

The Invisible Academy: nonlinear effects of linear learning

Mark Liberman
University of Pennsylvania

Outline

1. An origin myth: naming without Adam
a computer-assisted thought experiment
2. A little old-time learning theory
linear operator models of probability learning
and expected rate learning
3. Generalization:
Stochastic belief +
categorical perception +
social interaction ⇒
emergence of random shared beliefs
("culture" ?)

The problem of vocabulary consensus

- 10K-100K arbitrary pronunciations
- How is consensus established and maintained?

Genesis 2:19-20

And out of the ground the Lord God formed every beast of the field, and every fowl of the air; and brought them unto Adam to see what he would call them: and whatsoever Adam called every living creature, that was the name thereof. And Adam gave names to the cattle, and to the fowl of the air, and to every beast of the field...

Possible solutions

- Initial naming authority? *Implausible...*
 - Adam
 - L'académie paleolithique
- Natural names? *False to fact...*
 - evolved repertoire (e.g. animal alarm calls)
 - “ding-dong”
- ????
- **Emergent structure?**
 - begin with computer exploration of toy “agent-based” models
 - a thought experiment to explore the consequences of minimal, plausible assumptions
 - an interesting (?) idealization, not a realistic model!

Agent-based modeling

- AKA “individual-based modeling”

Ensembles of parameterized entities ("agents") interact in algorithmically-defined ways.

Individual interactions depend (stochastically) on the current parameters of the agents involved; these parameters are in turn modified (stochastically) by the outcome of the interaction.

Key ideas of ABM

- Complex structure emerges from the interaction of simple agents
 - Agents' algorithms evolve in a context they create collectively
 - Thus behavior is like organic form
- BUT
- ABM is a form of programming,
 - so just solving a problem via ABM has no scientific interest
 - We must prove a general property of some wide class of models
(or explain the detailed facts of a particular case)
 - Paradigmatic example of general explanation:
Axelrod's work on reciprocal altruism in the iterated prisoner's dilemma

Emergence of shared pronunciations

- Definition of success:
 - Social convergence
 (“people are mostly the same”)
 - Lexical differentiation
 (“words are mostly different”)
- These two properties
are required for successful communication

A simplest model

- Individual belief about word pronunciation:
vector of binary random variables
e.g. feature #1 is 1 with $p=.9$, 0 with $p=.1$
feature #2 is 1 with $p=.3$, 0 with $p=.7$
...
- (Instance of) word pronunciation: (random) binary vector
e.g. 1 0 ...
- Initial conditions: random assignment of values to beliefs of N agents
- Additive noise (models output, channel, input noise)
- Perception: assign input feature-wise to nearest binary vector
i.e. categorical perception
- Social geometry: circle of pairwise naming among N agents
- Update method: linear combination of belief and perception
belief is “leaky integration” of perceptions

Coding words as bit vectors

Morpheme template

$C_1V_1(C_2V_2)(\dots)$

Each bit codes for one feature in one position in the template,
e.g. “labiality of C_2 ”

Some 5-bit morphemes:

11111 g^wu

00000 $tæ$

01101 ga

10110 bi

C_1 labial?	1	0
C_1 dorsal?	1	0
C_1 voiced?	1	0
<i>more C_1 features ...</i>
V_1 high?	1	0
V_1 back?	1	0
<i>more V_1 features ...</i>
	$g^wu \dots$	$tæ \dots$

Belief about pronunciation as a random variable

Each pronunciation instance is an N-bit vector
(= feature vector = symbol sequence)

but belief about a morpheme's pronunciation is a
probability distribution over symbol sequences,
encoded as N independent bit-wise probabilities.

Thus [01101] encodes /ga/

but $\langle .1 \ .9 \ .9 \ .1 \ .9 \rangle$ is

[0 1 1 0 1] = ga with $p \approx .59$

[0 1 1 0 0] = gæ with $p \approx .07$

[0 1 0 0 1] = ka with $p \approx .07$

etc. ...

C1 labial?
C1 dorsal?
C1 voiced?
V1 high?
V1 back?

“lexicon”, “speaking”, “hearing”

Each agent’s “lexicon” is a matrix

- whose columns are template-linked features
 - e.g. “is the first syllable’s initial consonant labial?”
- whose rows are words
- whose entries are probabilities
 - “the 3rd word’s 2nd syllable’s vowel is back with $p=.973$ ”

MODEL 1:

To “speak” a word, an agent “throws the dice”
to chose a pronunciation (vector of 1’s and 0’s)
based on that row’s p values

Noise is added (random values like .14006 or .50183)

To “hear” a word, an agent picks the nearest vector of 1’s and 0’s
(which will eliminate the noise if it was $< .5$ for a given element)

Updating beliefs

When a word W_i is heard,
hearer “accommodates” belief about W_i
in the direction of the perception.

New belief is a linear combination
of old belief and new perception:

$$B_t = \alpha B_{t-1} + (1 - \alpha) P_t$$

Old belief = $\langle .1 \ .9 \ .9 \ .1 \ .9 \rangle$

Perception = $[1 \ 1 \ 1 \ 0 \ 1]$

New belief = $[.95*.1+.05*1 \ .95*.9+.05*1 \ . \ . \ .]$

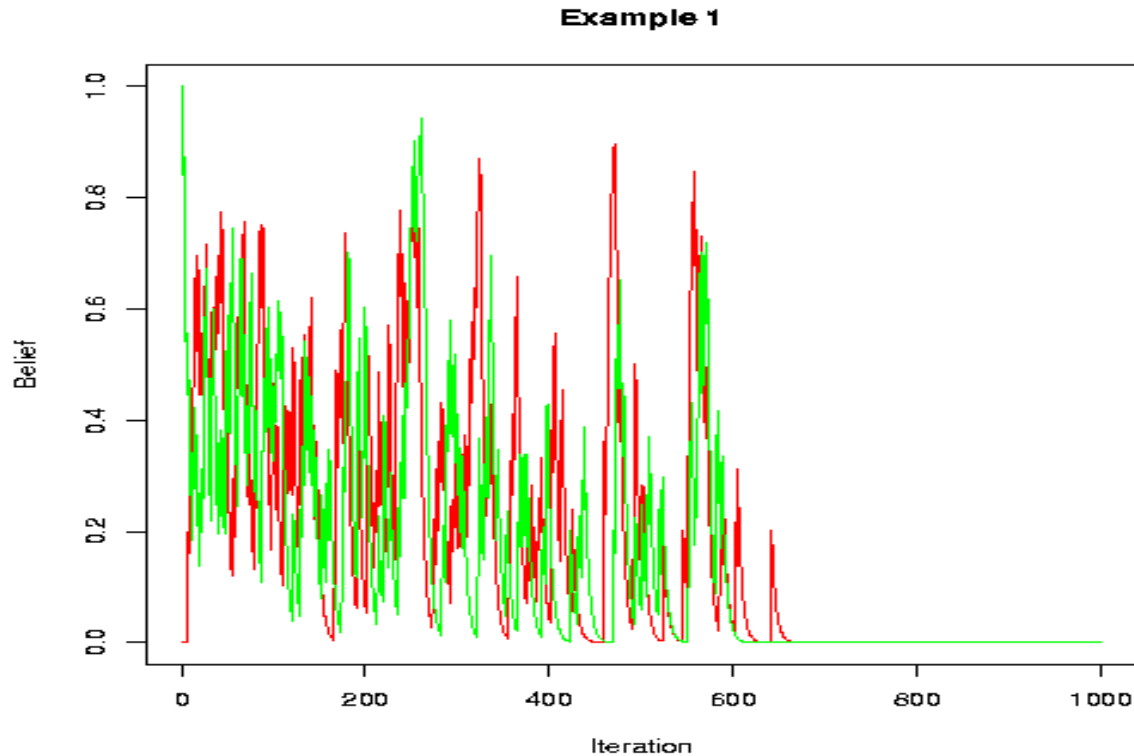
= $[.145 \ .905 \ \dots]$

Conversational geometry

- Who talks to whom when?
- How accurate is communication of reference?
- When are beliefs updated?
- Answers don't seem to be crucial
- In the experiments discussed today:
 - N (imaginary) people are arranged in a circle
 - On each iteration, each person “points and names” for her clockwise neighbor
 - Everyone changes positions randomly after each iteration
- Other geometries (grid, random connections, etc.) produce similar results
- Simultaneous learning of reference from collection of available objects (i.e. no pointing) is also possible

It works!

