of the other . . . It is not by reasoning, moreover, that we form the connection. All these operations are a species of natural instinct, which no reasoning or process of the thought and understanding is able either to produce or to prevent” (Section V, Part I).

Very important: Nonconscious and inaccessible to ratiocination.

Not Merely Bounded Rationality

…Social Science Without Choices

Perils of Fitness

“Survival circuits” (LeDoux 2012) conserved across vertebrate evolution.

Epstein (2013) “Pleistocene man never encountered a BMW, but *we* freeze when a car whips around the corner at us, just as *he* froze when huge animals charged suddenly from the tall brush. *We are harnessing the same innate fear-acquisition capacity—the same innate neurochemical computing architecture.* Miraculously, synaptic plasticity permits us to adapt the evolved machinery to encode novel threats.”

Invaluable but very dangerous…double-edged

Surprise + Salience → Strong Conditioning

CS	US	UR/CR
Light	Shock	Fear
Vietnamese Face	Ambush	My Lai Massacre
Arab Face	9/11	Koran as ISIS Ops-manual
Japanese Face	Pearl Harbor	Internment

Surprise + Salience → Strong Conditioning

CS	US	UR/CR
Light	Shock	Fear
Doctor	Tuskeegee	Distrust
MMR Vaccine	Autism	Vaccine refusal
Financial asset	Sudden devaluation	Panic

Associations also Over-General and Persistent

Should stay afraid of hippos.

Affect can remain above the threshold long after actual stimulus has stopped.

Stimulus stopped at t. Extinction may be far off.

Acquisition and Extinction not Symmetrical

Full Affective Trajectory

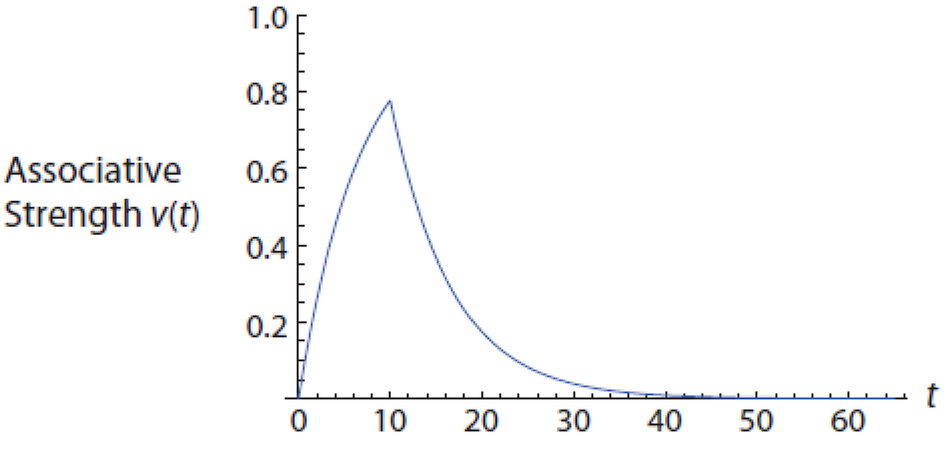
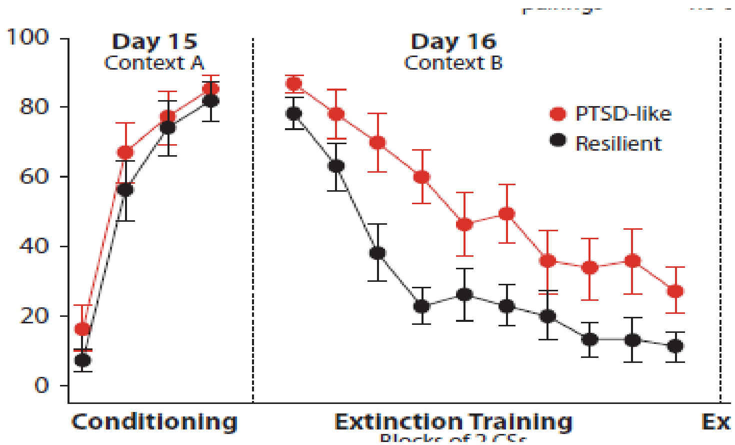


FIGURE 8. Acquisition and Extinction



Rats, predatory threat

With *t** the time at which trials cease, the full solution is then

$$v(t) = \lambda(H(t^* - t)(1 - e^{-\alpha\beta t}) + (1 - e^{-\alpha\beta t^*})H(t - t^*)e^{-\alpha\beta(t - t^*)})$$

We do not fear *what* the rat fears, but we fear *how* the rat fears (Epstein, 2013).

One Component is Fear, But Saying Contagious (ω)

Epstein, Parker, Cummings, Hammond (2008), “Coupled Contagion Dynamics of Fear and Disease.” PLoS_ONE

Generalized and fit to twitter Data by Smith and Broniatowsky (2016)

Dandy…but any neural basis/license?

One Component is Fear, But Saying Contagious (ω)

Epstein, Parker, Cummings, Hammond (2008), "Coupled Contagion Dynamics of Fear and Disease." PLoS_ONE

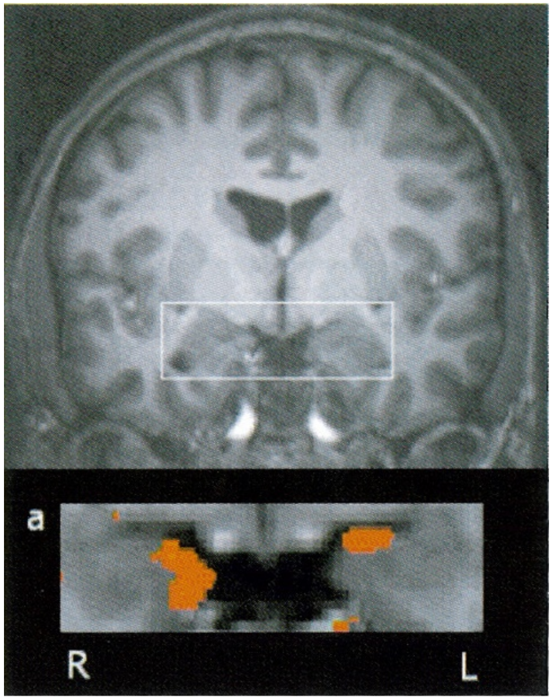
Generalized and fit to twitter Data by Smith and Broniatowsky (2016)

Dandy…but any neural basis/license?

Yes!

Observational Fear Conditioning*

Shown earlier : Fear-Conditioned human amygdala fMRI



US: Shock cuff

UR: Amygdala activation

CS: Blue Light (neutral)

CS-US PairingTrials

- Light…Shock
- Light…Shock
- Light…Shock
- Light alone ………….→

*Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social cognitive and affective neuroscience*, *2*(1), 3–11.

Is Fear Contagious?

