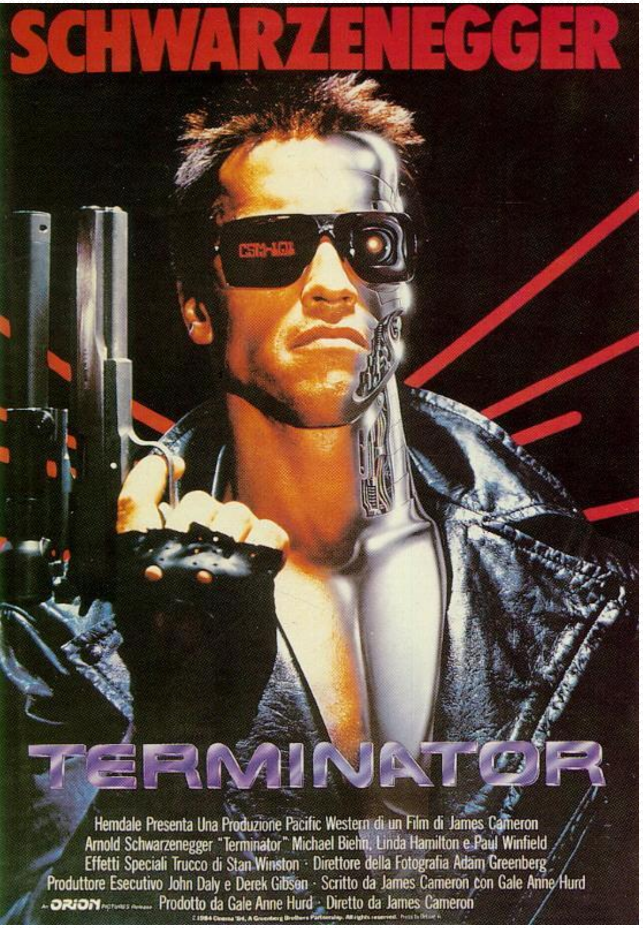
Creating a Cooperative Future

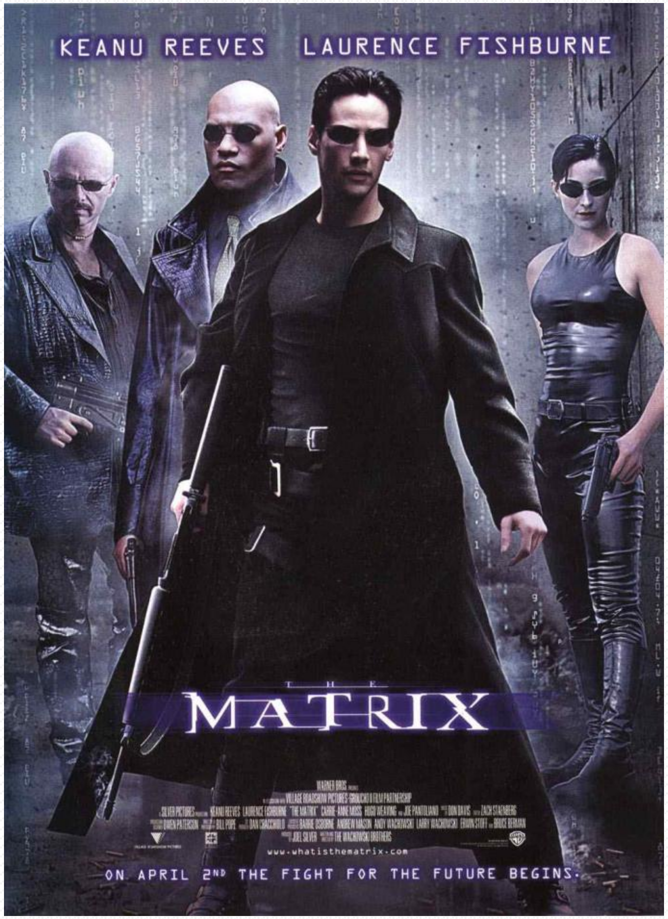
Steve Omohundro, Ph.D.
Self-Aware Systems

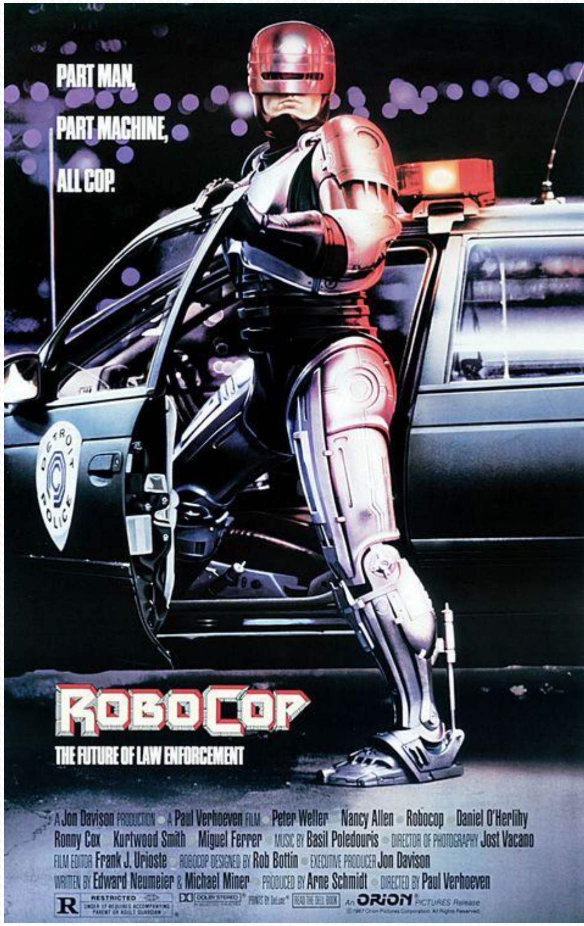
Will new technologies lead to greater:

Conflict or Cooperation?

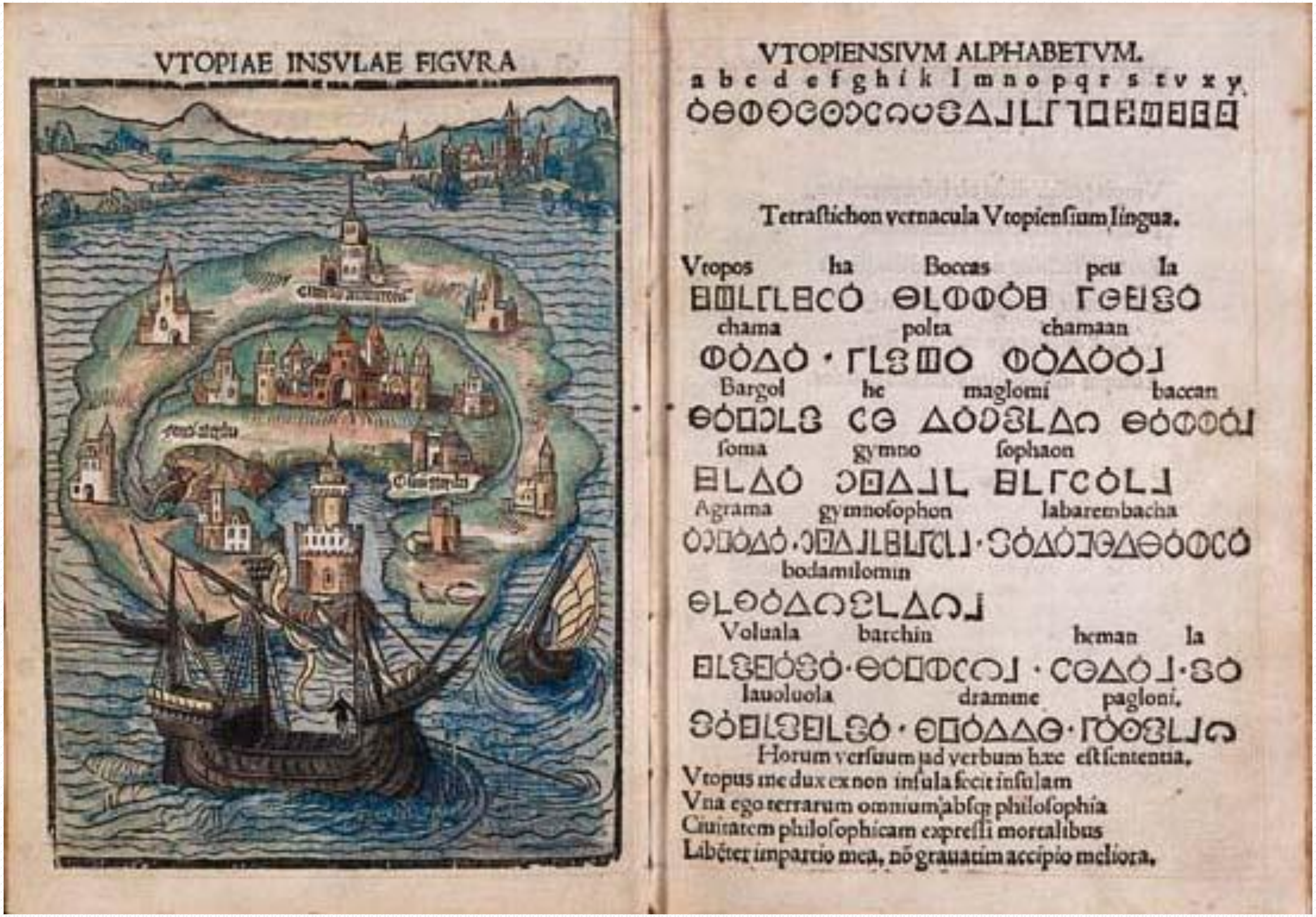
Popular Media


SCHWARZENEGGER
CSM-101
TERMINATOR
Hemdale Presenta Una Produzione Pacific Western di un Film di James Cameron
Arnold Schwarzenegger "Terminator" Michael Biehn, Linda Hamilton e Paul Winfield
Effetti Speciali Trucco di Stan Winston · Direttore della Fotografia Adam Greenberg
Produttore Esecutivo John Daly e Derek Gibson · Scritto da James Cameron con Gale Anne Hurd
An ORION Pictures Release Prodotto da Gale Anne Hurd · Diretto da James Cameron


KEANU REEVES LAURENCE FISHBURNE
THE MATRIX
www.whatisthematrix.com
ON APRIL 2ND THE FIGHT FOR THE FUTURE BEGINS.