- Channel noise = gaussian with $\sigma = .2$
- Update constant $\alpha = .8$
- 10 people
- one bit in one word for people #1 and #4 shown:



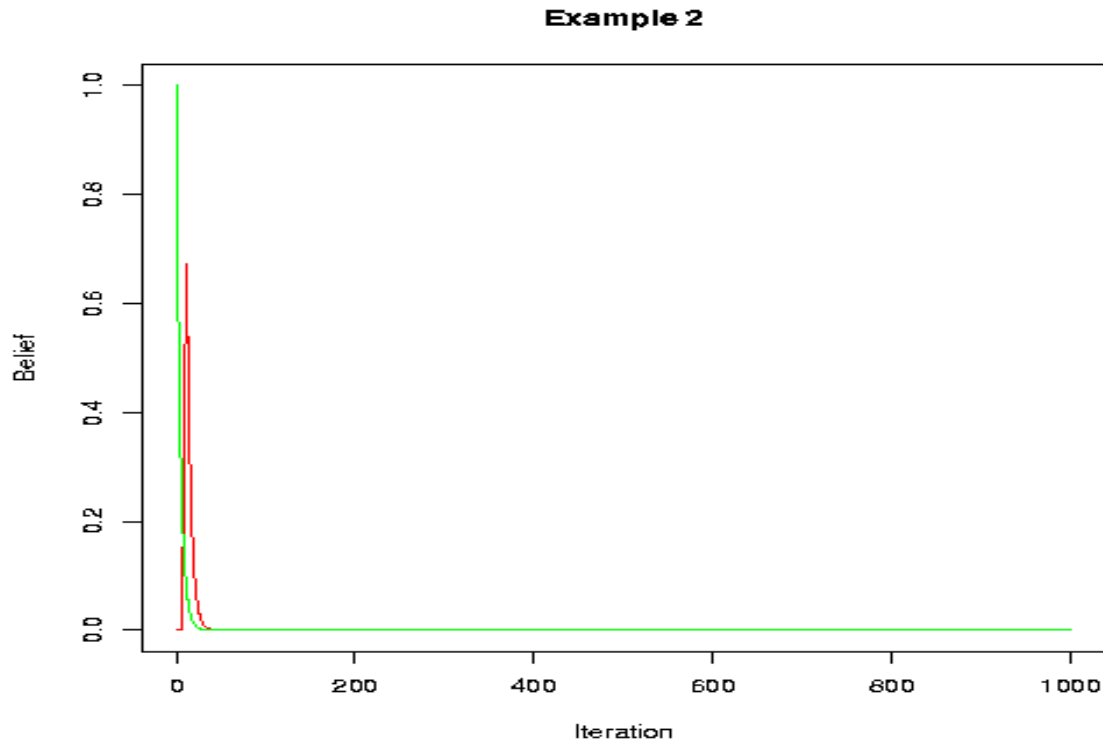
Chicago 5/24/2005

Gradient output = faster convergence

Instead of saying 1 or 0 for each feature, speakers emit real numbers (plus noise) proportional to their belief about the feature.

Perception is still categorical.

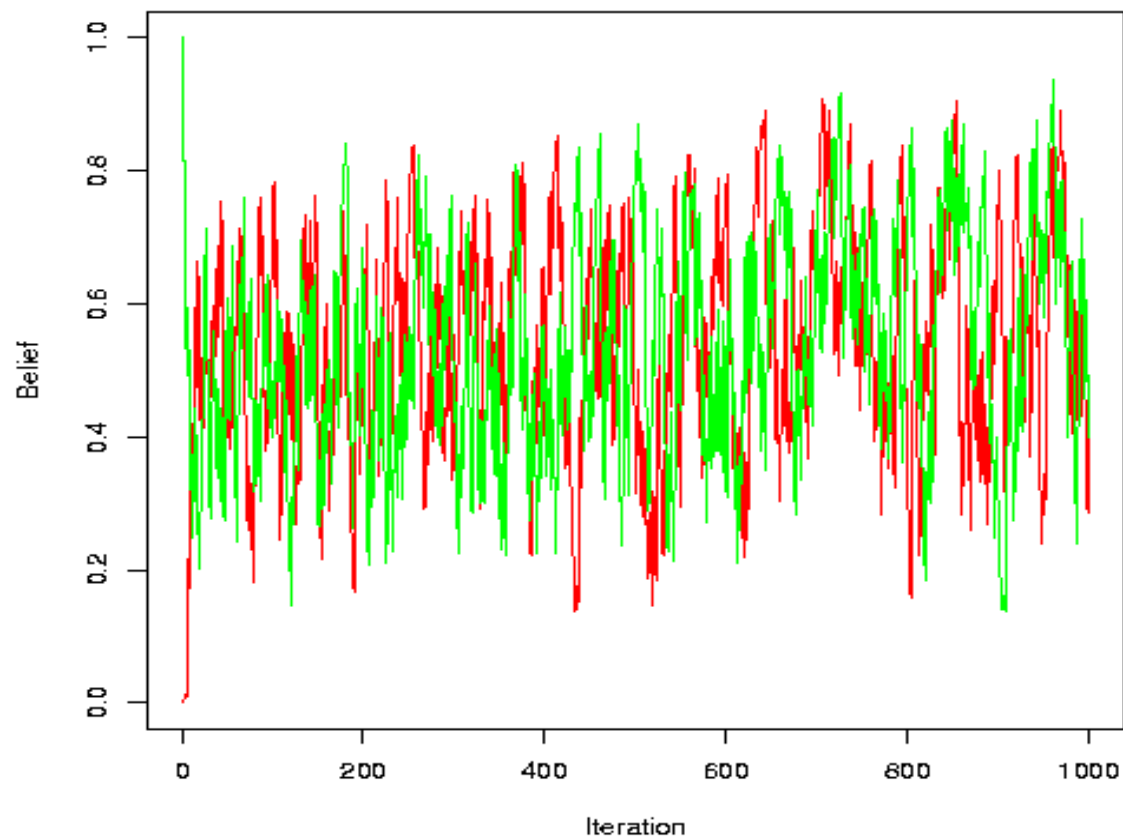
Result is faster convergence, because better information is provided about the speaker's internal state.



Gradient input = no convergence

If we make perception gradient (i.e. veridical),
then (whether or not production is categorical)
social convergence does not occur.

Example 3



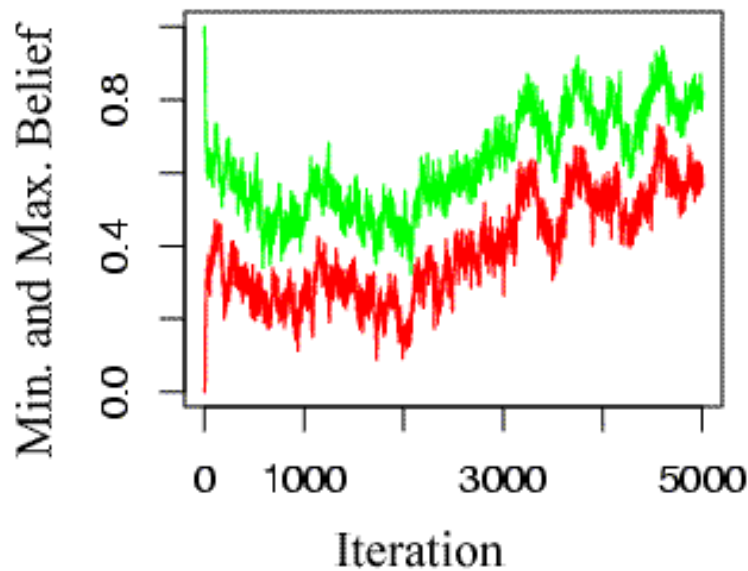
What's going on?

- Input categorization creates “attractors” that trap beliefs despite channel noise and initially random assignments
- Positive feedback creates social consensus
- Random effects generate lexical differentiation
- Assertions: to achieve social consensus with lexical differentiation, any model of this general type needs
 - stochastic (random-variable) beliefs
 - to allow learning
 - categorical perception
 - to create attractor to “trap” beliefs

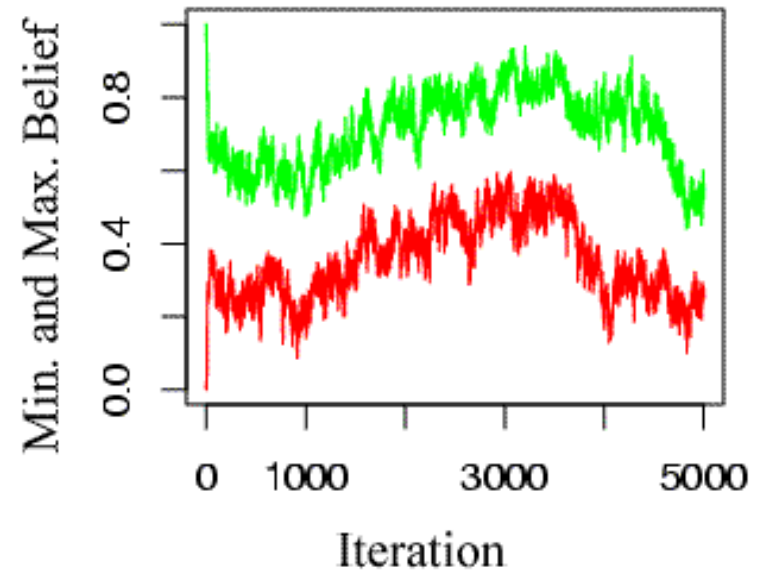
Divergence with population size

With gradient perception, it is not just that pronunciation beliefs continue a random walk over time. They also diverge increasingly at a given time, as group size increases.