Top Panel (a), fMRI of subject above

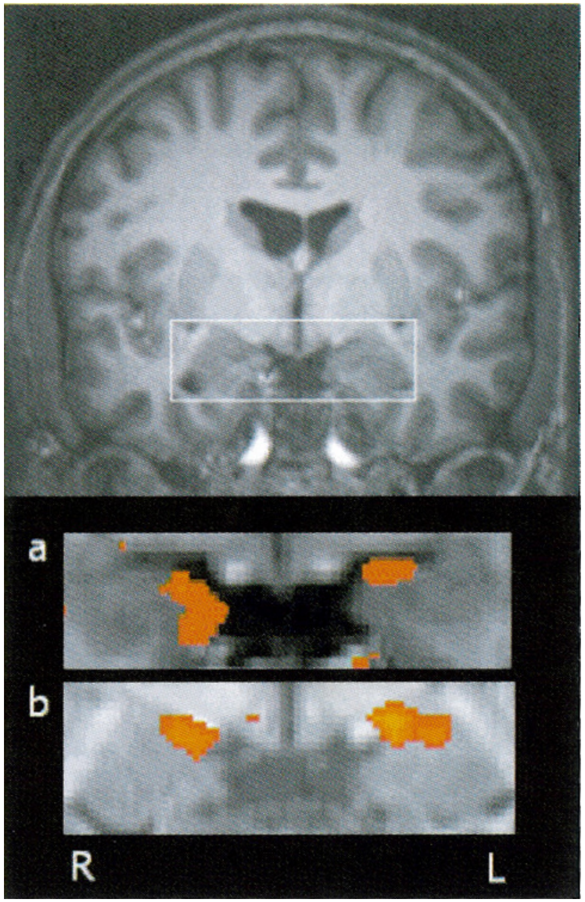
True Subject: Bottom Panel (b), fMRI of *observer.*

Watches the blue-shock pairings

- Then is shown blue light alone…
- Same fMRI as if conditioned!

Advantage clear

- Fear the fire by watching you get burned
- Downside is also clear: rapid nonconscious transmission of baseless fear.



Ingredient 1: Emotion

Introduce a generalized version of the classic (1972) Rescorla-Wagner model and emotional contagion through weights (ultimately functions of affect, so not a state variable).

Reason may be "a slave to the passions," but reasoning sometimes happens…

Typically we have incomplete and imperfect information

Make systematically erroneous appraisals of it.

Robustly documented errors:

- Representativeness heuristic
  - My local sample represents population
- Base rate neglect
  - P(+|sick) radically different than P(sick|+)

Agent_Zero (local relative frequentist) exhibits both.

To Make Matters Worse…

Agents driven by strong (unconscious) emotions (like fear), doing bad statistics on incomplete and biased data, also *influence* one another.

Conformist pressures can then produce widespread convergence on counter-productive behavior.

Conformity effects are documented in many spheres (since Asch 1958).

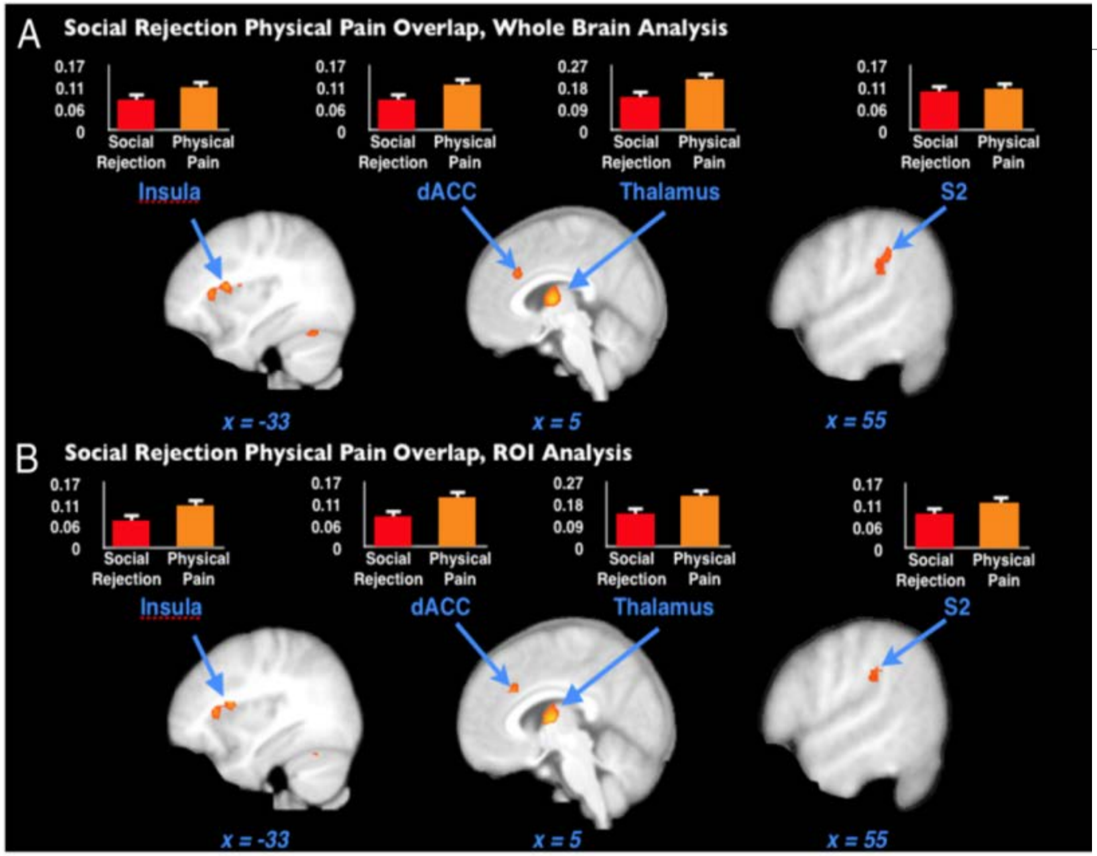
Again, a neural basis?

Yes: Nonconformity Hurts!

Kross et al (*PNAS* 2011) "…when rejection is powerfully elicited…areas that support the sensory components of *physical pain* (secondary somatosensory cortex; dorsal posterior insula) become active."

Illustrated in fMRIs below.

Neural Drivers of Conformity



Neural overlap between social rejection and physical pain.

Bar graph: no statistically significant difference between (βs of) rejection and physical pain. **Positive predictive value = 88%.**

Source: Kross, E., Berman, M. G., Mischel, W., Smith, E. E., & Wager, T. D. (2011). Social rejection shares somatosensory representations with physical pain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6270–6275.

Conform Because Rejection Hurts.

As they write, “These results give new meaning to the idea that rejection ‘hurts’…rejection and physical pain are similar not only in that they are distressing—they share a common somatosensory representation as well.”

Ingredient 3: Network Weights

Agents experience a weighted sum of the affective and deliberative states of others

Weights are actually endogenous in model--affective homophily *generates* networks…more after some runs

Given these components…

Logic of the Model:

- Disposition
- Threshold
- Action

Agent-Based Model Runs

Three Blue *Agent_Zeros* Occupy Yellow Country

V(t): emotion

- Yellow passive, Orange stimulus trial
- Level of conditioned aversive affect toward indigenous population (Generalized RW)