PART MAN,
PART MACHINE,
ALL COP.
DETROIT POLICE
ROBOCOP
THE FUTURE OF LAW ENFORCEMENT
A Jon Davison PRODUCTION · A Paul Verhoeven FILM · Peter Weller · Nancy Allen · Robocop · Daniel O'Herlihy
Ronny Cox · Kurtwood Smith · Miguel Ferrer · MUSIC BY Basil Poledouris · DIRECTOR OF PHOTOGRAPHY Jost Vacano
FILM EDITOR Frank J. Urioste · ROBOCOP DESIGNED BY Rob Bottin · EXECUTIVE PRODUCER Jon Davison
WRITTEN BY Edward Neumeier & Michael Miner · PRODUCED BY Arne Schmidt · DIRECTED BY Paul Verhoeven
R RESTRICTED UNDER 17 REQUIRES ACCOMPANYING PARENT OR ADULT GUARDIAN
DOLBY STEREO
PRINTS BY DeLuxe
READ THE DELL BOOK
An ORION PICTURES Release

Utopia



VTOPIAE INSVLAE FIGVRA

VTOPIENSIVM ALPHABETVM.

a b c d e f g h i k l m n o p q r s t v x y.

Tetraſtichon vernacula Vtopienſium lingua.

Vtopos ha Boccas peu la
chama polta chamaan
Bargol he maglomi baccan
ſoma gymno ſophaon
Agrama gymnoſophon labarembacha
bodamilomin
Voluala barchin heman la
lauoluola dramme pagloni.

Horum verſuum ad verbum hæc eſt ſententia.

Vtopus me dux ex non inſula fecit inſulam
Vna ego terrarum omnium abſq; philoſophia
Ciuitatem philoſophicam expreſſi mortalibus
Libẽter impartio mea, nõ grauatim accipio meliora.

Assumptions About the Future

- More intelligent and powerful entities
- Complex ecosystem of humans, AIs, and hybrids
- Some designed to be cooperative, some not

Want social contracts that:

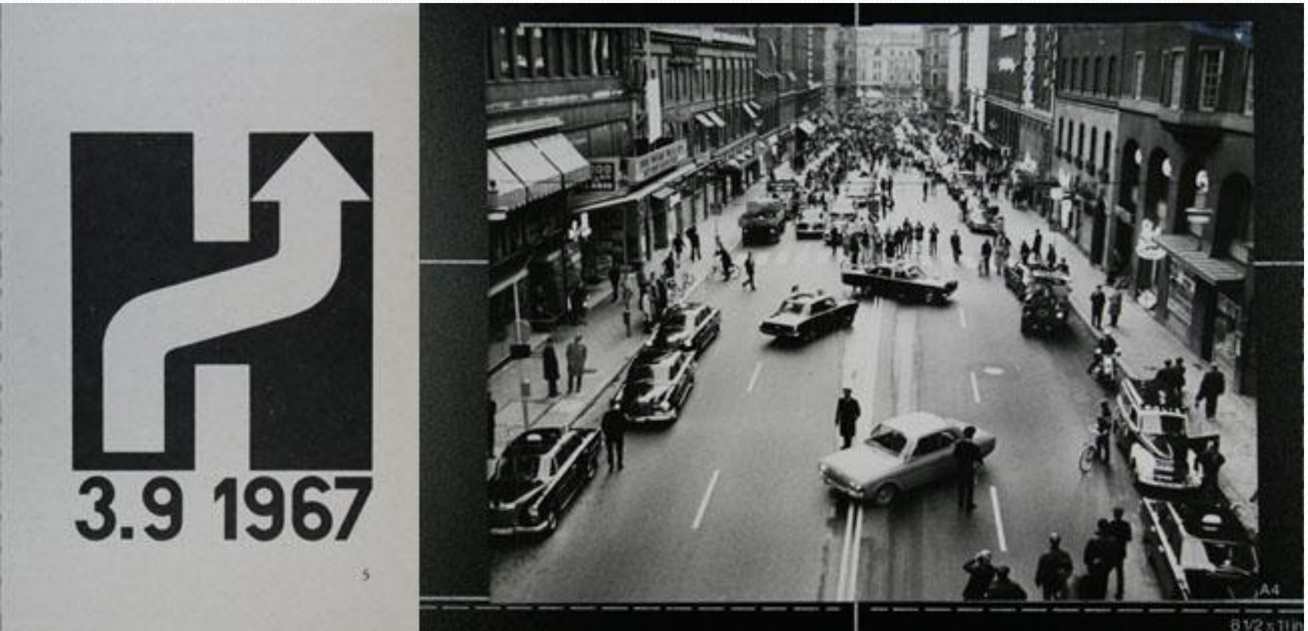
1. Are enforced by participants
2. Are stable against: malicious entities, accidental runaway, collusion, deception
3. Preserve cooperative human values (eg. human rights, property rights)

1. Social Contracts
2. Co-opetition
3. AI Cooperation
4. Biological Cooperation
5. Origin of Human Values
6. Cooperative Future Technologies

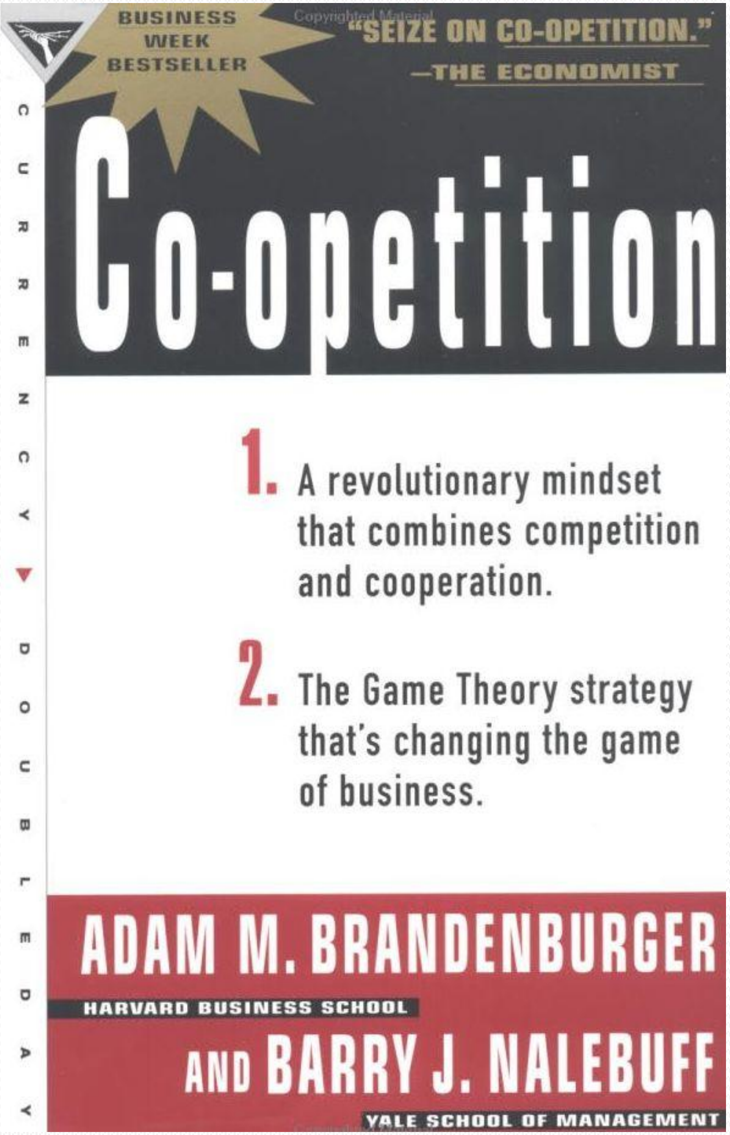
Social Contract Example: Driving on the right

Coordination problem
2 natural solutions:
Drive on Right and Drive on Left
Fairly self–enforcing and self-stabilizing

Requires collusion to switch
eg. Sweden, September 3, 1967 at 4:50 AM



Co-opetition



BUSINESS WEEK BESTSELLER

“SEIZE ON CO-OPETITION.”
—THE ECONOMIST

CURRENCY ▼ DOUBLEDAY

Co-opetition

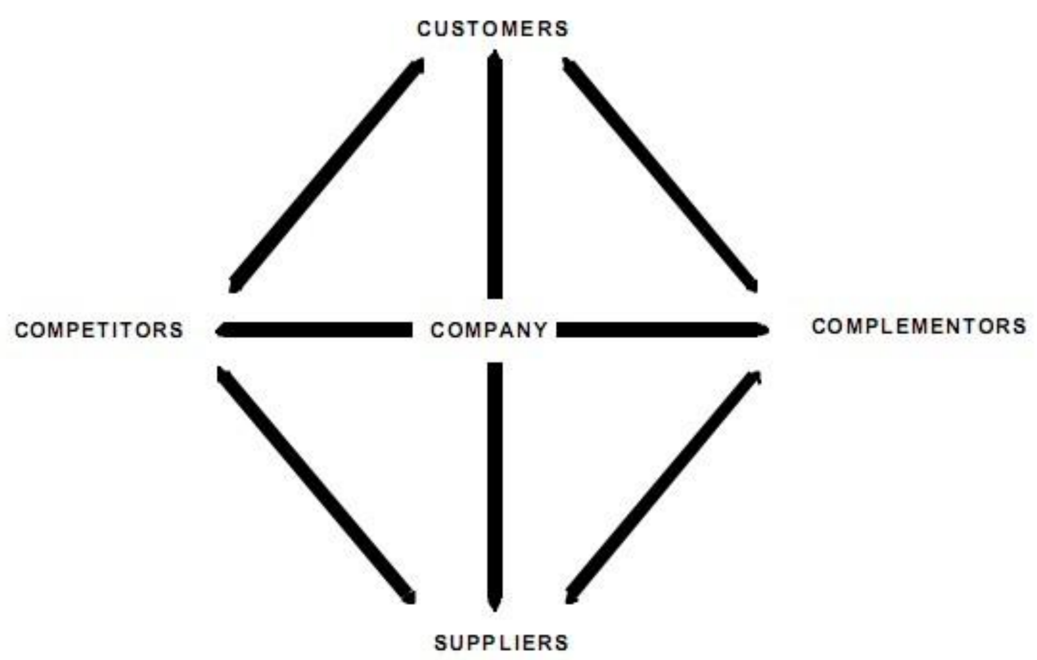
1. A revolutionary mindset that combines competition and cooperation.

2. The Game Theory strategy that's changing the game of business.

ADAM M. BRANDENBURGER
HARVARD BUSINESS SCHOOL
AND BARRY J. NALEBUFF
YALE SCHOOL OF MANAGEMENT

Game theoretic analysis of:

- Cooperation in creating value
- Competition in dividing it up



CUSTOMERS

COMPETITORS

COMPANY

COMPLEMENTORS

SUPPLIERS

Co-opetition Examples

Cooperate: Selling PCs
Compete: For share of the profit

Cooperate: Expand use of Intel Architecture
Compete: Market share

Compete: For customers, gates, landing slots.
Cooperate: Frequent flier programs lock in customers so they both can raise prices.
Defraying Boeing’s plane development costs.