20 people:



40 people:



Pronunciation differentiation

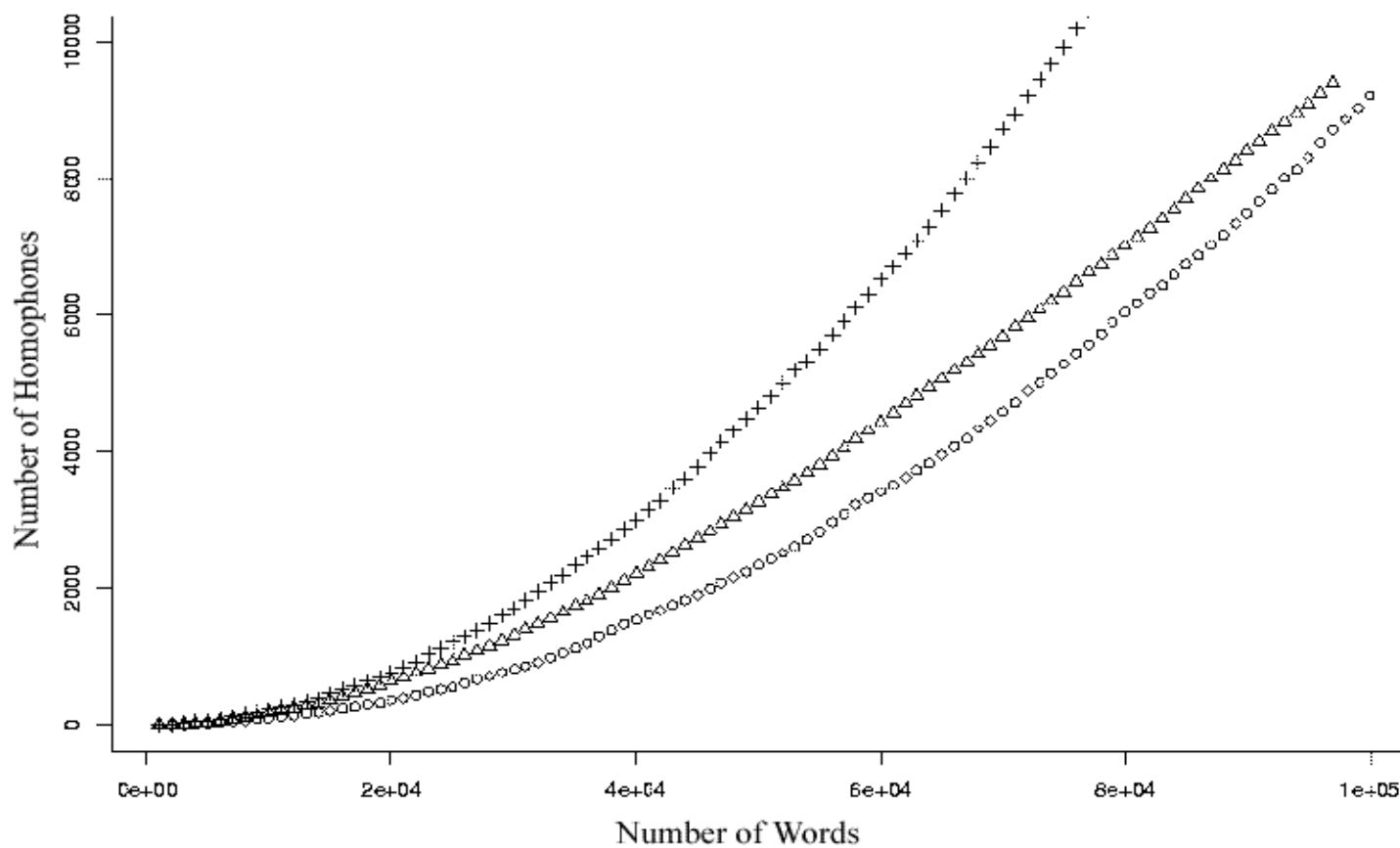
- There is nothing in this model to keep words distinct
- But words tend to fill the space randomly
(vertices of an N-dimensional hypercube)
- This is fine if the space is large enough
- Behavior is rather lifelike with word vectors of 19-20 bits

Homophony comparison

English is plotted with triangles (97K pronouncing dictionary).

Model vocabulary with 19 bits is X's.

Model vocabulary with 20 bits is O's.



But what about using a purely digital representation of belief about pronunciation? What's with these (pseudo-) probabilities? Are they actually important to "success"?

In a word, yes. To see this, let's explore a model in which belief about the pronunciation of a word is a binary vector rather than a discrete random variable -- or in more anthropomorphic terms, a string of symbols rather than a probability distribution over strings of symbols.

If we have a very regular and reliable arrangement of who speaks to whom when, then success is trivial. Adam tells Eve, Eve tells Cain, Cain tells Abel, and so on. There is a perfect chain of transmission and everyone winds up with Adam's pronunciation. The trouble is that less regular less reliable conversational patterns, or regular ones that are slightly more complicated, result in populations whose lexicons are blinking on and off like Christmas tree lights. Essentially, we wind up playing a sort of ***Game of Life***.

Consider a circular world, permuted randomly after each conversational cycle, with values updated at the end of each cycle so that each speaker copies exactly the pattern of the "previous" speaker on that cycle. Here's the first 5 iterations of a single feature value for a world of 10 speakers. Rows are conversational cycles, columns are speakers (in "canonical" order).

```
0 1 0 1 1 1 0 1 0 0
1 0 1 0 0 0 1 1 0 1
1 1 0 1 1 0 0 1 0 0
1 0 1 1 1 0 0 0 1 0
1 0 0 0 1 1 0 1 0 1
```

Here's another five iterations after 10,000 cycles -- no signs of convergence:

```
0 1 1 1 1 0 0 0 1 0
1 0 1 0 1 0 0 1 1 0
1 0 0 1 0 1 1 1 0 0
1 1 0 0 1 1 1 0 0 0
0 1 1 0 0 1 0 1 0 1
```

Even with a combination of update algorithm and conversational geometry that converges, such a system will be fragile in the face of occasional incursions of rogue pronunciations.

Conclusions of part 1

For “naming without Adam”, it’s sufficient that

- perception of pronunciation be categorical
- belief about pronunciation be stochastic

Are these are also necessary?

No! But...

Outline

1. An origin myth: naming without Adam
a computer-assisted thought experiment
2. **Some old-time learning theory**
**linear operator models of probability learning
and expected rate learning**
3. Some morals:
 - Another advantage of categorical perception
 - Grammatical beliefs as random variablesStochastic belief + categorical perception + social interaction
= emergence of coherent shared grammar

Summary of next section

- Animals (including humans) readily learn stochastic properties of their environment
- Over 100 years, several experimental paradigms have been developed and applied to explore such learning
- A simple linear model gives an excellent qualitative (and often quantitative) fit to the results from this literature
- This linear learning model is the same as the “leaky integrator” model used in our simulations
- Such models can predict either probability matching or “maximization” (i.e. emergent regularization), depending on the structure of the situation
- In *reciprocal* learning situations with discrete outcomes, this model predicts ***emergent regularization***.

Probability Learning

On each of a series of trials, the S makes a choice from ... [a] set of alternative responses, then receives a signal indicating whether the choice was correct... [E]ach response has some fixed probability of being ... indicated as correct, regardless of the S's present or past choices...

[S]imple two-choice predictive behavior ... show[s] close approximations to probability matching, with a degree of replicability quite unusual for quantitative findings in the area of human learning...

Probability matching tends to occur when the ... task and instructions are such as to lead the S simply to express his expectation on each trial... or when they emphasize the desirability of attempting to be correct on *every* trial...

“Overshooting” of the matching value tends to occur when instructions indicate... that the S is dealing with a random sequence of events... or when they emphasize the desirability of maximizing successes over blocks of trials.

-- Estes (1964)

Chicago 5/24/2005

26

Contingent correction: When the “reinforcement” is made contingent on the subject’s previous responses, the relative frequency of the two outcomes depends jointly on the contingencies set up by the experimenter and the responses produced by the subject.

Nonetheless... on the average the S will adjust to the variations in frequencies of the reinforcing events resulting from fluctuations in his response probabilities in such a way that his probability of making a given response will tend to stabilize at the unique level which permits matching of the response probability to the long-term relative frequency of the corresponding reinforcing event.

-- Estes (1964)

In brief: people learn to predict event probabilities pretty well.

Expected Rate Learning

[W]hen confronted with a choice between alternatives that have different expected rates for the occurrence of some to-be-anticipated outcome, animals, human and otherwise, proportion their choices in accord with the relative expected rates...

-- Gallistel (1990)

Maximizing vs. probability matching: a classroom experiment

A rat was trained to run a T maze with feeders at the end of each branch. On a randomly chosen 75% of the trials, the feeder in the left branch was armed; on the other 25%, the feeder in the right branch was armed. If the rat chose the branch with the armed feeder, it got a pellet of food. ... Above each feeder was a shielded light bulb, which came on when the feeder was armed. The rat could not see the bulb, but the [students in the classroom] could. They were given sheets of paper and asked to predict before each trial which light would come on.

Under these noncorrection conditions, where the rat does not experience reward at all on a given trial when it chooses incorrectly, the rat learns to choose the higher rate of payoff... [T]he strategy that maximizes success is always to choose the more frequently armed side...

The undergraduates, by contrast, almost never chose the high payoff side exclusively. In fact, as a group their percentage choice of that side was invariably within one or two points of 75 percent... They were greatly surprised to be shown... that the rat's behavior was more intelligent than their own. We did not lessen their discomfiture by telling them that if the rat chose under the same conditions they did... it too would match the relative frequencies of its... choices to the relative frequencies of the payoffs.

-- Gallistel (1990)

But from the right perspective,

Matching and maximizing
are just two words describing one outcome.

-Herrnstein and Loveland (1975)

If you don't get this, wait-- it will be explained in detail in later slides.