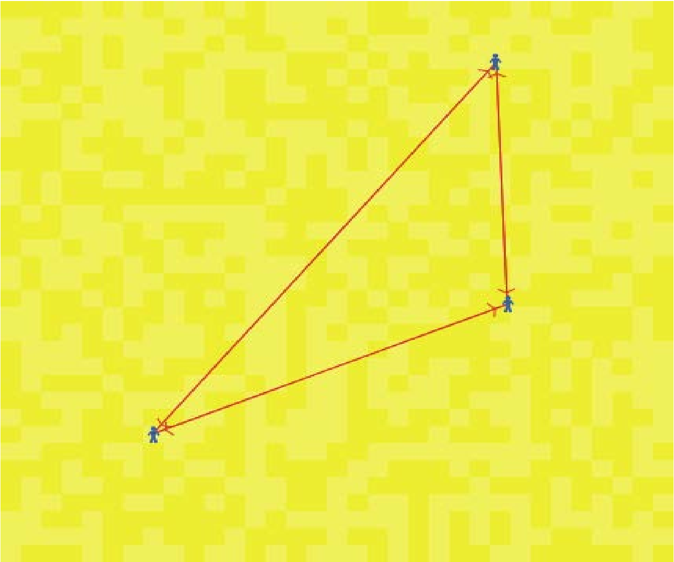
P(t): deliberation

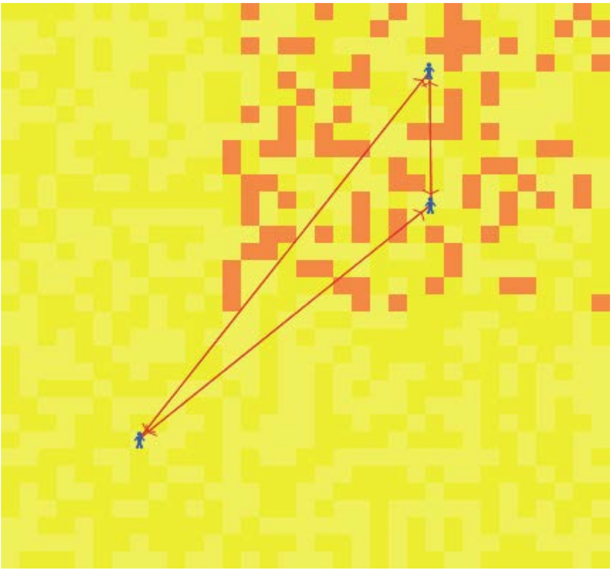
- Estimated probability that a random indigenous is in fact an enemy (e.g., random muslim is a terrorist; random vaccine is dangerous)
- Moving Average of Orange Relative Frequency

Social:

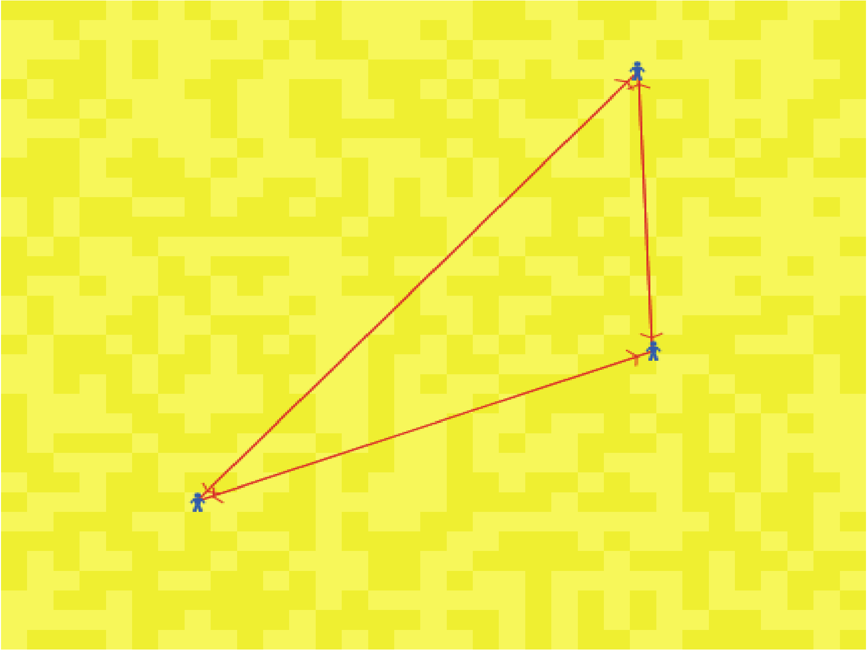
$$D_i^{tot}(t) = D_i^{solo}(t) + \sum_{j \neq i} \omega_{ji} D_j^{solo}(t)$$

Landscape and Trials:
Agent 0 Fixed and Mobile Rovers





Agent_Zero Joins Without Direct Stimulus



V=P=0, since no stimulus within sensory radius.

La Condition Humaine

Why? $D_{total} > \tau > D_{solo}$

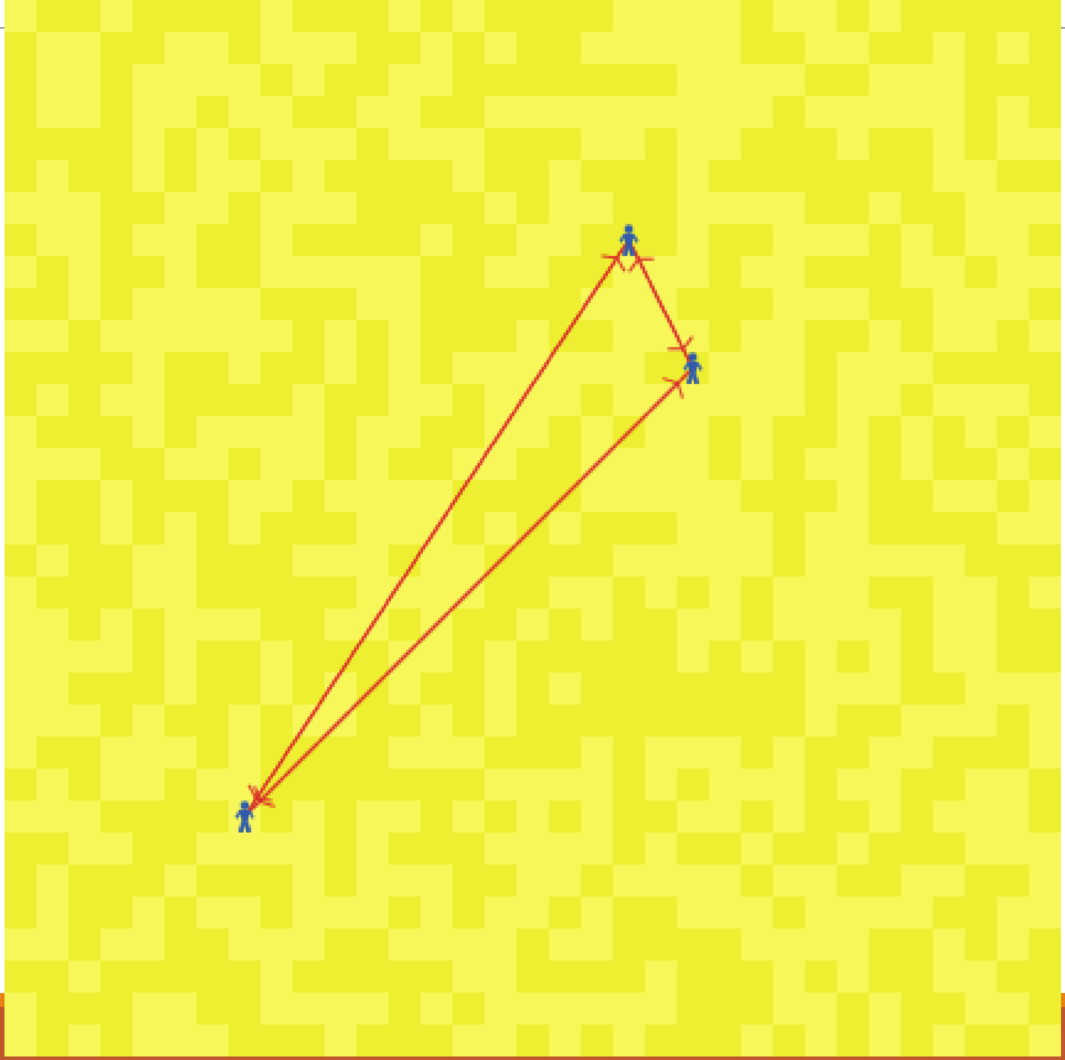
You take action in group (because $D_{total} > \tau$) that you would not take alone (because $D_{solo} < \tau$).