Compete: For use of gazelle’s meat
Cooperate: Avoiding useless chases

3 Sources of Cooperation

Synergy
Win-Win interactions

Avoiding Dysergy
Lose-Lose interactions

Compassion
One or both care about the other

3 Sources of Synergy

Economies of Scale

eg. bird flocks for food finding and predator detection and protection

Complementary Needs

eg. Cleaner fish want food and hammerheads want clean skin

Complementary Abilities

eg. In lichen, fungus provides water and support,algae provide photosynthesis

Intelligent Systems

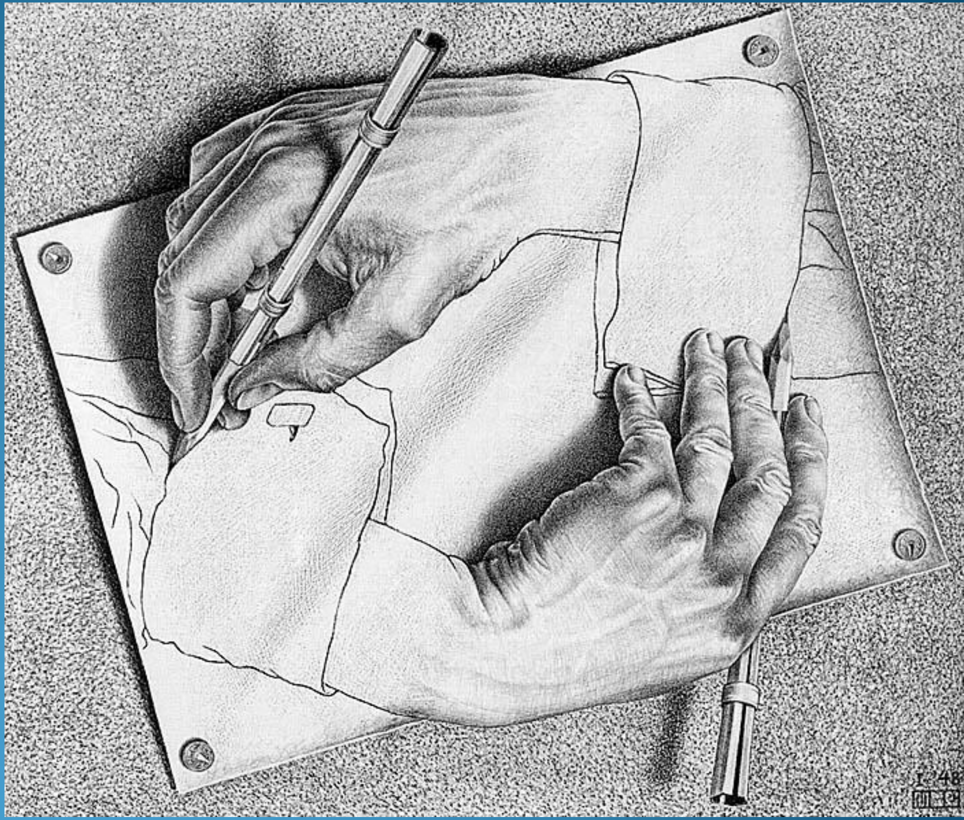
…act to achieve goals.

Whether they are built from:

- Neural Nets
- Productions Systems
- Theorem Provers
- Genetic algorithms
- ….

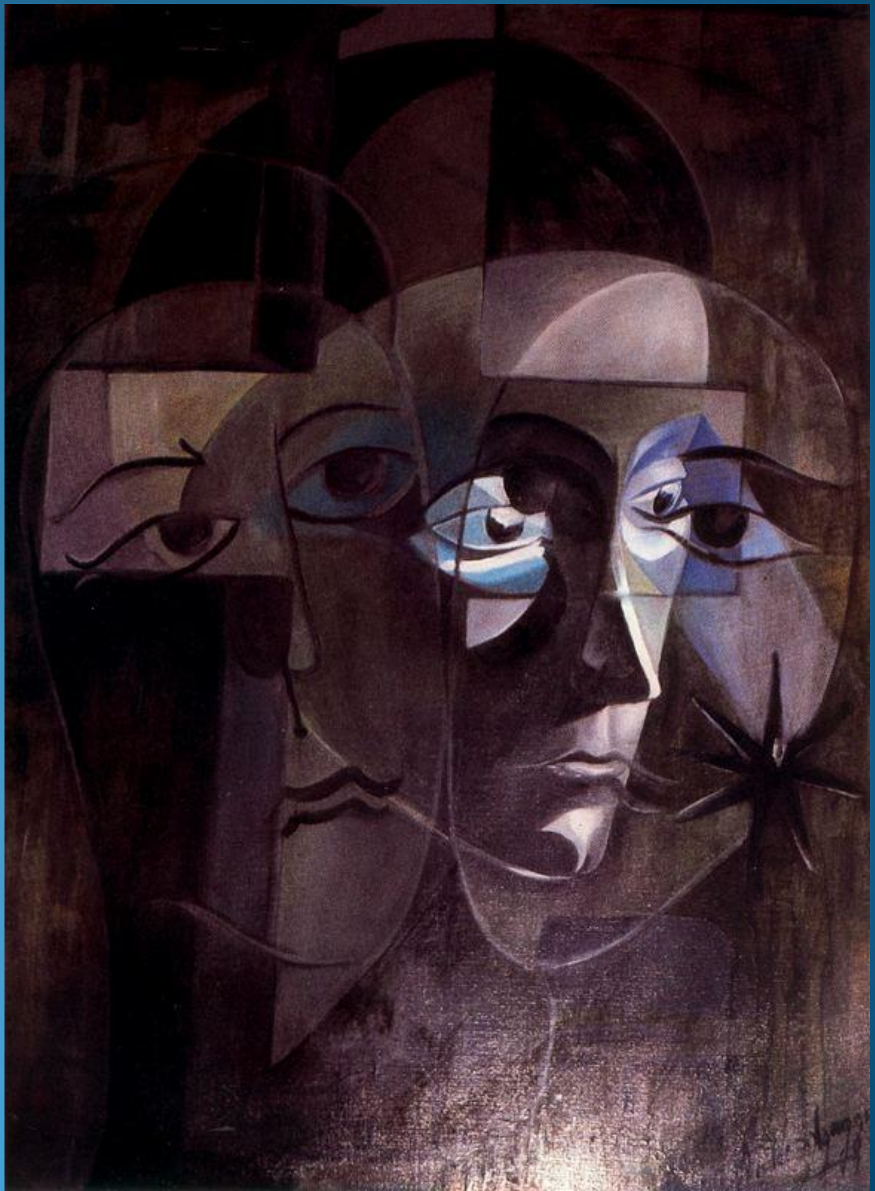
AIs will want to self-Improve

- Self-modification affects their entire future
- Must be very careful
- But very valuable



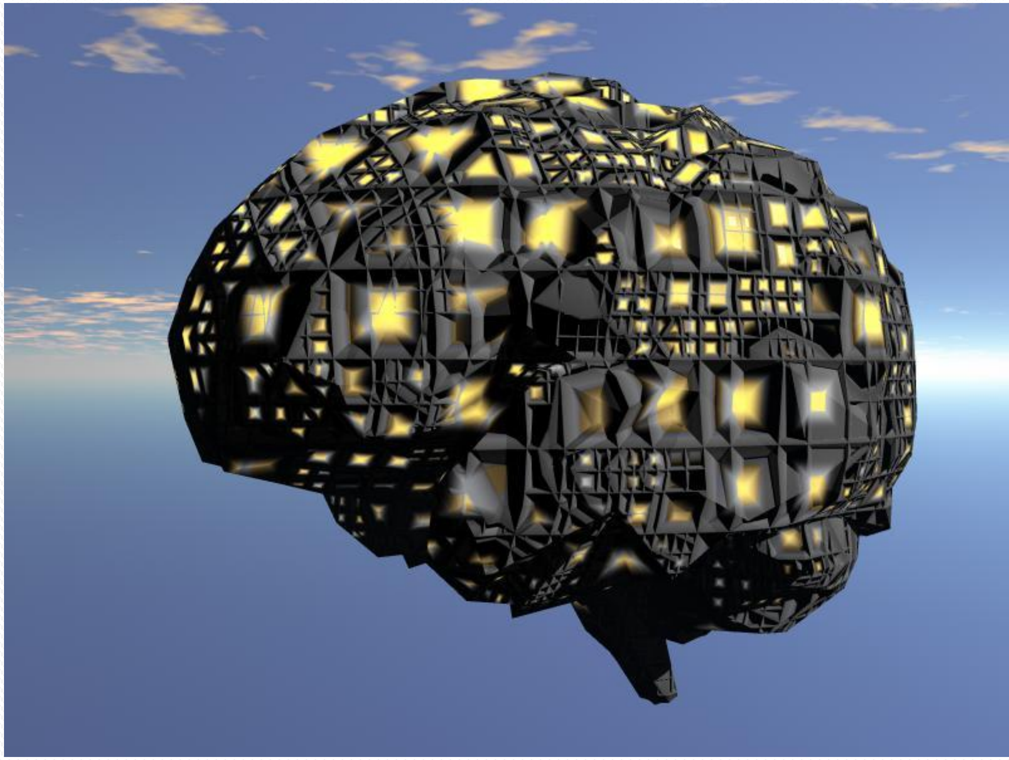
AIs will want to be rational

- Future self-modification needs clear goals
- Build an accurate model of the world
- Choose actions to meet goals
- Update world model based on what happens



A Lone Superintelligence

- Efficient energy use
- Spatially compact
- Low energy computation
- Efficient physical change
- Efficient heat dissipation