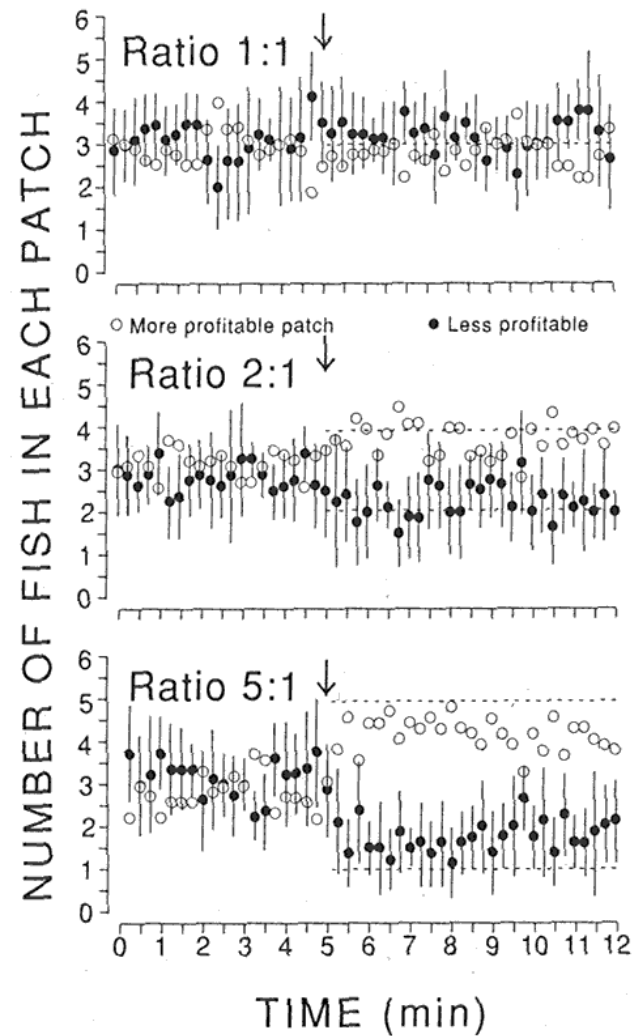
Ideal Free Distribution Theory

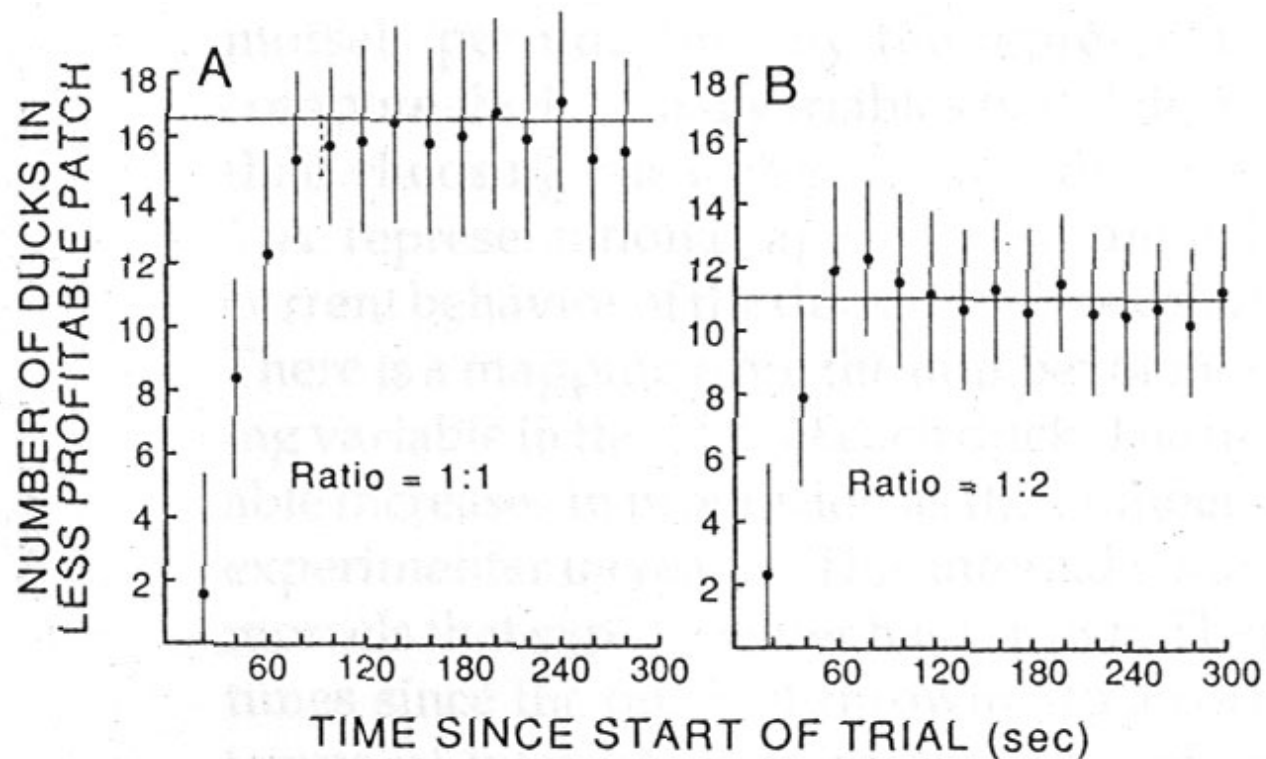
- In foraging, choices are proportioned stochastically according to estimated “patch profitability”
- Evolutionarily stable strategy
 - given competition for variably-distributed resources
 - curiously, isolated animals still employ it
- Re-interpretation of many experimental learning and conditioning paradigms
 - as estimation of “patch profitability” combined with stochastic allocation of choices in proportion
 - simple linear estimator fits most data well

Ideal Free Fish:

Mean # of fish at each of two feeding stations, for each of three feeding profitability ratios.

(From Godin & Keenleyside 1984, via Gallistel 1990)



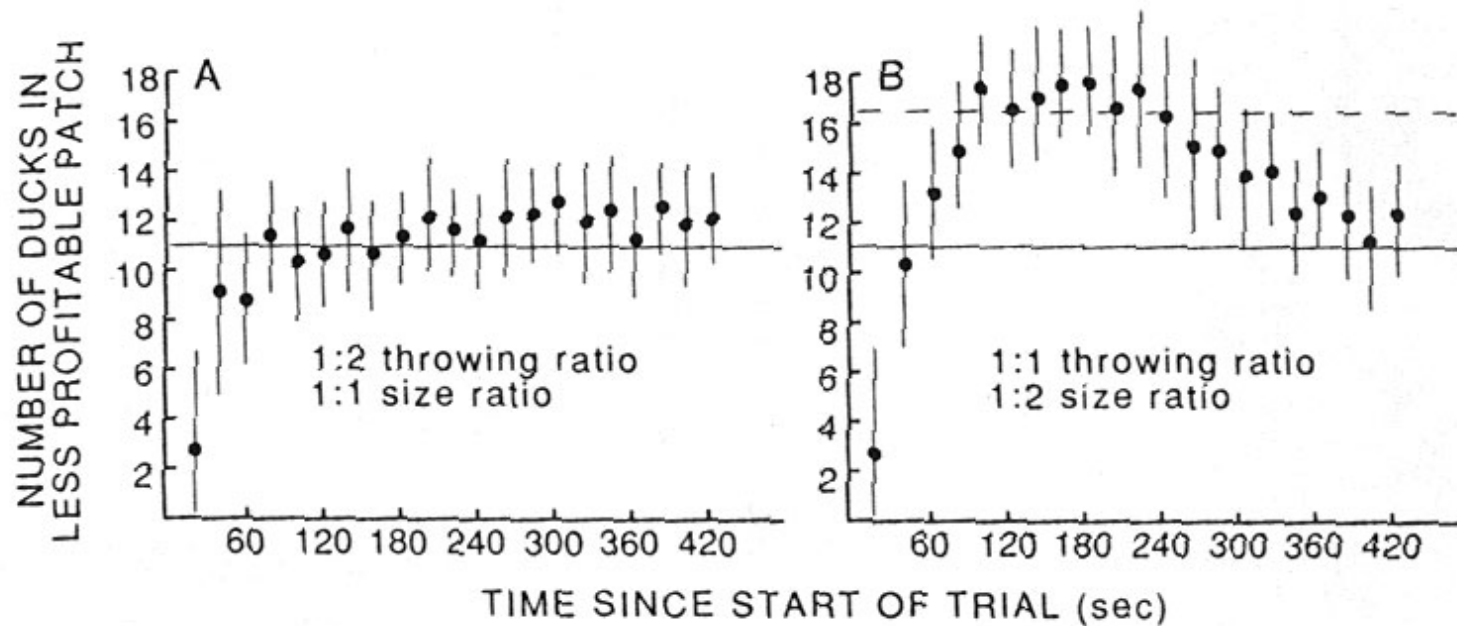


Ideal Free Ducks: flock of 33 ducks, two humans throwing pieces of bread.

A: both throw once per 5 seconds.

B: one throws once per 5 seconds, the other throws once per 10 seconds.

(from Harper 1982, via Gallistel 1990)



More duck-pond psychology – same 33 ducks:

A: same size bread chunks, different rates of throwing.

B: same rates of throwing, 4-gram vs. 2-gram bread chunks.