Indeed, you may be the only agent with this ordering.

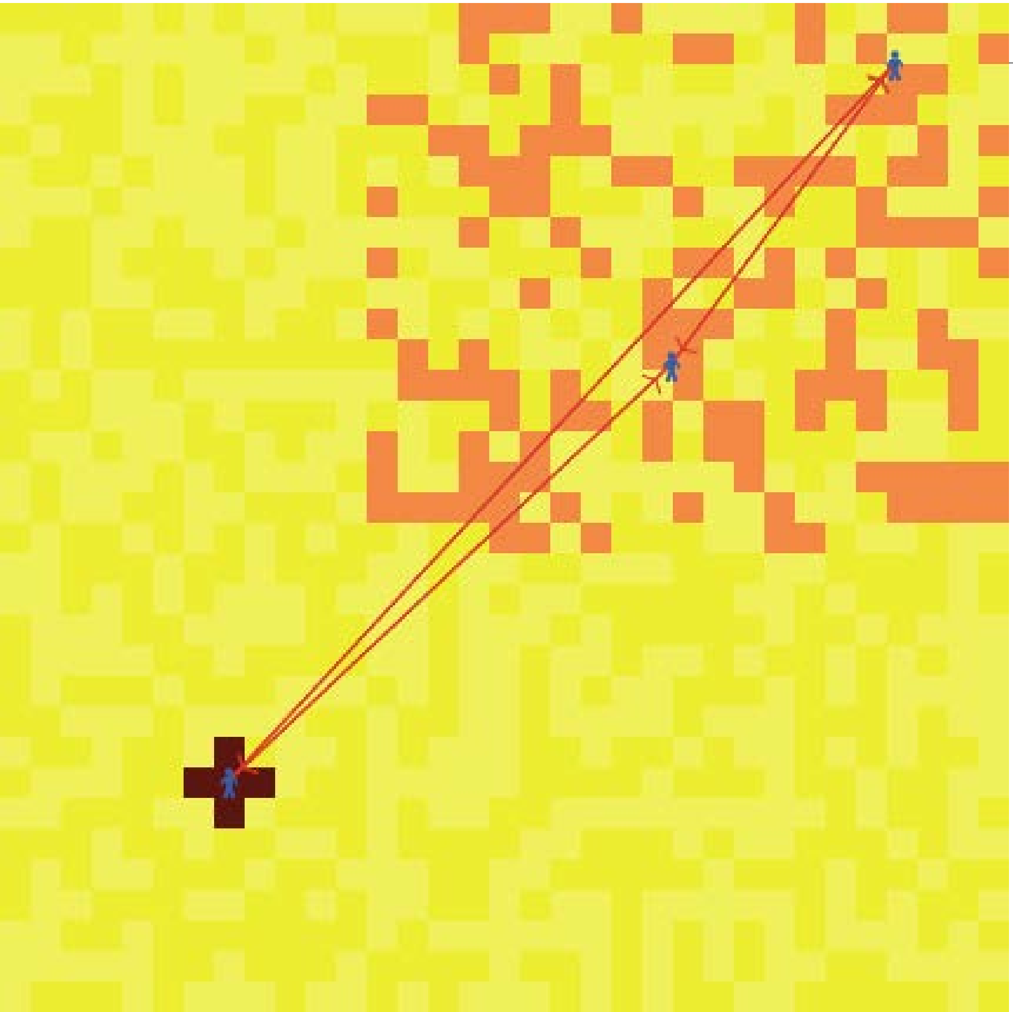
Despite being negatively disposed* you act first!

* $D_{solo} - \tau < 0$

Central Computational Parable: *Agent_Zero* Initiates Action



Core Parable: Agent_Zero Goes First Without Stimulus



Leadership or Susceptibility?

Not behavioral imitation.

- If 1st, nobody to imitate
- Deeper: Observable behavior not part of the Action formula. Just disposition

Leader, or just most susceptible to D-contagion?

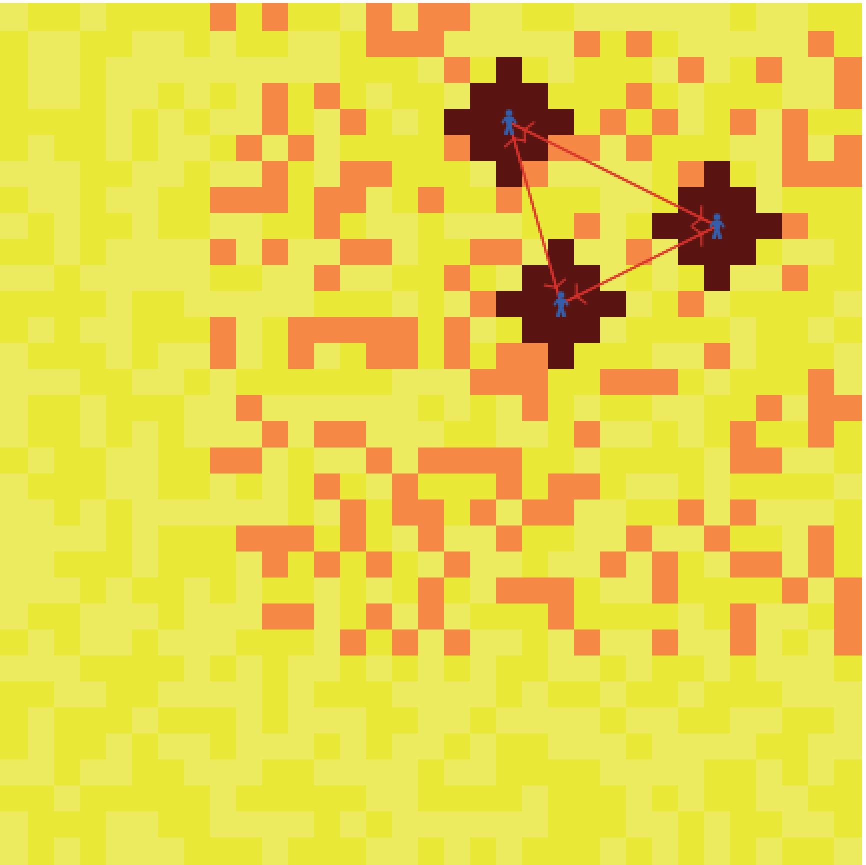
Tolstoy's answer: 'A king is history's slave, performing for the swarm life.' (War and Peace, 1896)