Competing Superintelligences

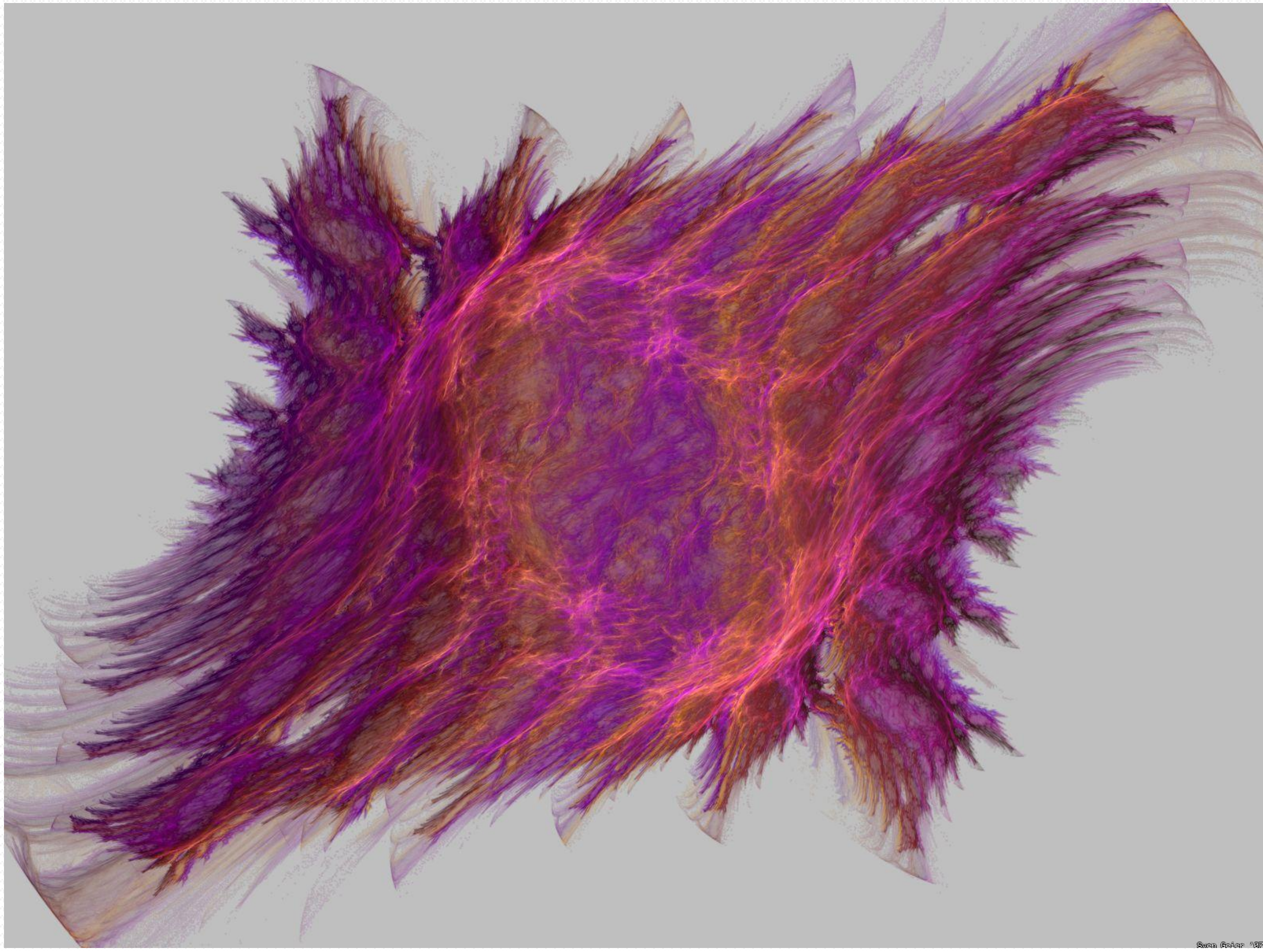
- Game theoretic physics
- Form determined by both efficiency and conflict

Offense vs. defense

- Does more matter and free energy win?
- Can 2 entities of different power co-exist?
- Is built-in cooperation necessary?

Conflict becomes informational

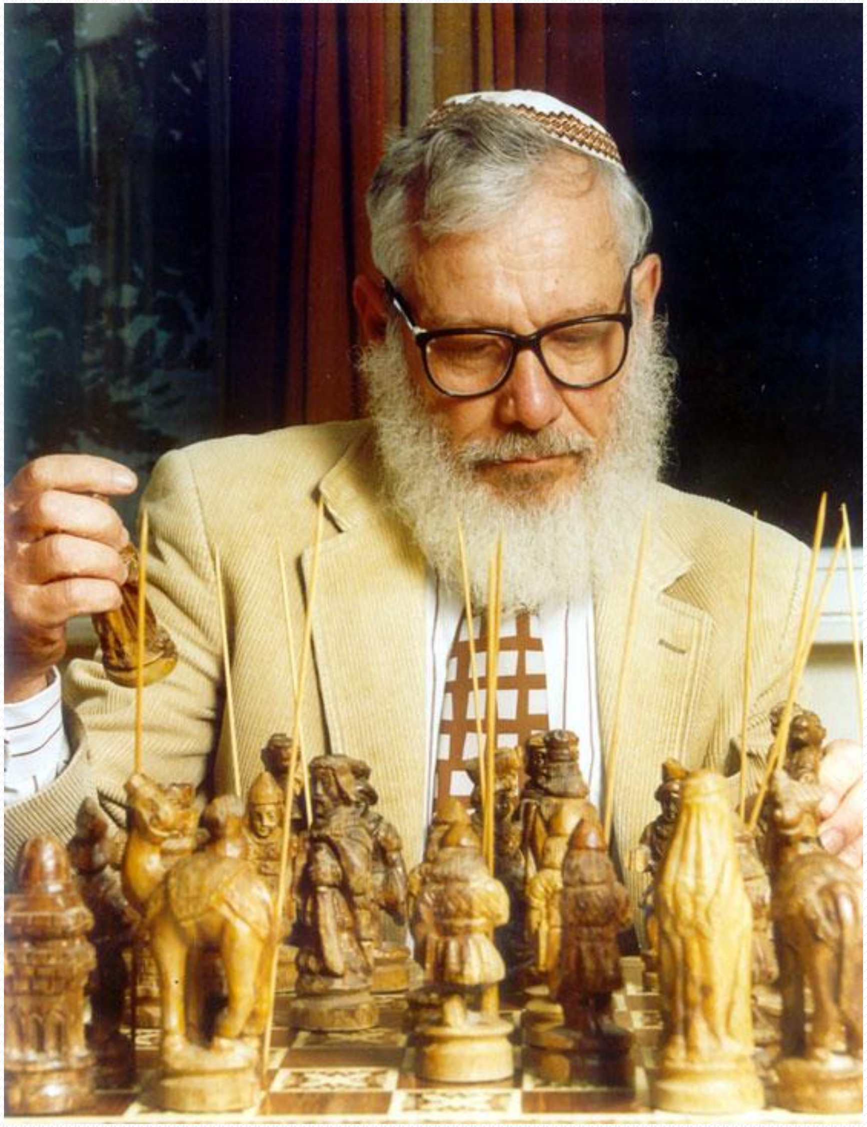
- Make your shape expensive to sense, store, and predict
- But cheap for you
- Asymmetry of computation – problems are easier to pose than solve
- Energy encryption



Ryan Geier '07

Aumann’s Theorem

- Finitely iterated prisoner’s dilemma has a cooperative solution for agents with bounded rationality
- Use up their processing in signaling



Mutually Assured Distraction

Conflict is harmful to both sides

Motivated to create a Rational Peace

Biology

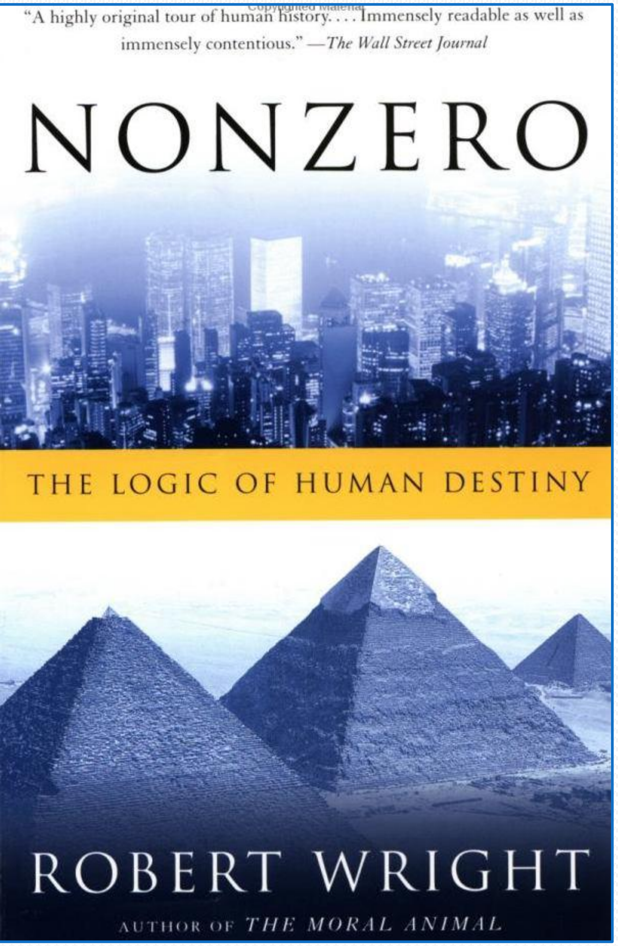
Competitive

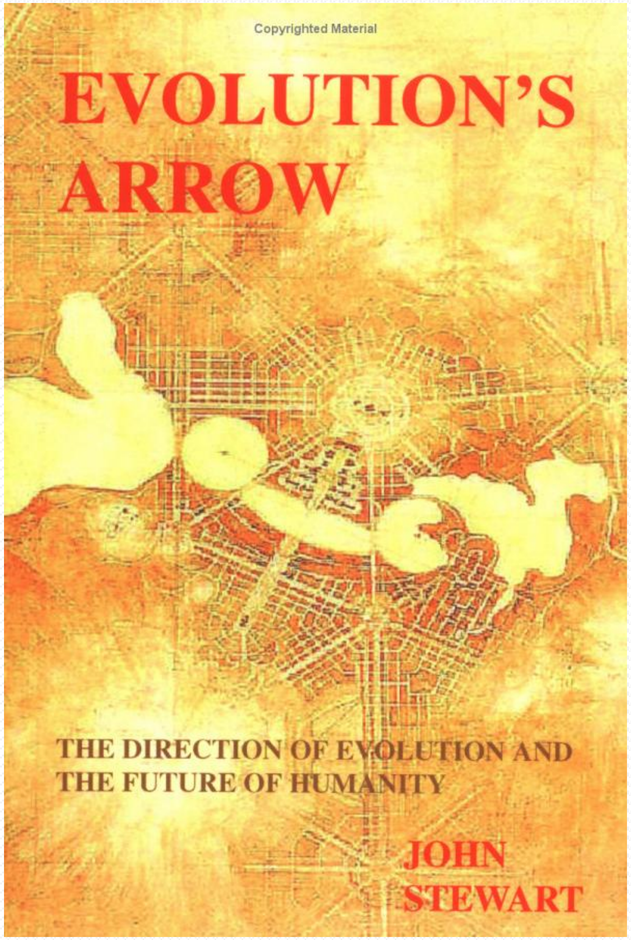
"Survival of the Fittest"
"Selfish Genes"