Linear operator model

- The animal maintains an estimate of resource density for each patch (or response frequency in p-learning)
- At certain points, the estimate is updated
- The new estimate is a linear combination of the old estimate and the “current capture quantity”

Updating equation:
$$E_n = wE_{n-1} + (1 - w)C$$

w “memory constant”

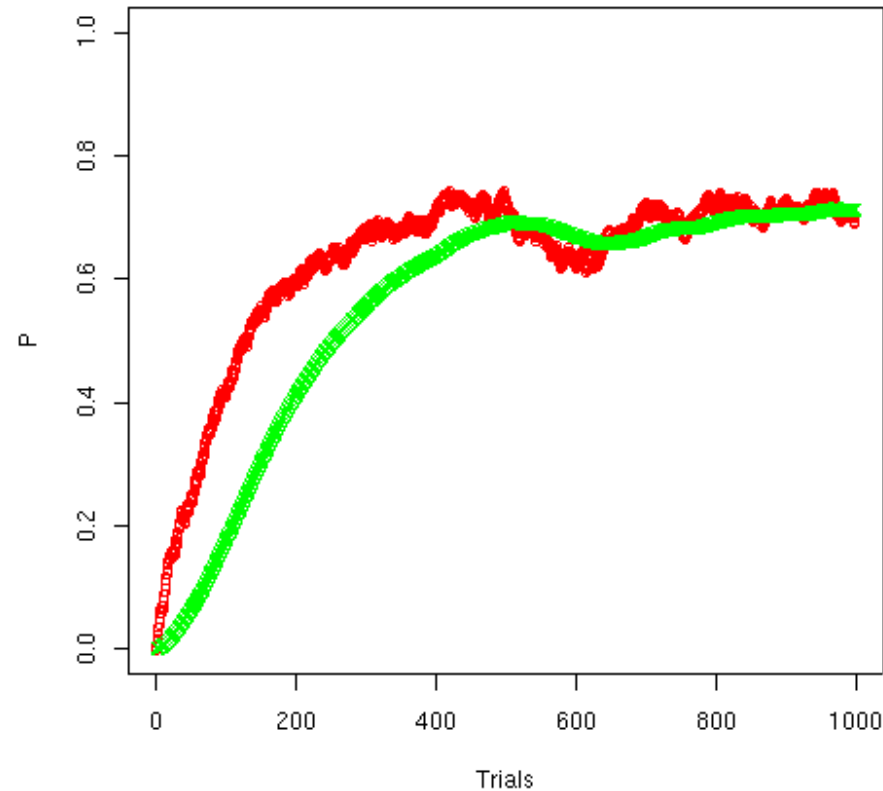
C “current capture quantity”

Bush & Mosteller (1951), Lea & Dow (1984)

What is E?

- In different models:
 - Estimate of resource density
 - Estimate of event frequency
 - Probability of response
 - Strength of association
 - ???

Simulation of Probability Matching by "Linear Operator Models"
(Learning $P=0.7$ from start at $P=0.0$)



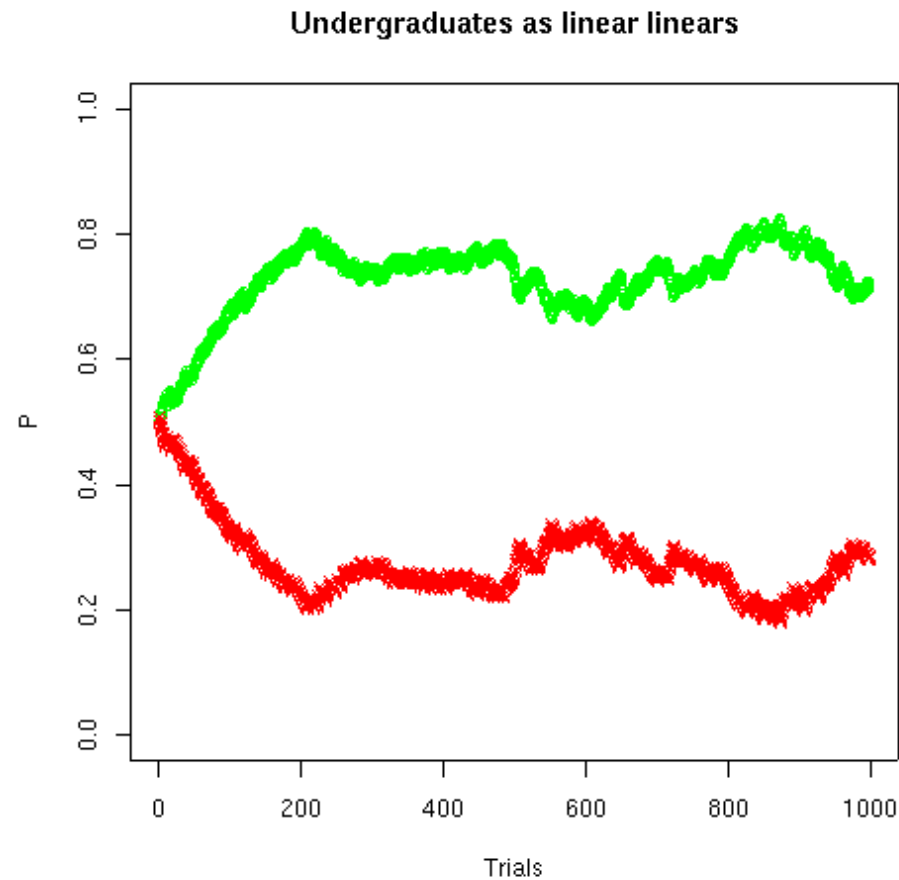
On each trial, "current capture quantity" is 1 with $p=.7$, 0 with $p=.3$

Red and green curves are "leaky integrators" with different time constants, i.e. different values of w in the updating equation.

Linear-operator model of the undergraduates' estimation of 'patch profitability':

On each trial, one of the two lights goes on, and each side's estimate is updated by 1 or 0 accordingly.

Note that the estimates for the two sides are complementary, and tend towards .75 and .25.

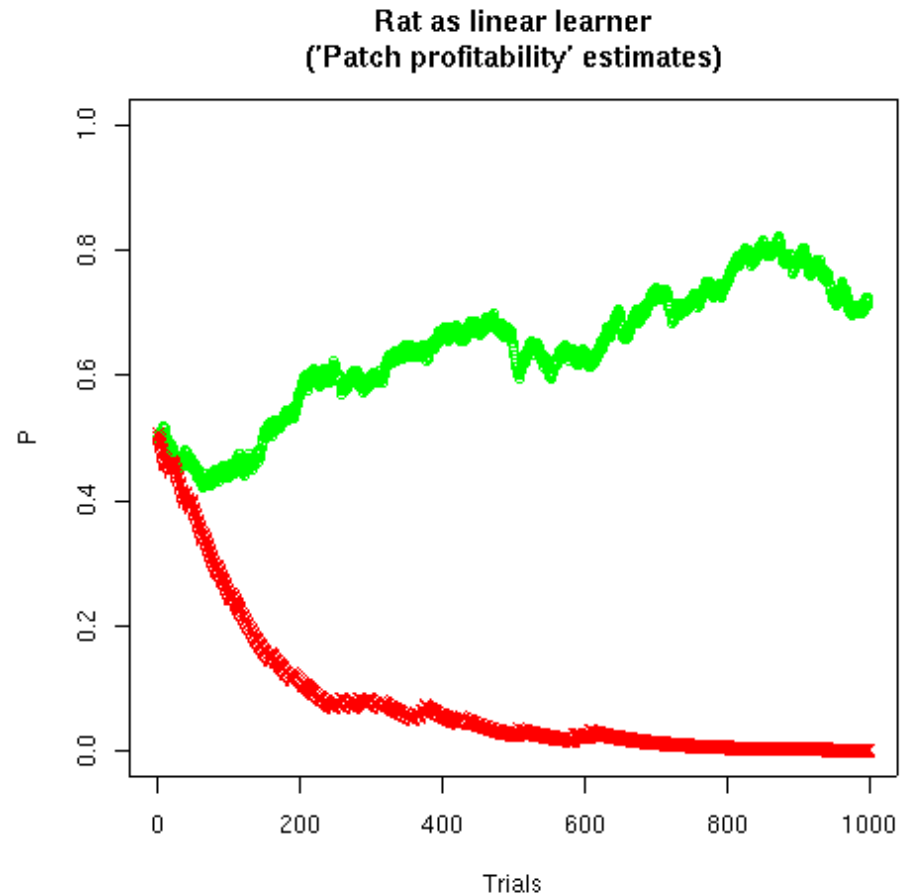


Linear-operator model of the rat's estimate of 'patch profitability':