Unsettling Picture

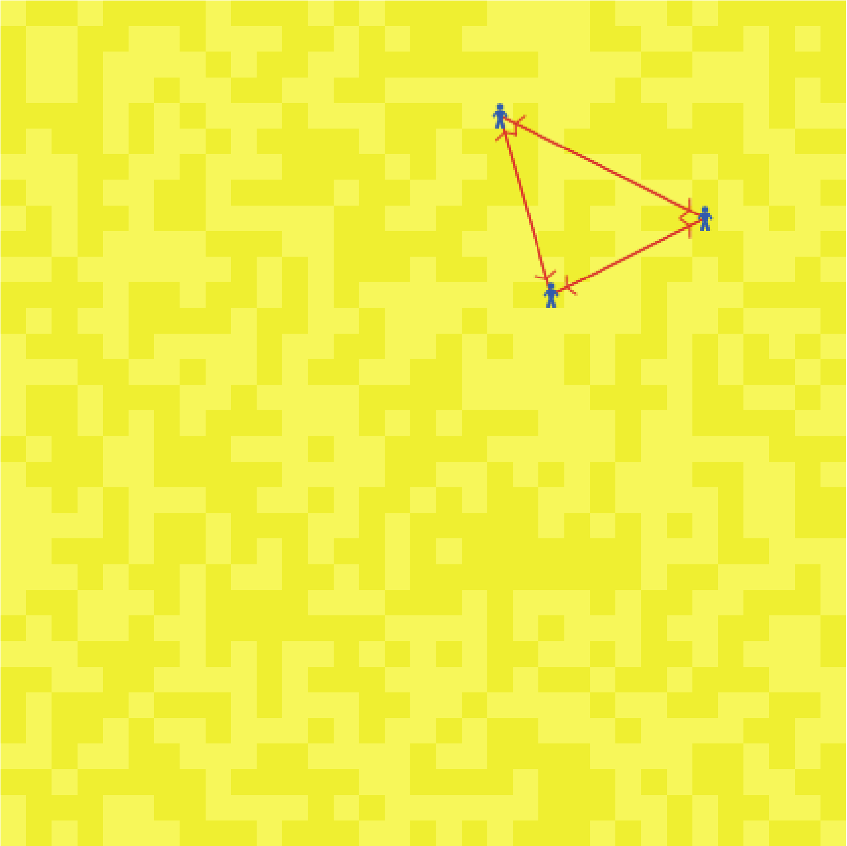
“The overall picture of Homo sapiens reflected in these interpretations of *Agent_Zero* is unsettling: Here we have a creature evolved (that is, selected) for high susceptibility to unconscious fear conditioning. Fear (conscious or otherwise) can be acquired rapidly through direct exposure or indirectly, through fearful others. This primal emotion is moderated by a more recently evolved deliberative module, which, at best, operates suboptimally on incomplete data, and whose risk appraisals are normally biased further by affect itself. Both affective and cognitive modules, moreover, are powerfully influenced by the dispositions of similar—equally limited and unconsciously driven—agents. Is it any wonder that collectivities of interacting agents of this type—the *Agent_Zero* type—can exhibit mass violence, dysfunctional health behaviors, and financial panic?”

Fight vs. Flight

Fight



Flight (Syrian Bombing and Refugees?)



Networks Implicated

How do network weights change?

Why do networks happen?

Endogenous Network Weights Affective Homophily

Affective homophily. Affects changing. So try: $\left|v_i(t)-v_j(t)\right|$

Problem: equals zero when identical; want 1.0 when equal.

OK, so as homophily, use: $1-\left|v_i(t)-v_j(t)\right|$

Problematic as a weight: nudniks (v=0) same strength as crusaders (v=1). So, scale by total strength $\omega_{ji}(t)=[v_i(t)+v_j(t)](1-|v_i(t)-v_j(t)|)$

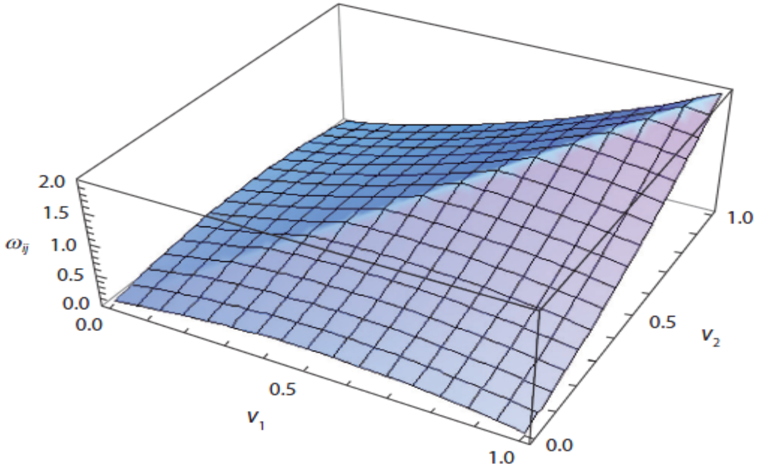


FIGURE 63. Weight Surface

Grow The Arab Spring
Case 1: No Communication

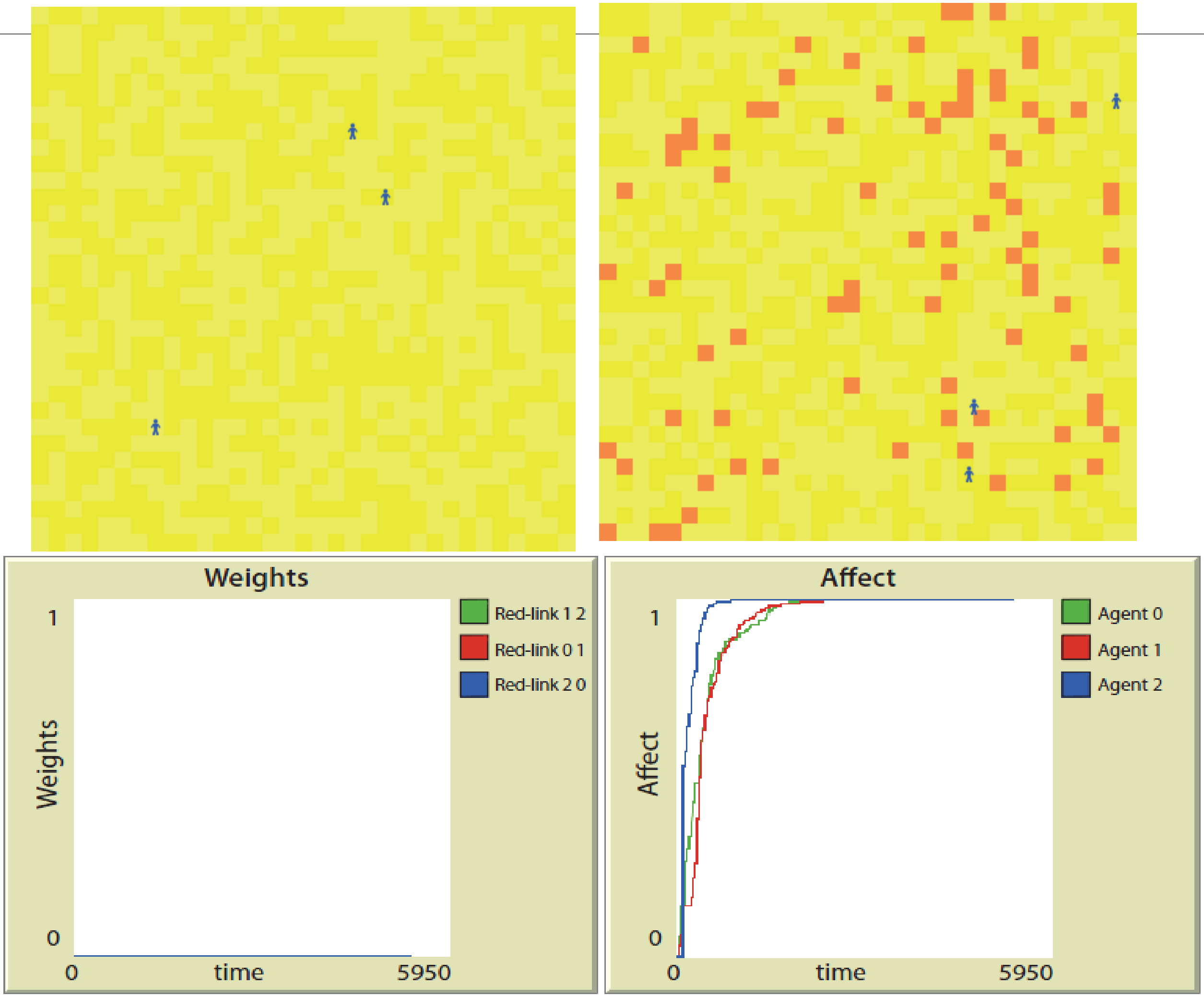
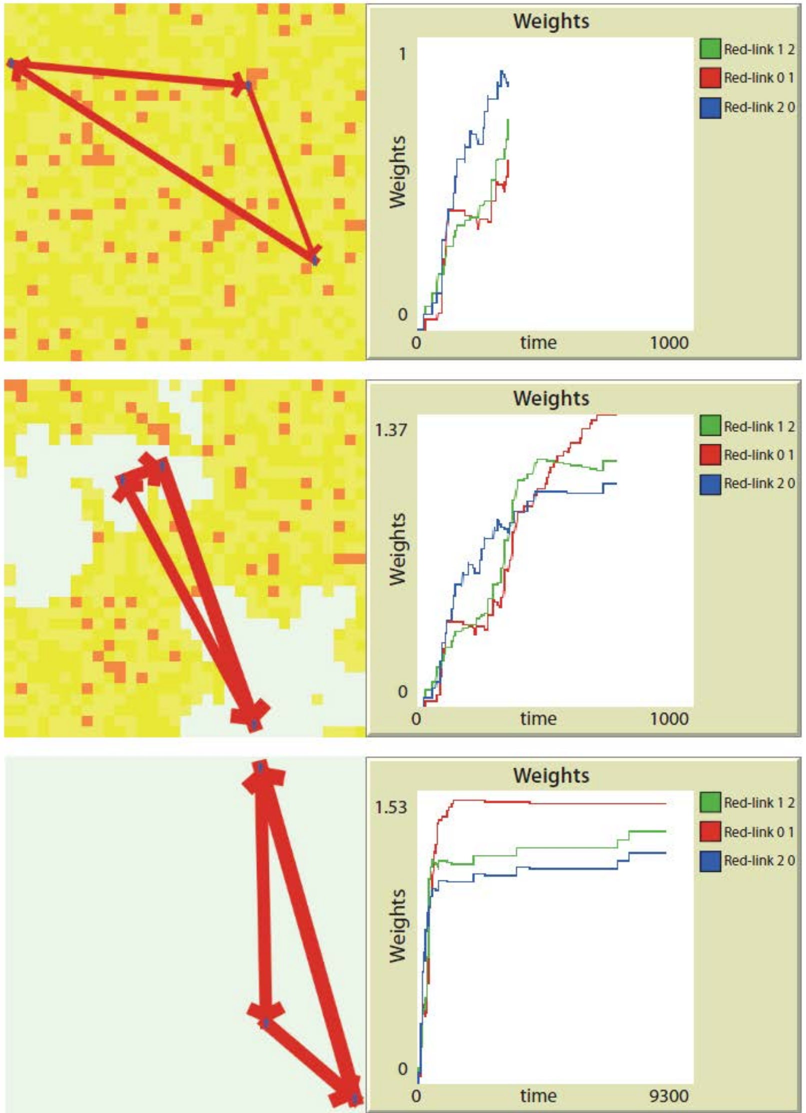