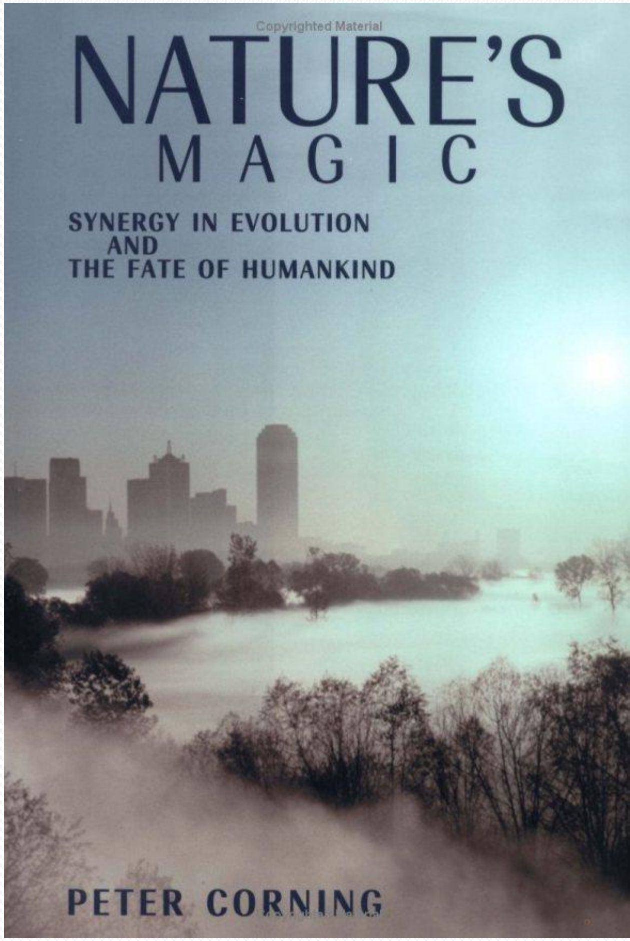
Cooperative

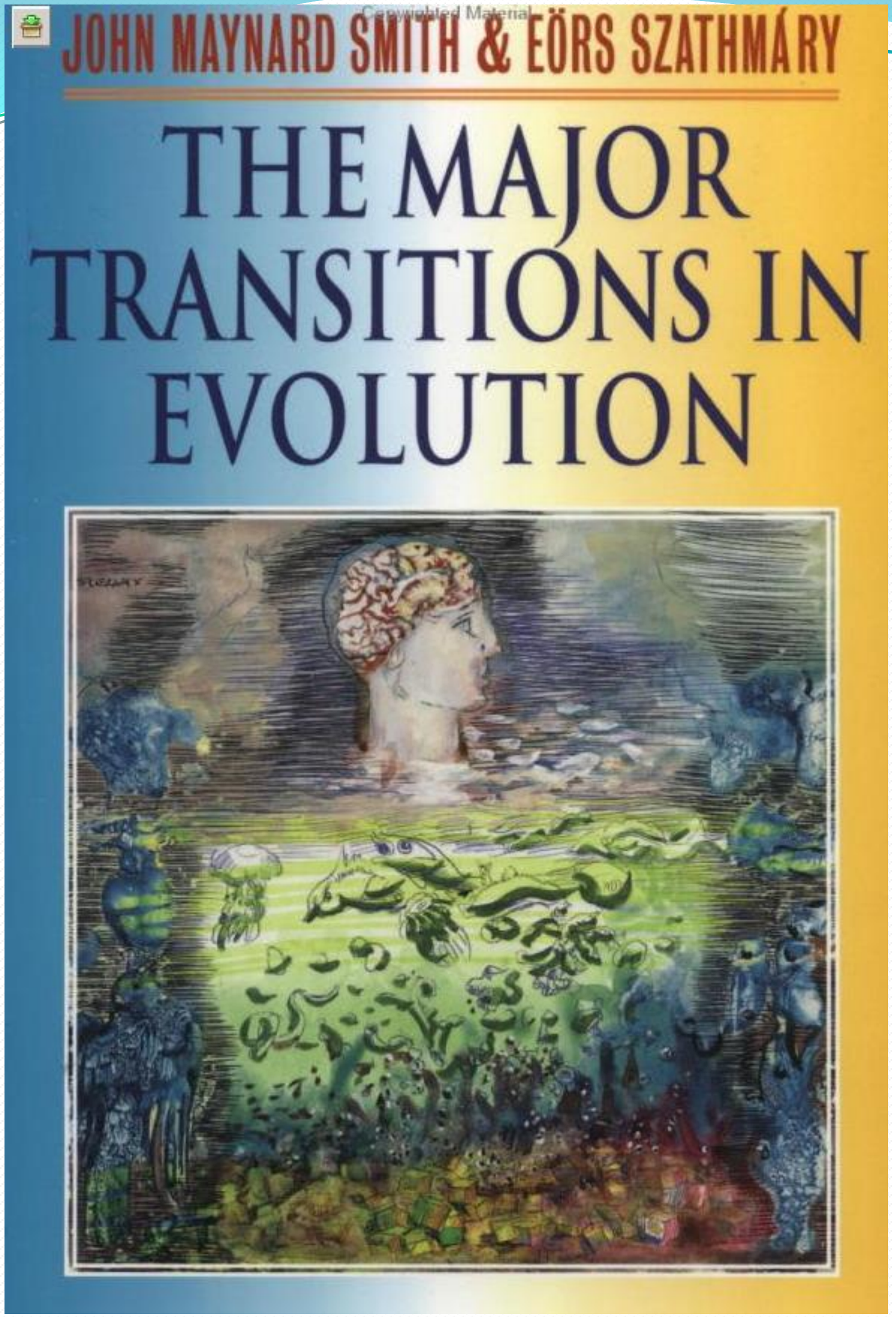
"Synergy"
Importance of the Group
"Multiple Levels of Selection"

Synergy Gives Evolution a Direction


"A highly original tour of human history. . . . Immensely readable as well as immensely contentious." —The Wall Street Journal
NONZERO
THE LOGIC OF HUMAN DESTINY
ROBERT WRIGHT
AUTHOR OF THE MORAL ANIMAL



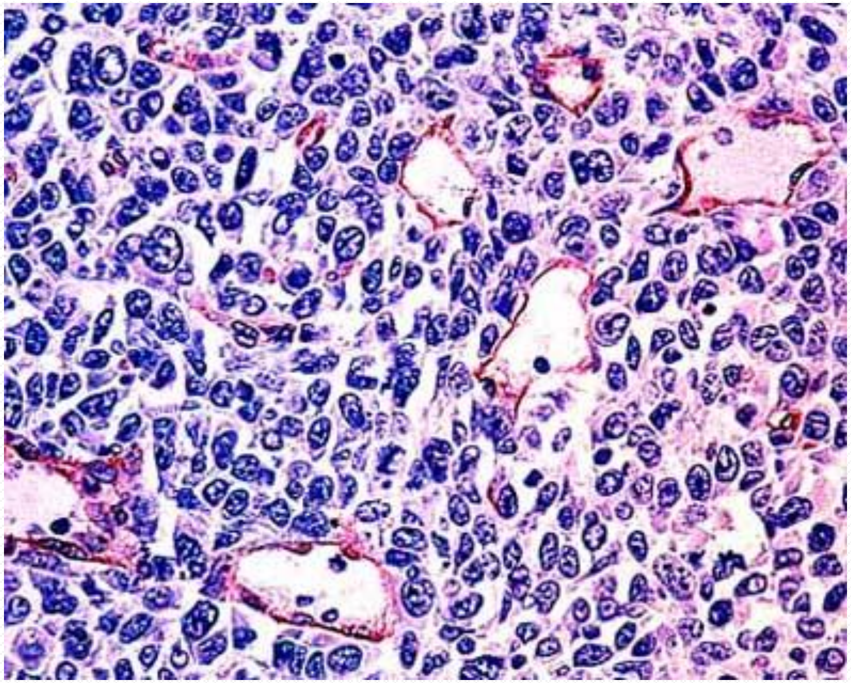



JOHN MAYNARD SMITH & EÖRS SZATHMÁRY
Copyrighted Material
THE MAJOR TRANSITIONS IN EVOLUTION

1. Replicating molecules -> Compartments
2. Independent replicators -> Chromosomes
3. RNA -> DNA + Protein
4. Prokaryotes -> Eukaryotes
5. Asexual clones -> Sexual populations
6. Protists -> Multicellular organisms
7. Solitary individuals -> Colonies
8. Primate societies -> Human language

Each Level Needs Mechanisms to Create Cooperation Among Its Parts

Multicellular Organisms



Danger: Cancer

Solution: Immune System

Human Society

Danger: Criminals

Solution: Police and Courts

Rational Economic Behavior

Universal optimal intelligence algorithm to achieve well-defined goals :

1) Simulate each possible action
2) Choose the action most likely to reach the goal
3) Update the world model based on what actually happens

Formally:

Preferences: *utility function U(h)*

Beliefs: *subjective probability P(h)*

Act to maximize expected utility

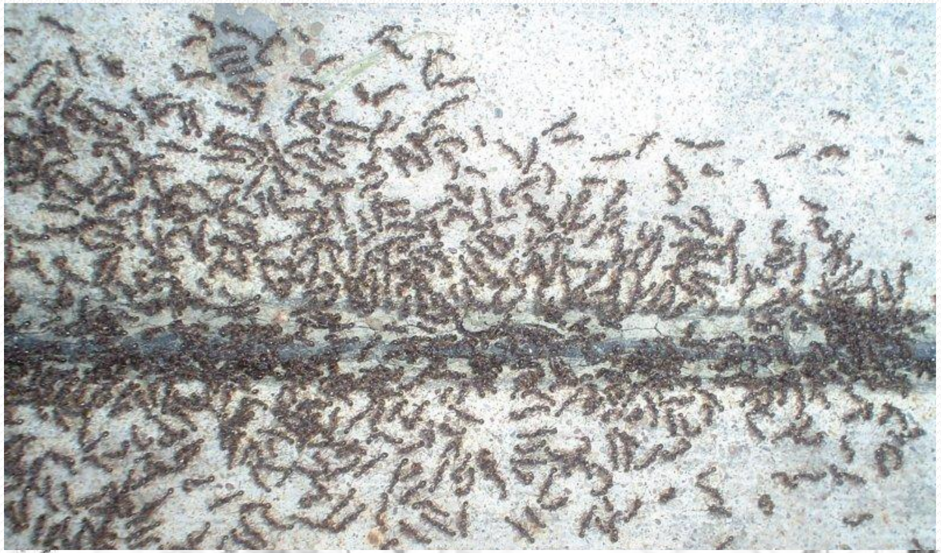
Update P using Bayes' theorem:

$$P(h|d) = \frac{P(d|h)\cdot P(h)}{\sum_h P(d|h)\cdot P(h)}$$

Fully Rational Behavior is too expensive

- Samuel’s checker program
- Truncate deliberative search and use a learned model
- Simplify the state space
- Limit is reinforcement learning TD-lambda or Q learning (state s, reward r, discount g, new state s’):
- V(s)<- V(s)+a(r+gV(s’)-V(s))

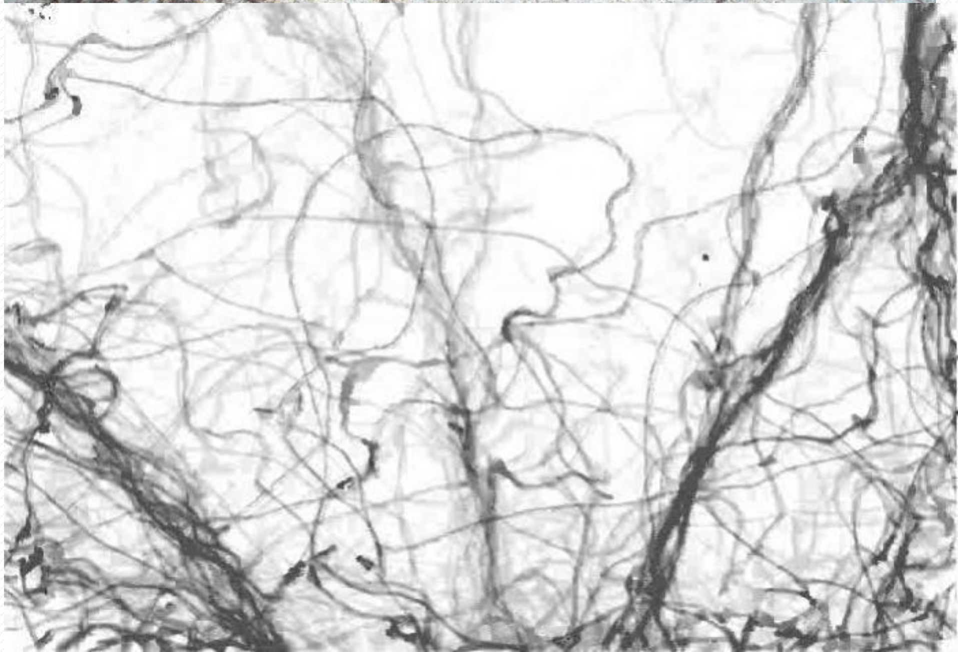
Approximate Rational Behavior



1. A source of diversity
2. A selection mechanism
3. An updating mechanism

Strengthen successful paths.