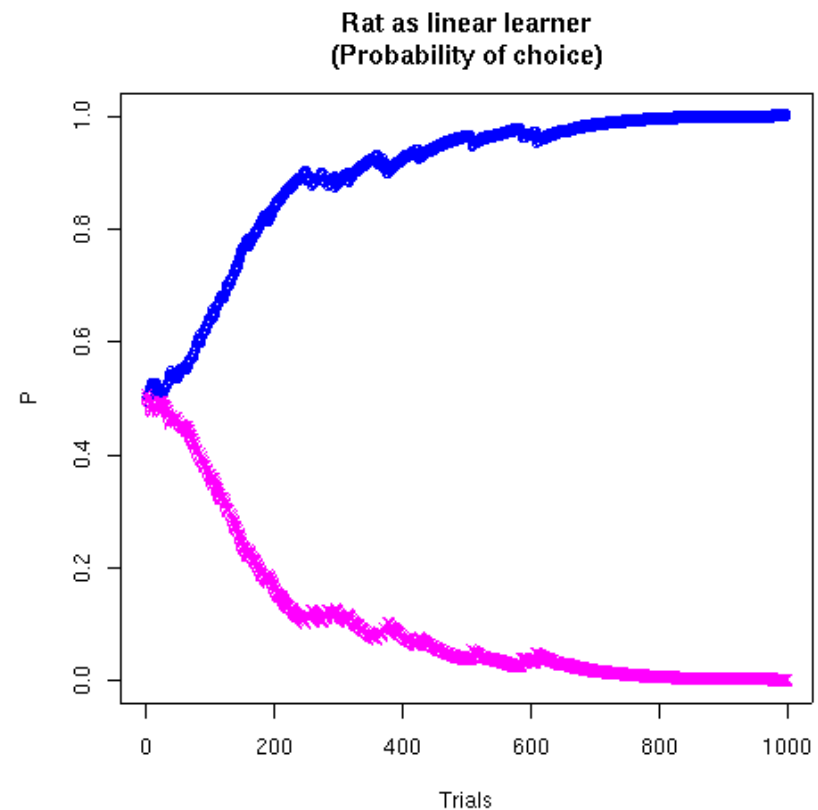
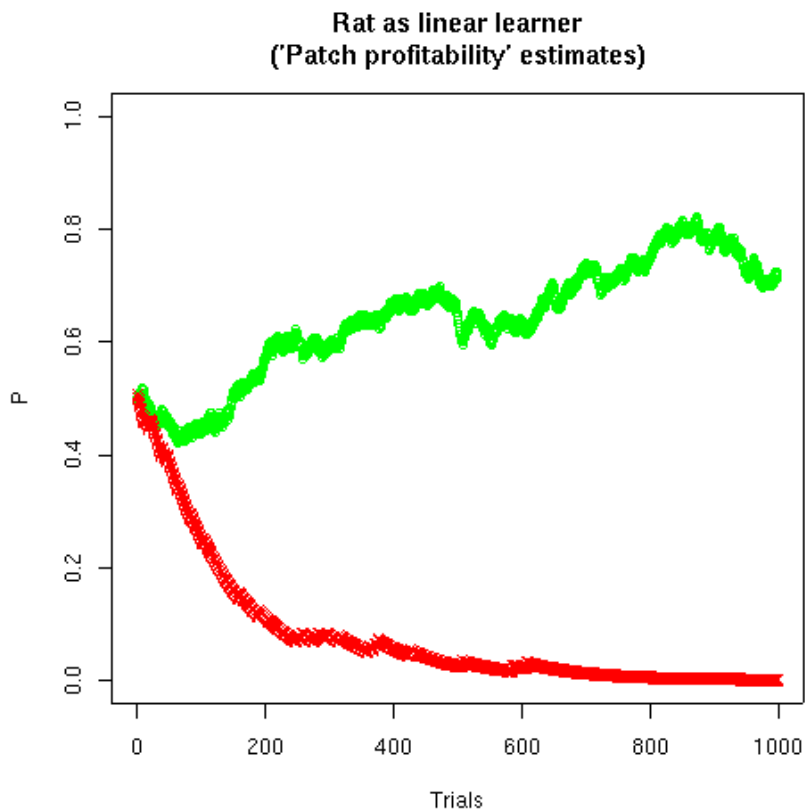
If the rat chooses correctly, the side chosen gets 1 and the other side 0.

If the rat chooses wrong, both sides get 0 (because there is no feedback).

Note that the estimates for the two sides are not complementary. The estimate for the higher-rate side tends towards the true rate (here 75%). The estimate for the lower-rate side tends towards zero (because the rat increasingly chooses the higher-rate side).



Since **animals ... proportion their choices in accord with the relative expected rates**, the model of the rat's behavior tends quickly towards maximization. Thus in this case (single animal without competition), less information (i.e. no feedback) leads to a higher-payoff strategy.

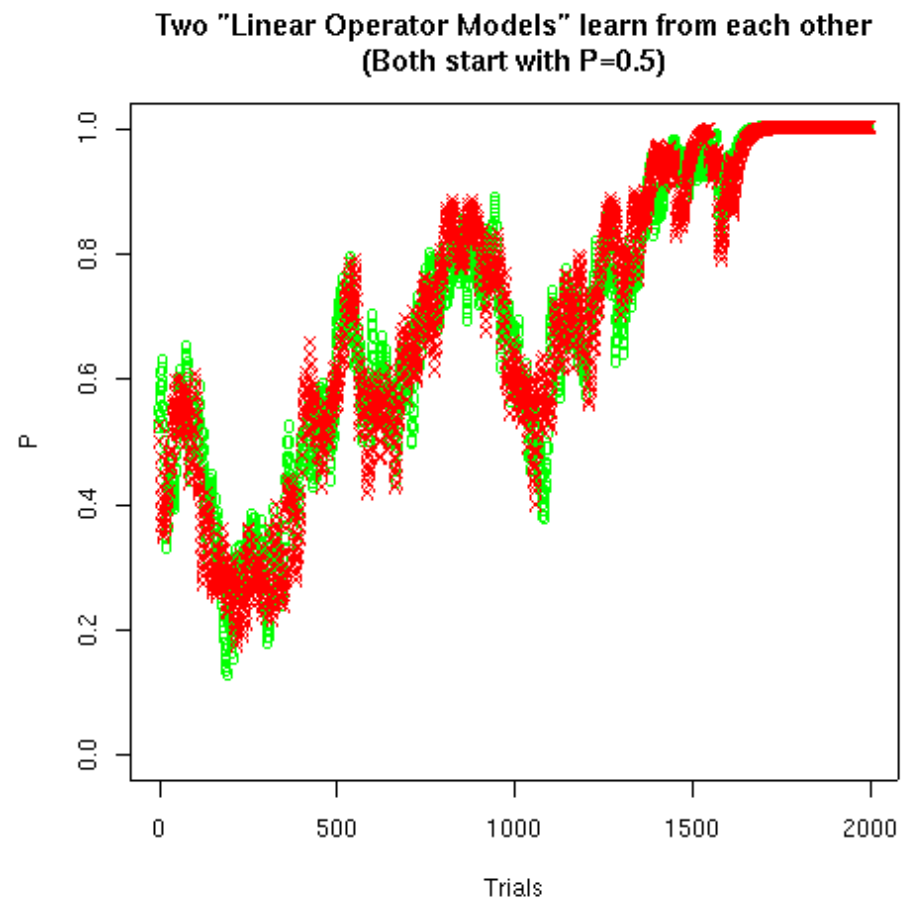


The rat's behavior influences the evidence that it sees. This feedback loop drives its estimate of food-provisioning probability in the lower-rate branch to zero.

If the same learning model is applied to a two-choice situation in which the evidence about both choices is influenced by the learner's behavior – as in the case where two linear-operator learners are estimating one another's behavioral dispositions – then the same feedback effect will drive the estimate for one choice to one, and the other to zero.

However, it's random which choice goes to one and which to zero.

Two models, each responding to the stochastic behavior of the other (green and red traces):



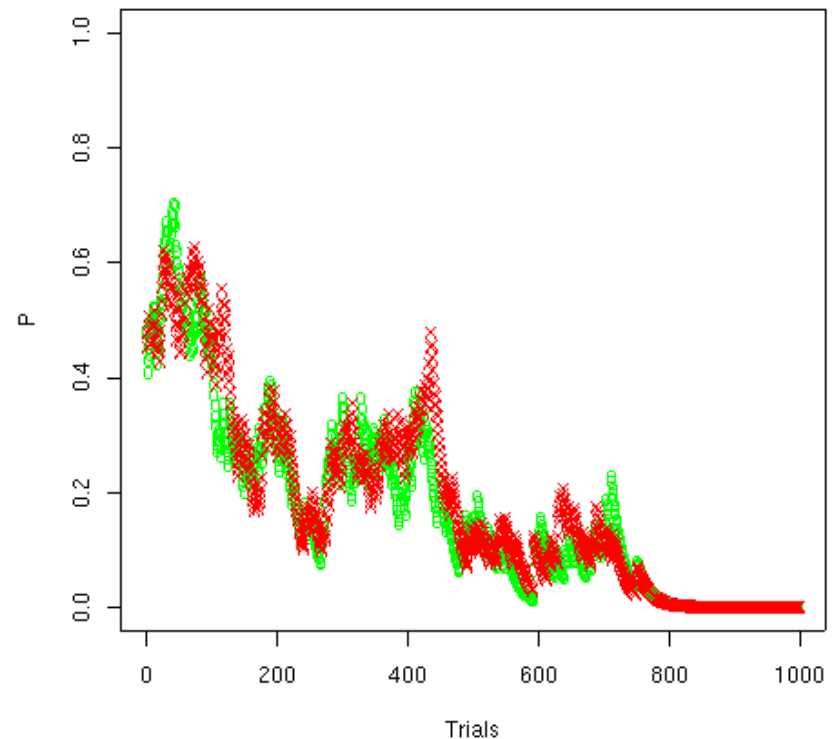
Another run, with a different random seed, where both go to zero rather than to one:

If this process is repeated for multiple independent features, the result is the emergence of random but shared structure.

Each feature goes to 1 or 0 randomly, for both participants.

The process generalizes to larger “communities” of social learners; this is just what happened in the naming model.

Two “Linear Operator Models” learn from each other
(Different seed)



The learning model, though simplistic, is plausible as a zeroth-order characterization of biological strategies for frequency estimation.

This increases the motivation for exploring the rest of the naming model.