FIGURE 73. Weights and Affect with No Social Media

Arab Spring (Jasmine Revolutions)
Case 2: Communication/Dispositional Amplification



Revolt of the Swarm

Leaderless Revolutions

- No Mao, Lenin

More like an immune response than a top-down organized assault.

Final Run: Universal Self-Batrayal…

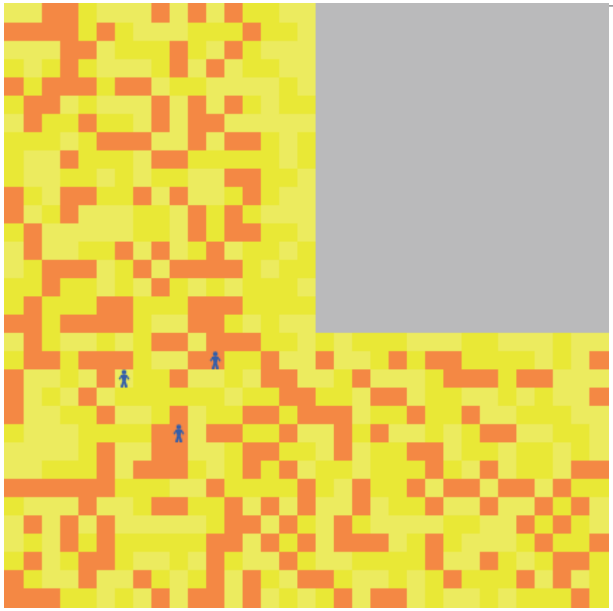
Jury Dynamics:

Pre-Trial: General landscape of stimuli about OJ’s guilt. Initial dispositions to convict are formed. Jurors strangers. All weights off.

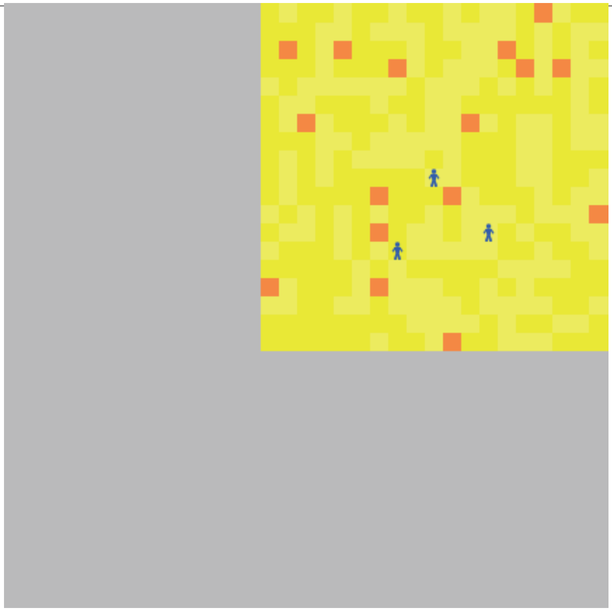
Trial: Competing stimuli sets (Prosecution and Defense). Dispositions are updated. Jurors do not communicate. Weights still off.

Sequestration: Now homophily dynamics and network effects operate strongly.

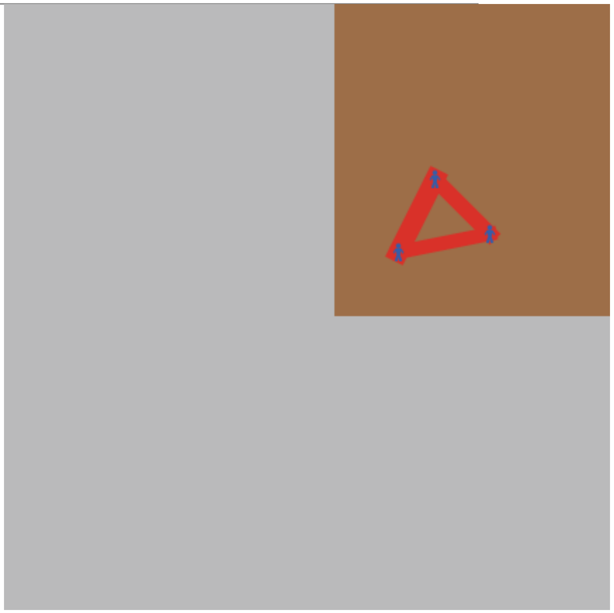
Three Phased Trial.
Pre-trial, Courtroom, Jury Phases