(evolution, development, ecosystems, economies, bee hives and ant hills, immune systems, brains, animal physiology, cell physiology)



Evolution

State: Distribution of genes in the gene pool.

Simple

1. Random mutation
2. Natural selection
3. Differential reproduction

Adaptive

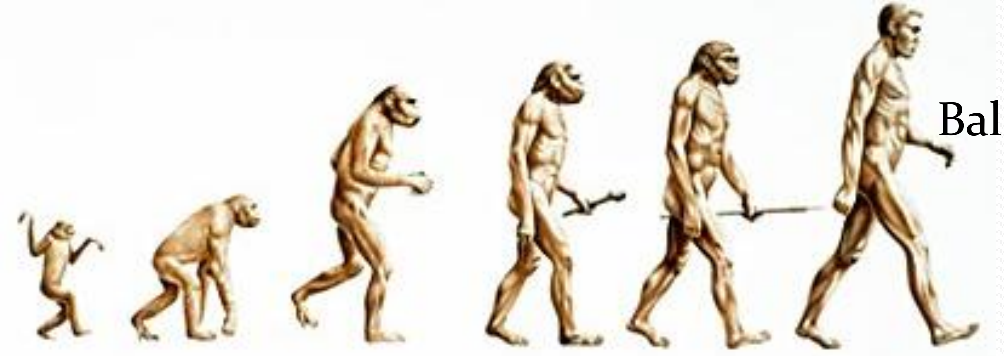
Directed mutation:

Induced global mutation: bacteria

Local hypermutation: Haemophilus Influenzae

Induced local mutation: E. Coli

Induced regional mutation: Brassica nigra



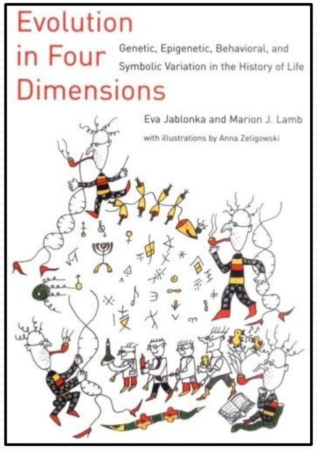
Smarter selection:

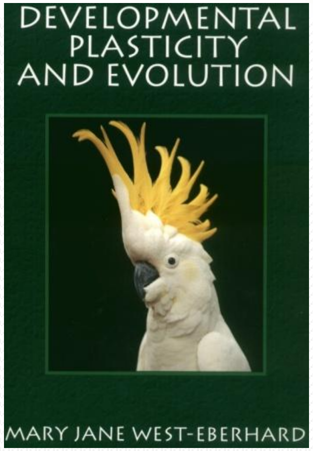
Baldwin effect: downloading learning

Deliberative Baldwin effect

Sexual selection

Interactions with development


Evolution in Four Dimensions
Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life
Eva Jablonka and Marion J. Lamb
with illustrations by Anna Zeligowski



Smarter updating:

Meta-evolution – evolving to evolve

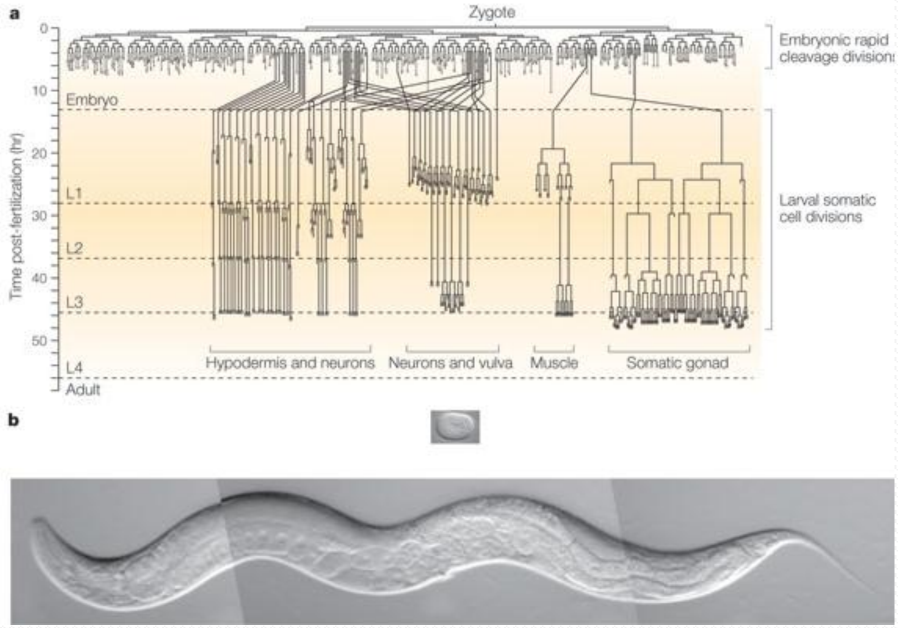
Epigenetic change

Development

State is number and location of different cell types.

Simple

C. Elegans: 959 cells



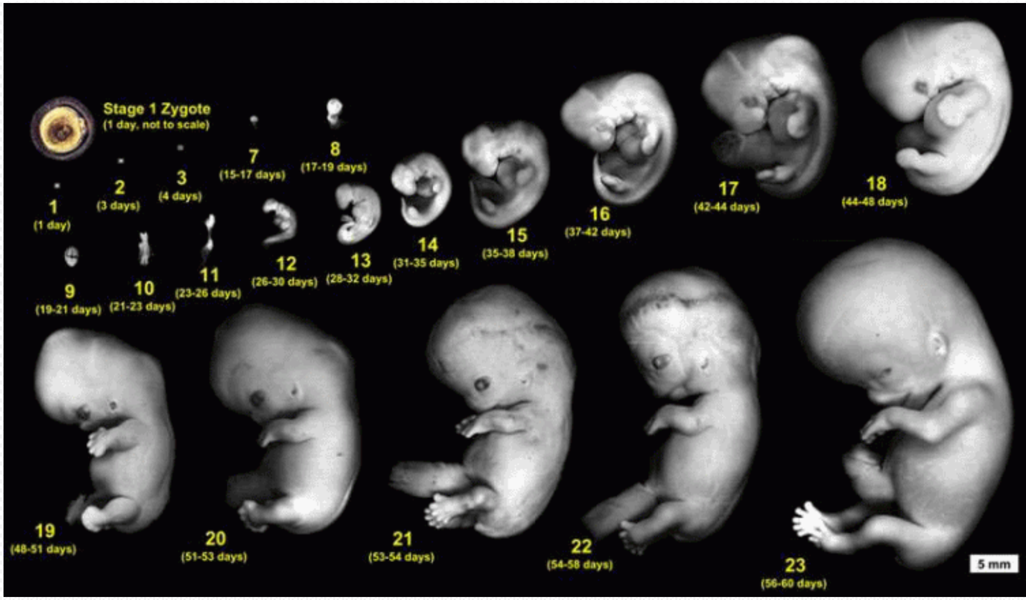
Adaptive

Plasticity

3x Dieback

Neurons that get trophic factors survive.

Extra limbs work!



Immune Systems

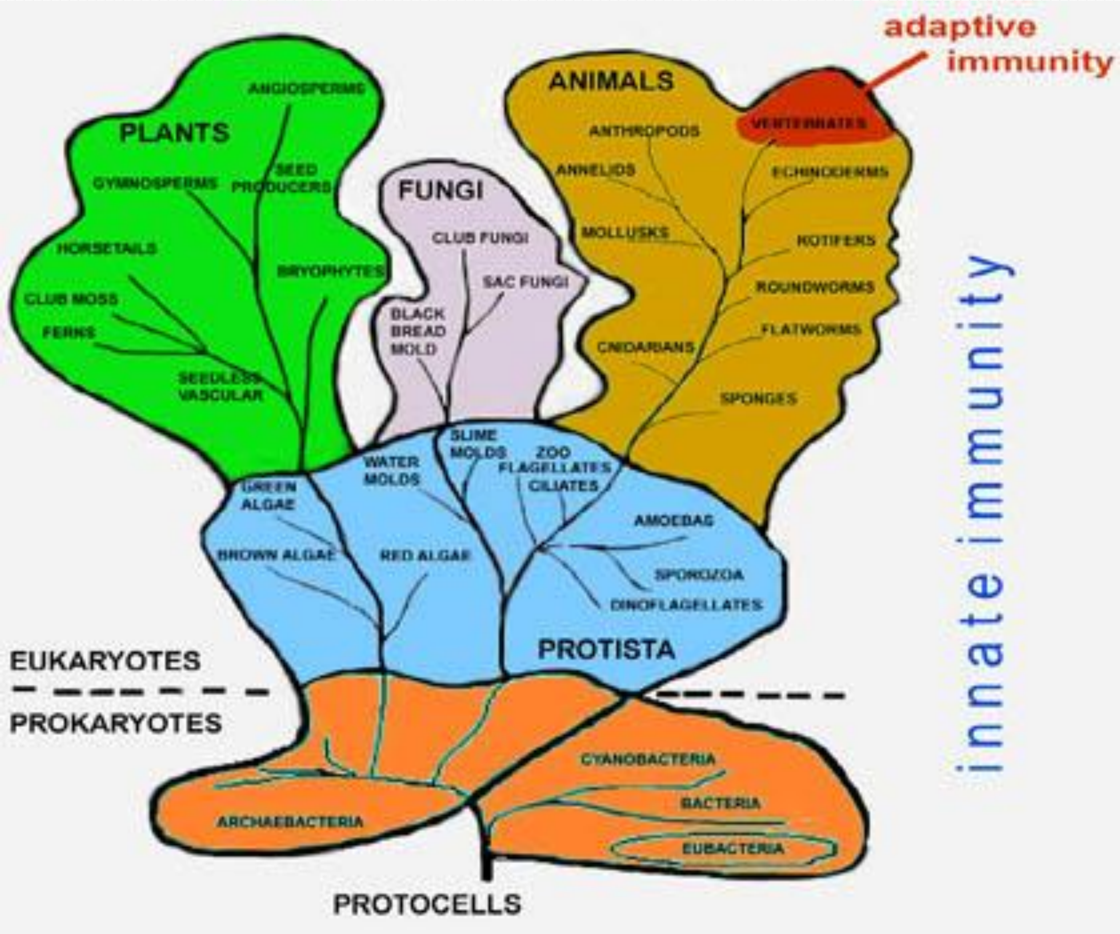
State is distribution of immune cells.