Outline

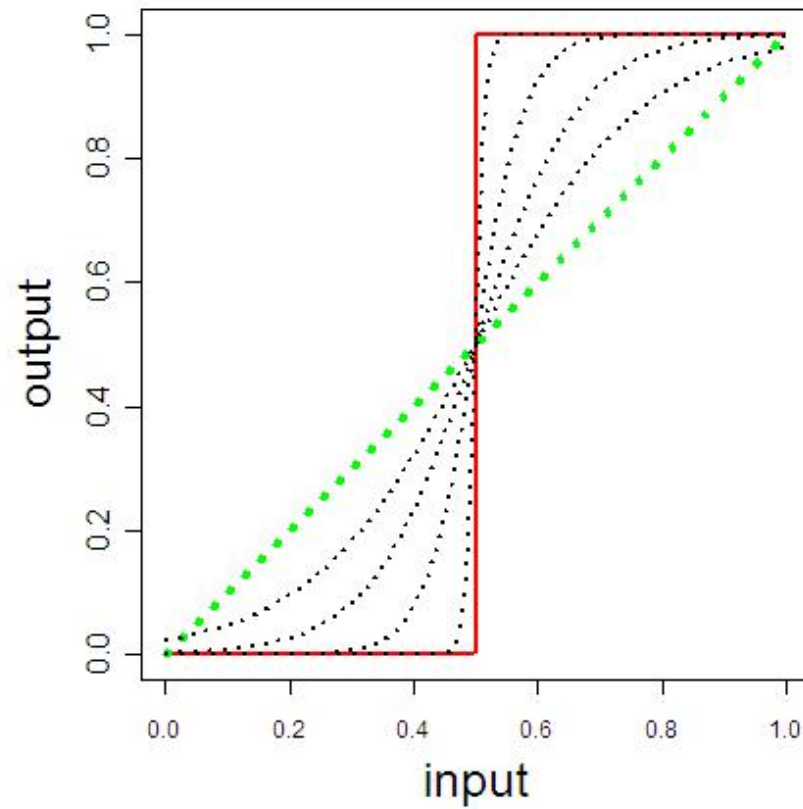
1. An origin myth: naming without Adam
a computer-assisted thought experiment
2. That old-time learning theory
linear operator models of probability learning
and expected rate learning
3. **Some morals:**
 - **Another advantage of categorical perception**
 - **Grammatical beliefs as random variables**

Stochastic belief + categorical perception + social interaction
= emergence of coherent shared grammar

Perception of pronunciation must be categorical

- Categorical (i.e. digital) perception is crucial for a communication system with many well-differentiated words
- Arguments based on “error correction”
 - digital transmission avoids accumulation of noise along multi-step transmission paths
 - permits redundant coding for correction of digital errors
- Equally strong arguments based on social convergence?
 - categorization is the nonlinearity that creates the attractors in the iterated map of reciprocal learning
 - milder nonlinearities would also work here
- Note that perceptual orthogonality of phonetic dimensions was also assumed
 - Orthogonality is not essential, but:
 - multiple dimensions are needed for adequate size of lexical space given modest number of distinct values on each dimension
 - orthogonal binary variables make the model simple

From veridical to categorical



Chicago 5/24/2005

Beliefs about pronunciation must be stochastic

- “Pronunciation field” of an entry in the mental lexicon may be viewed as a random variable,
i.e. a distribution over possible pronunciations
- Evidence from variability in performance
 - probabilities traditionally placed in rules or constraints (or competition between whole grammars) rather than in lexical forms themselves
- A new argument based on social convergence?
 - underlying lexical forms as distributions over symbol sequences rather than symbol sequences themselves
 - allows learning to “hill climb” in the face of social variation and channel noise
- Note that computational linguists now routinely assume that syntactic beliefs are random variables in a similar sense

Other ideas about linguistic variation

- variable rules
 - estimated by logistic regression on conditioning of alternatives
- “competing grammars”
 - linear combination of overall categorical systems
- stochastic ranking of OT constraints
- In the models discussed today
 - beliefs about the pronunciation of individual words are random variables, with parameters estimated from utterance-by-utterance experience by a simple and general learning process
 - stochastic rules or constraints produce similar behavior but have different learning properties (because they generalize across words)
 - Paradoxically, stochastic beliefs about individual lexical items are seen here as essential to the categorical coherence of linguistic knowledge in a speech community

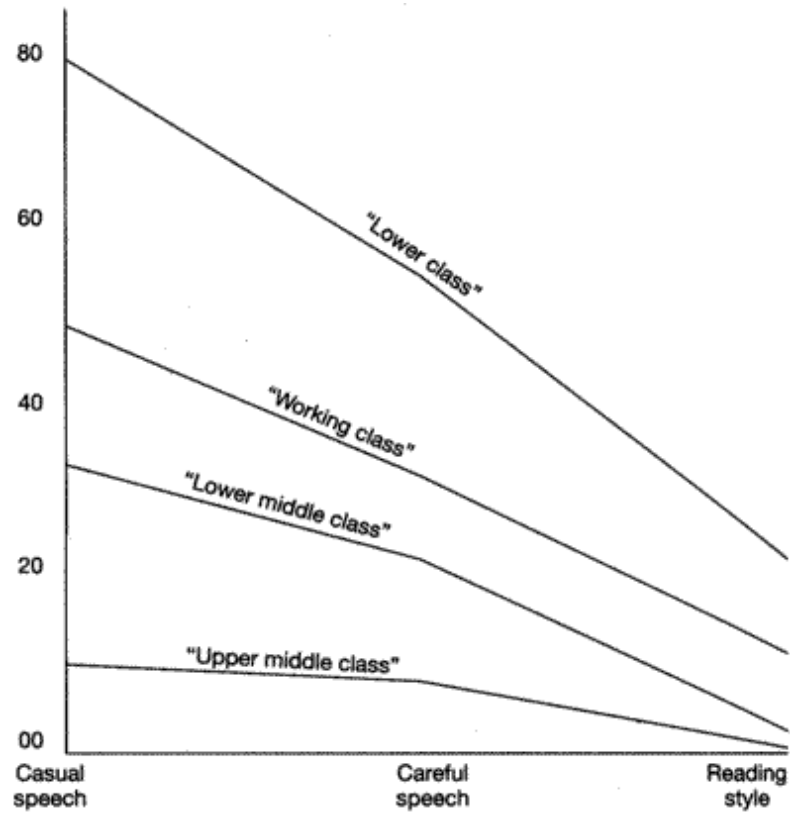
A note on evolutionary plausibility?

- Learned stochastic beliefs are the norm
 - no special pleading needed here
- Perceptual factoring of phonetic dimensions is helpful for vocal imitation
 - factors complex learning problem into several simple ones
- What about categorical perception?
 - natural nonlinearities?
 - scaling of psychometric functions?
 - semi-categorical functions also provide positive feedback that creates attractors in the iterated map of reciprocal learning
 - more categorical → better communication

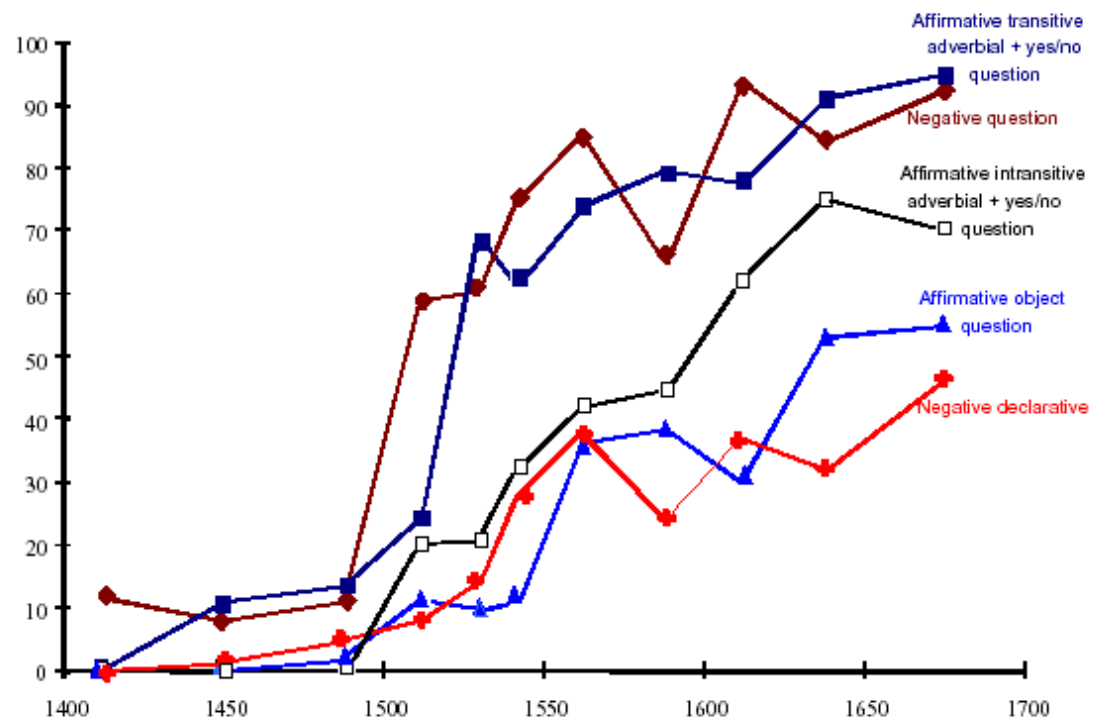
Comparison to Collective Intelligence in Social Insects

Self-organization was originally introduced in the context of physics and chemistry to describe how microscopic processes give rise to macroscopic structures in out-of-equilibrium systems. Recent research that extends this concept to ethology, suggests that it provides a concise description of a wide range of collective phenomena in animals, especially in social insects. This description does not rely on individual complexity to account for complex spatiotemporal features which emerge at the colony level, but rather assumes that interactions among simple individuals can produce highly structured collective behaviors.