Pre-trial: S1>0, W=0

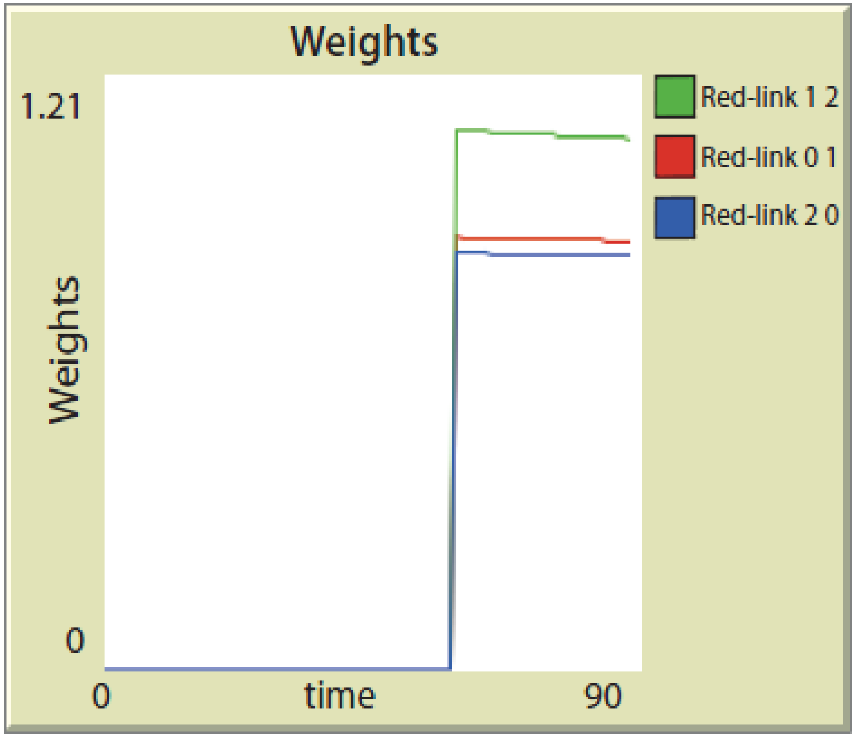


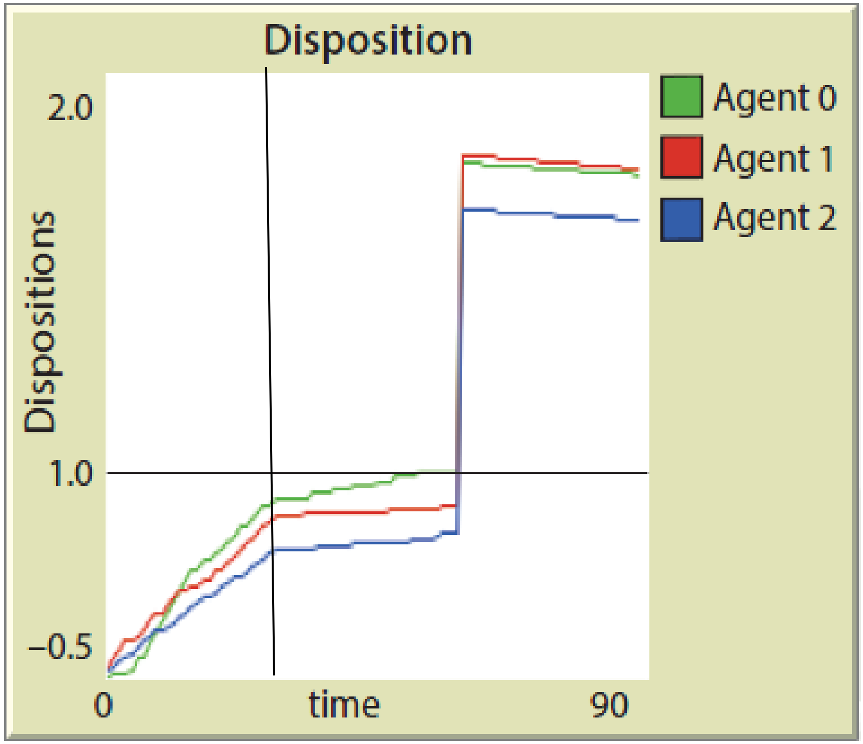
Courtroom: S2>0, W=0



Jury Phase: S=0, W>0

Weights Jump in Jury Chamber. Drive Dispositions to Convict





Now Universal Self-Betrayal

$$D^{total} > \tau > D^{solo}$$

No *jurors would have convicted before the jury phase, but they are unanimous in rendering a guilty verdict, having interacted directly*.

Networks are Two-sided: no-one alone would join the Arab Spring. But maybe no-one alone would join the ethnic cleansing either!

Immune Response

A bullet stops the bear. But it does not stop a leaderless swarm of bees.

How does one enable an immune response from the embedding society?

If they are Agent_Zeros, and the act is to condemn and resist ISIS, how does one reduce their action threshold? Connect them and magnify their weights?

Bottom-up social science might be informative.

Further Extensions in….

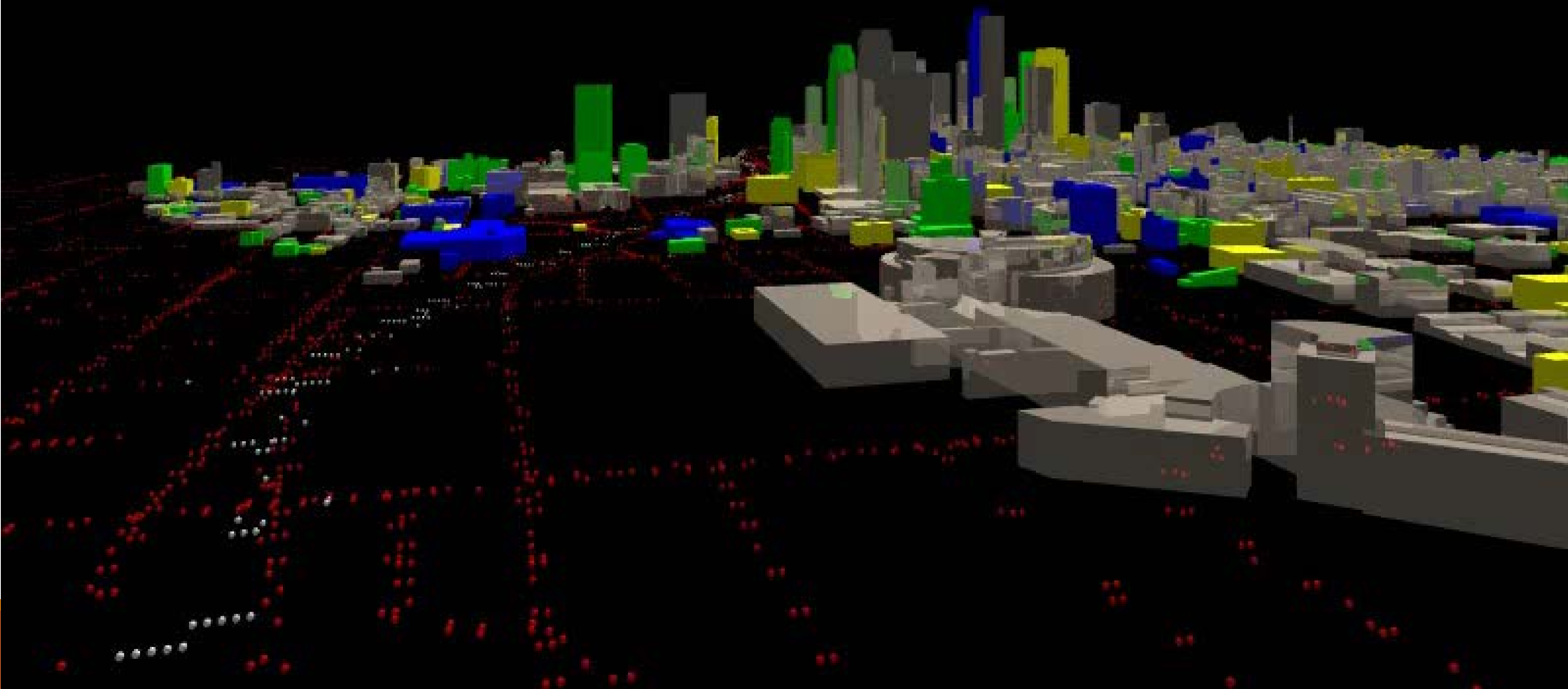
Epstein, JM and Chelen J , “Advancing Agent Zero” in Kirman A and Wilson DS, eds. *Complexity and Evolution: Toward a New Synthesis for Economics* (MIT Press, 2016)