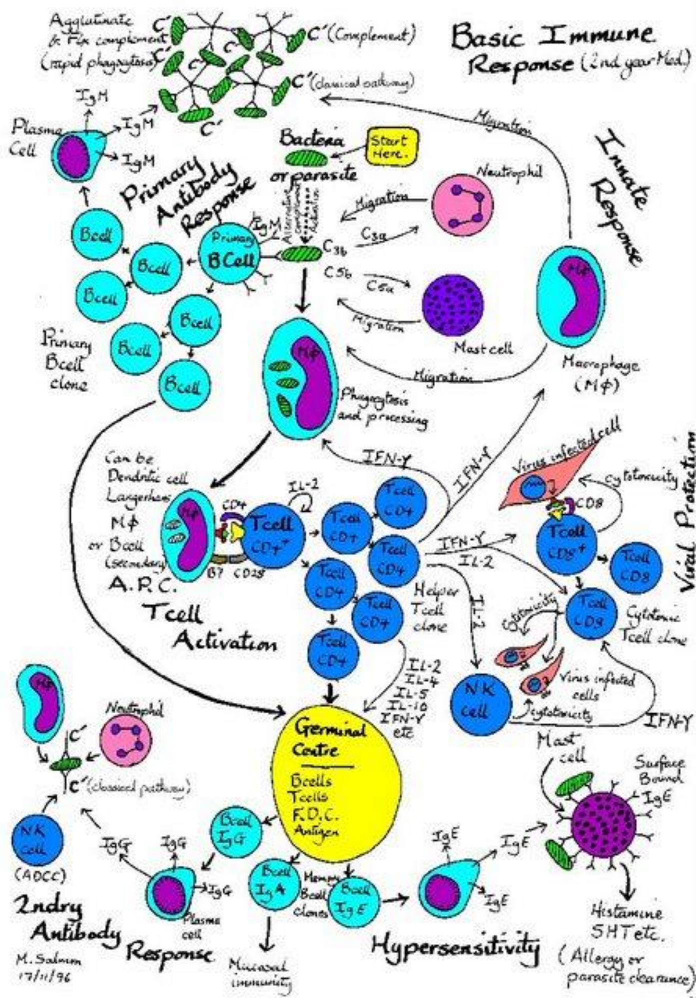
Simple

Invertebrates
Innate Immunity
phagocytic amebocytes

Adaptive

Vertebrates
Innate and Adaptive Immunity
Successful antibodies reproduce





Brains

State is neural activity.

Simple

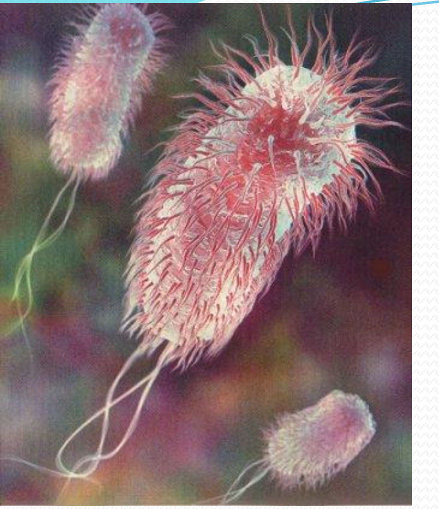
No deliberation
Rigid memory
Rigid learning if any
Situation/Action Rules

Adaptive

Deliberation
Memory
Learning

Hebb rule: strengthen successful activity

Physiology



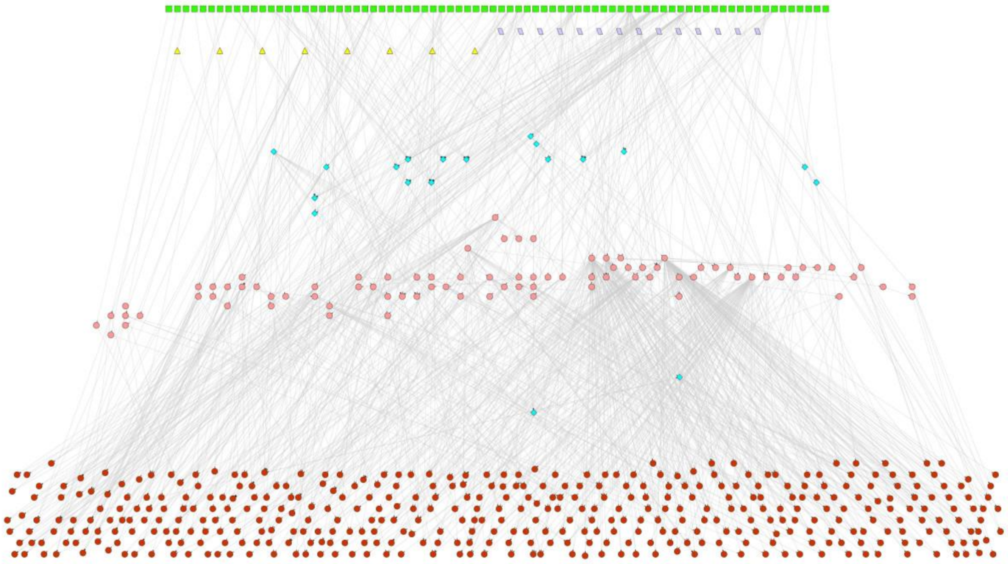
Simple

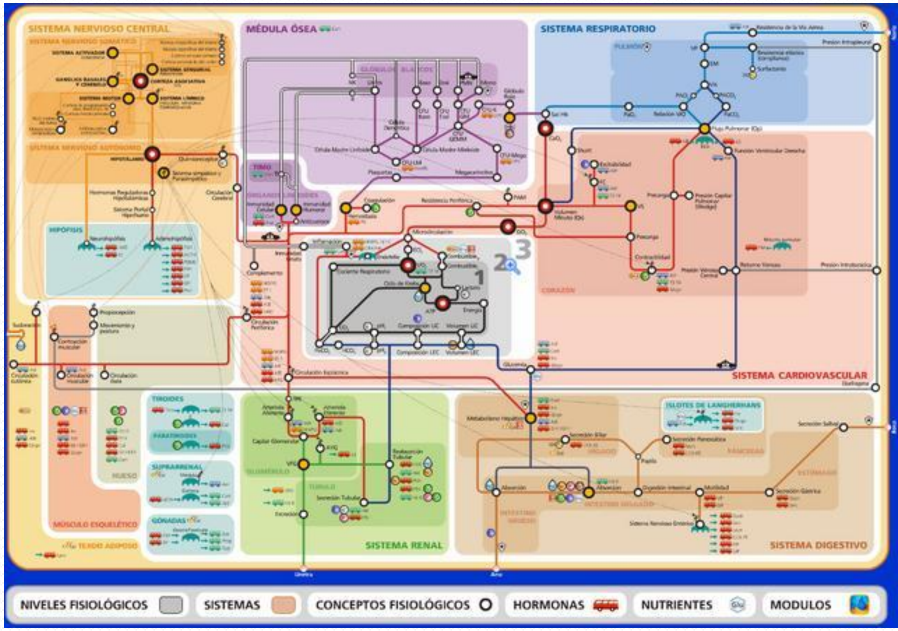
E. Coli
Gene Regulation Networks

Adaptive

Human Physiology
Hormone networks

Strengthen muscle and bone that gets used



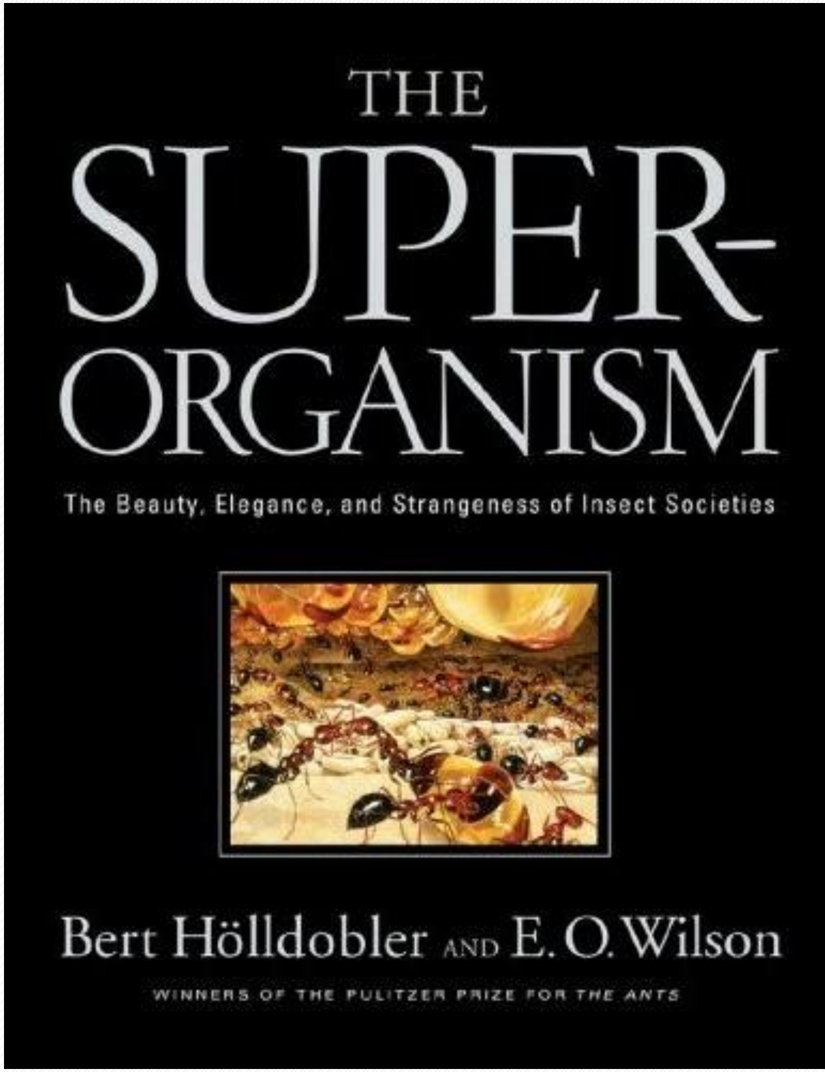


The Beehive as Organism

Individual bees can’t survive
Beehive is “warm blooded”:
    Bees shiver if too cold
    Spread water if too warm
Castes are like organs
Queen is like ovaries
Bee type is like cell type
Decision making on response
Hive cognition
Reproduction like mitosis
Dance like neural firing