E. Bonabeau et al., *Self-Organization in Social Insects*, 1997

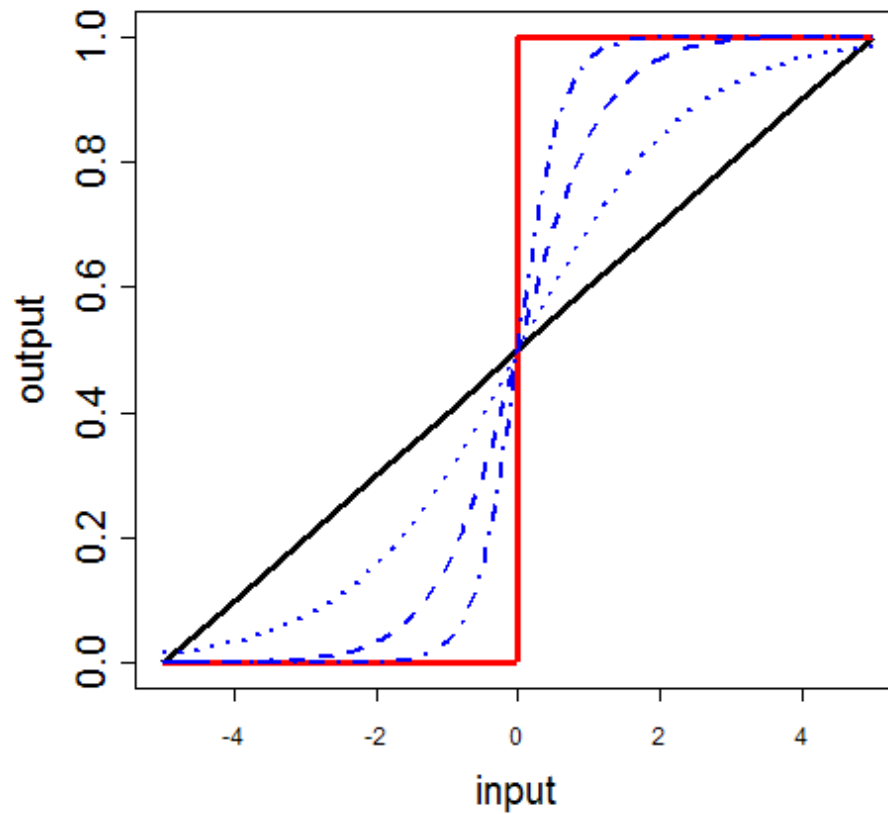


Percentage of “g-dropping” by formality & social class
(NYC data from Labov 1969)

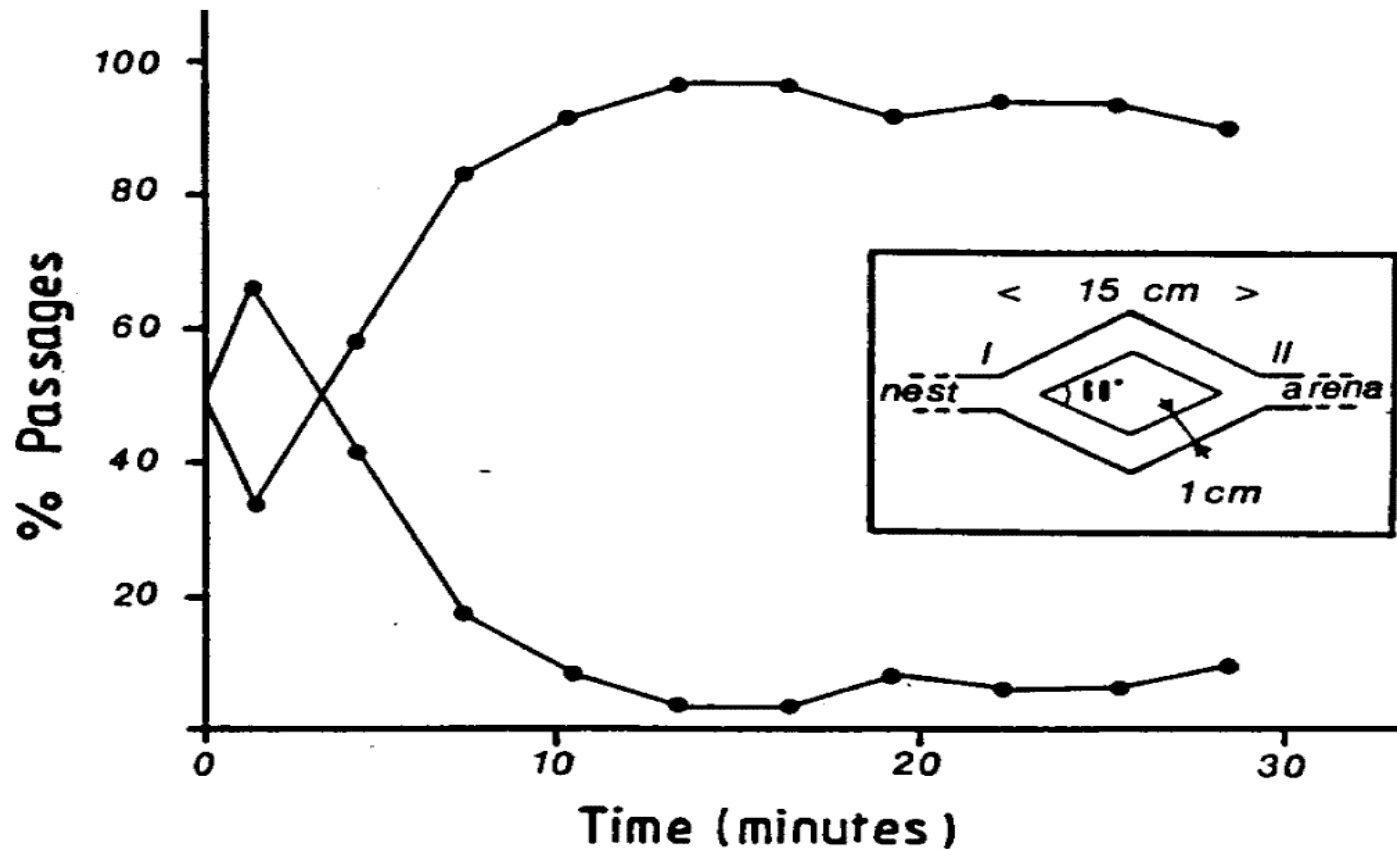


The rise of periphrastic *do*
 (from Ellegård 1953 via Kroch 2000).

From linear to categorical perception

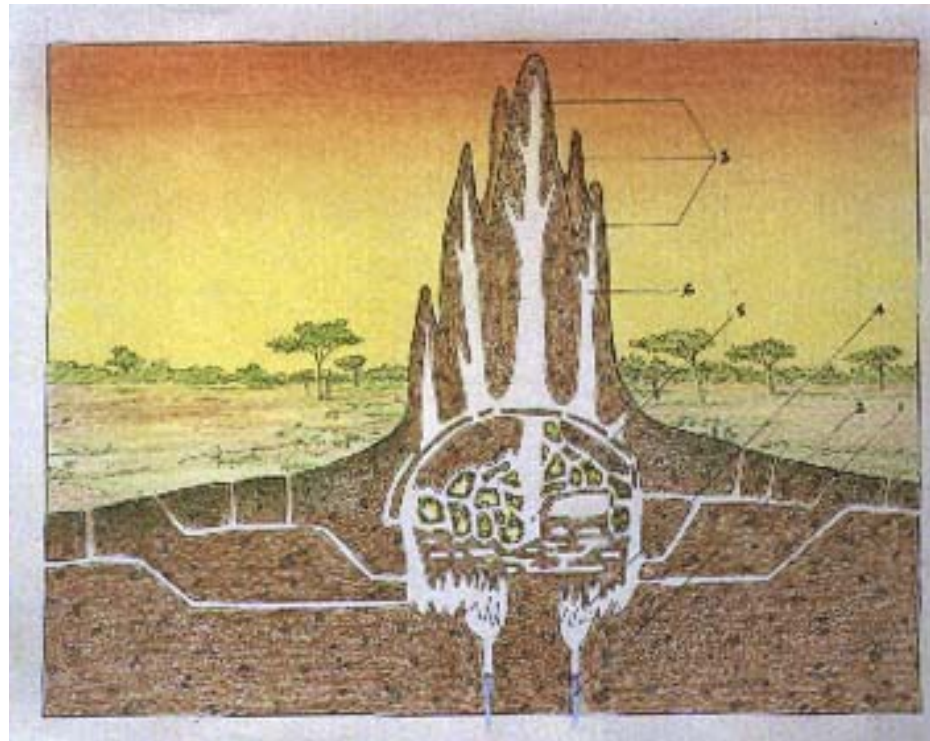


Buridan's Ants make a decision



Percentage of *Iridomyrex Humulis* workers passing each (equal) arm of bridge per 3-minute period

More complex emergent structure: termite mounds...



Termite Theory:

Bruinsma (1979): positive feedback mechanisms, involving responses to a short-lived pheromone in deposited soil pellets, a long-lived pheromone along travel paths, and a general tendency to orient pellet deposition to spatial heterogeneities; these lead to the construction of pillars and roofed lamellae around the queen.

Deneubourg (1977): a simple model with parameters for the random walk of the termites and the diffusion and attractivity of the pellet pheromone, producing a regular array of pillars.

Bonabeau et al. (1997): air convection, pheromone trails along walkways, and pheromones emitted by the queen; "under certain conditions, pillars are transformed into walls or galleries or chambers", with different outcomes depending not on changes in behavioral dispositions but on environmental changes caused by previous building. Thus "nest complexity can result from the unfolding of a morphogenetic process that progressively generates a diversity of history-dependent structures."

Similar to models of embryological morphogenesis.