Main Agenda

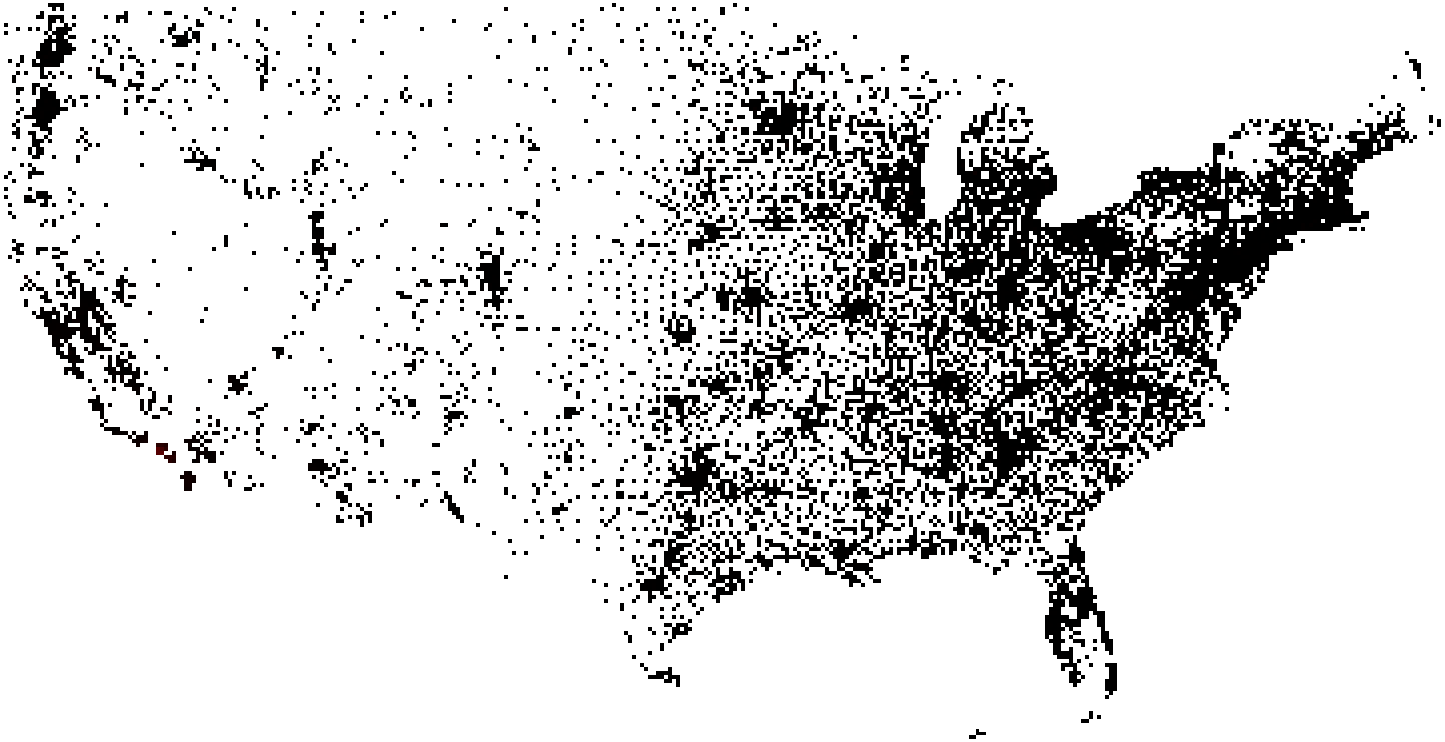
Deepen: Improve the components neuro-scientifically

Scale-Up: Populate large models

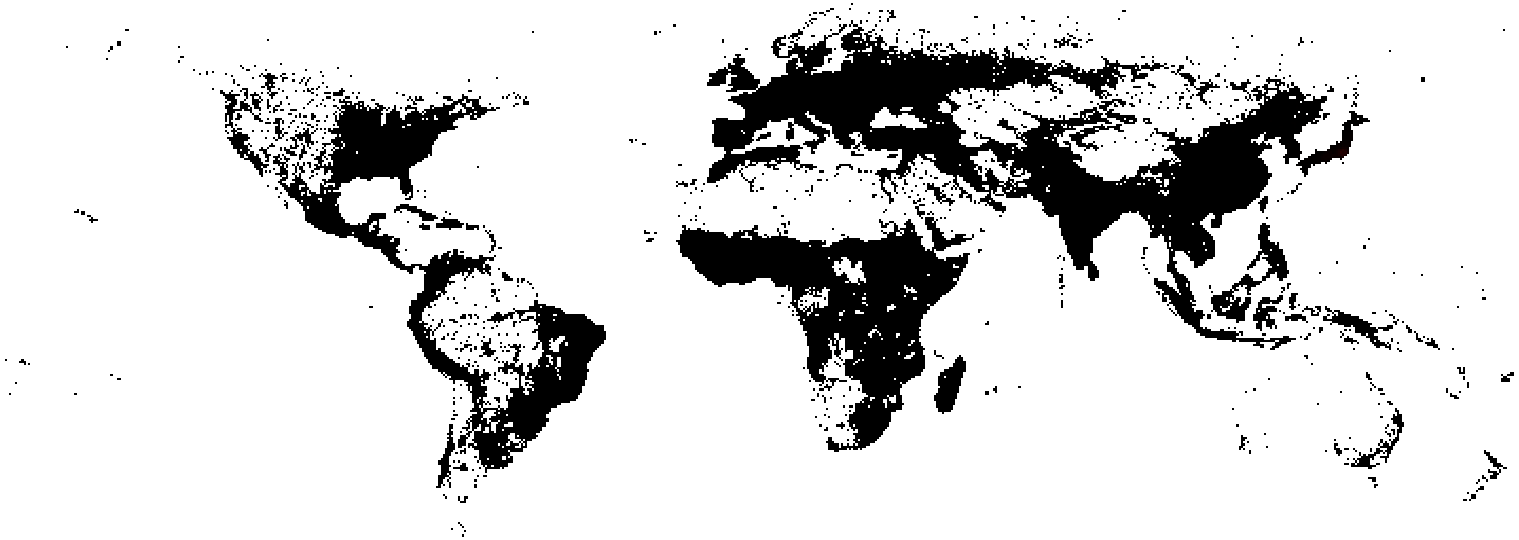
- LA Plume-Agent Hybrid
- US National
- Global



A US Run (300m agents)



A Global Run (6.5b agents)



Agent_Zero

A neurocognitively grounded agent capable of generating a wide range of important social phenomena.

A mathematically explicit functioning alternative to the rational actor.

Foundation for Generative Social Science.

Thank You!