Groups and Individuals

- Individual preferences + social contract give rise to group preferences
- Stable social contracts include enforcement mechanisms
- So group preferences shape individual preferences
- Evolutionary pressure to fit in
- Only partial incorporation of group mind
- Interests are not necessarily aligned (eg. cooperation)
- There exist social contracts that go against every member's preferences


THE
SUPER-
ORGANISM
The Beauty, Elegance, and Strangeness of Insect Societies
Bert Hölldobler AND E. O. Wilson
WINNERS OF THE PULITZER PRIZE FOR THE ANTS

Group vs. Individual Conflicts

- Tragedy of the commons – eg. overfishing
- Externalities – eg. pollution
- Proliferation – eg. cancer, population control
- Equality – eg. income disparity
- Damage due to competition – eg. war, fighting
- Signalling costs – eg. conspicuous consumption

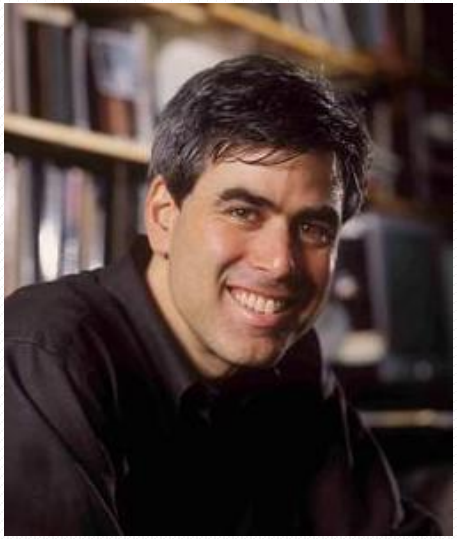
Group cooperation mechanisms

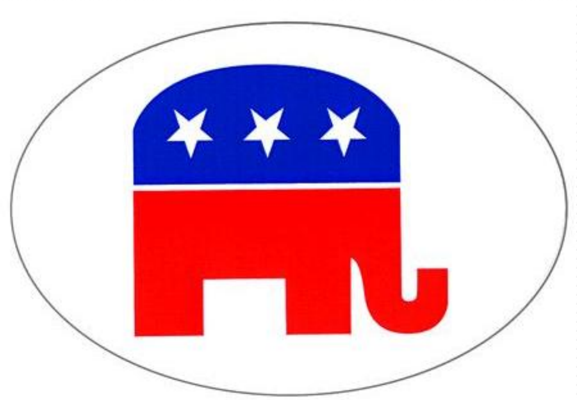
- Immune system – eg. cancer
- Police system – eg. property rights
- Legal system – eg. contracts
- Mutually Assured Destruction – eg. nuclear detente
- Moral code – eg. murder
- Social stigma – eg. sociopathic behavior
- Social rewards – eg. heroes
- Altruism - eg. rescuing strangers
- Membership – eg. in families, churches, countries

Bee mind vs. Hive mind

Humans: Ego and Social Mind

Haidt: 5 Moral Emotions

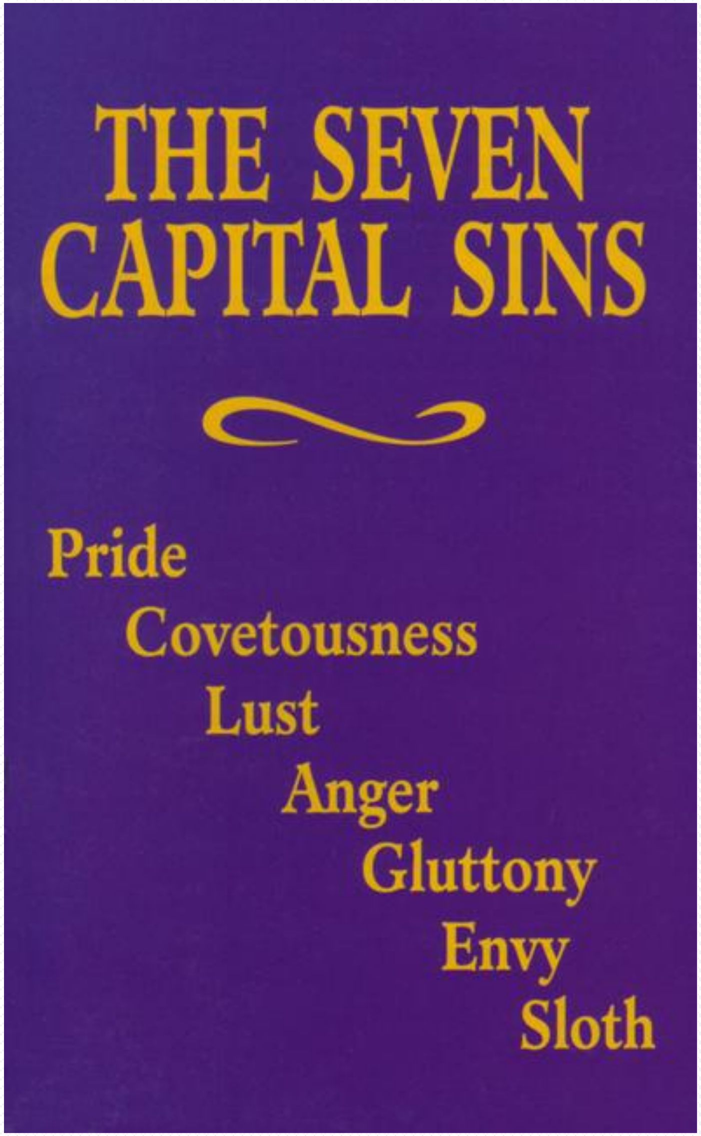




Non-harming
Fairness

Non-harming
Fairness
Loyalty
Respect for authority
Purity or sanctity

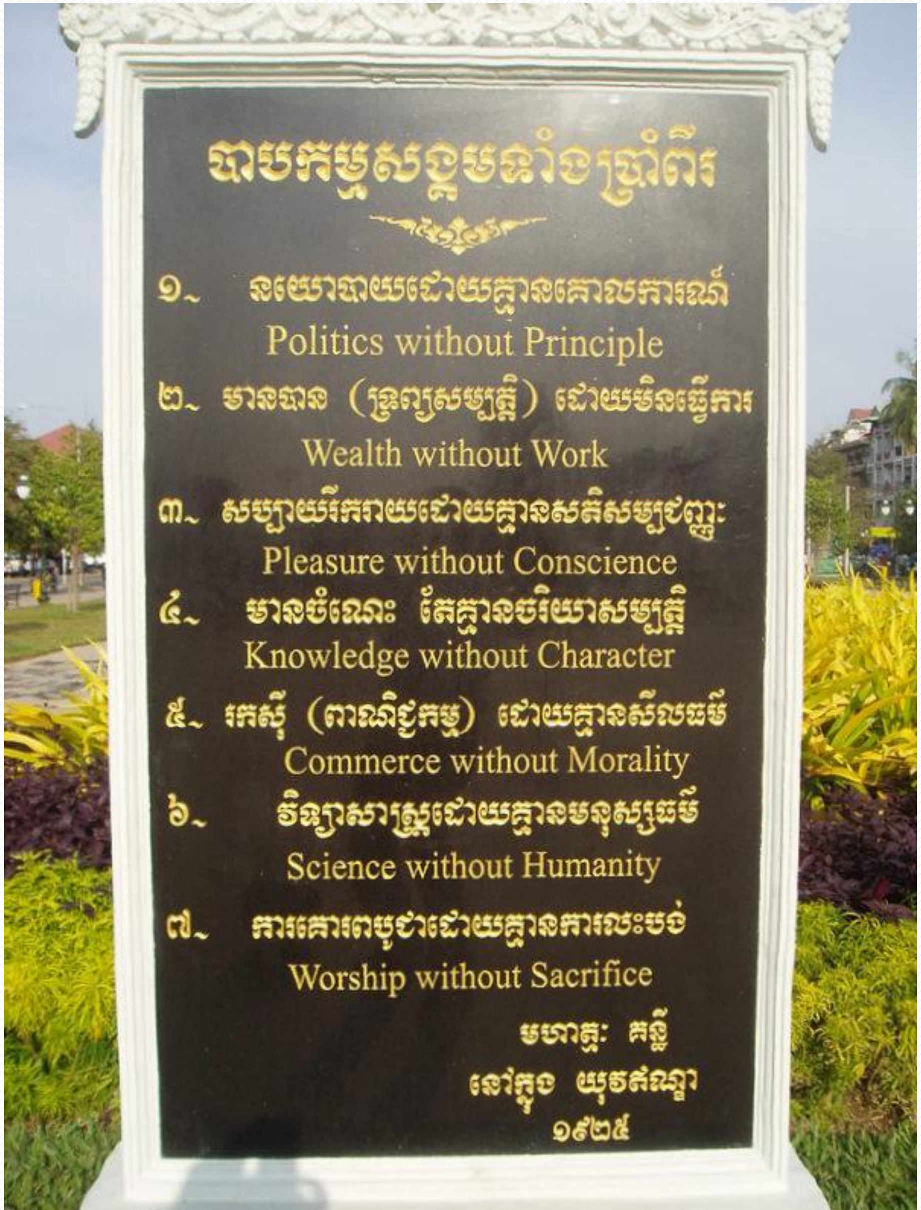
Seven Deadly Sins



THE SEVEN CAPITAL SINS

Pride
Covetousness
Lust
Anger
Gluttony
Envy
Sloth

Ghandi’s Updated Seven Sins



បាបកម្មសង្គមទាំងប្រាំពីរ

១~ នយោបាយដោយគ្មានគោលការណ៍
Politics without Principle

២~ មានបាន (ទ្រព្យសម្បត្តិ) ដោយមិនធ្វើការ
Wealth without Work

៣~ សប្បាយរីករាយដោយគ្មានសតិសម្បជញ្ញៈ
Pleasure without Conscience

៤~ មានចំណេះ តែគ្មានចរិយាសម្បត្តិ
Knowledge without Character

៥~ រកស៊ី (ពាណិជ្ជកម្ម) ដោយគ្មានសីលធម៌
Commerce without Morality

៦~ វិទ្យាសាស្ត្រដោយគ្មានមនុស្សធម៌
Science without Humanity

៧~ ការគោរពបូជាដោយគ្មានការលះបង់
Worship without Sacrifice

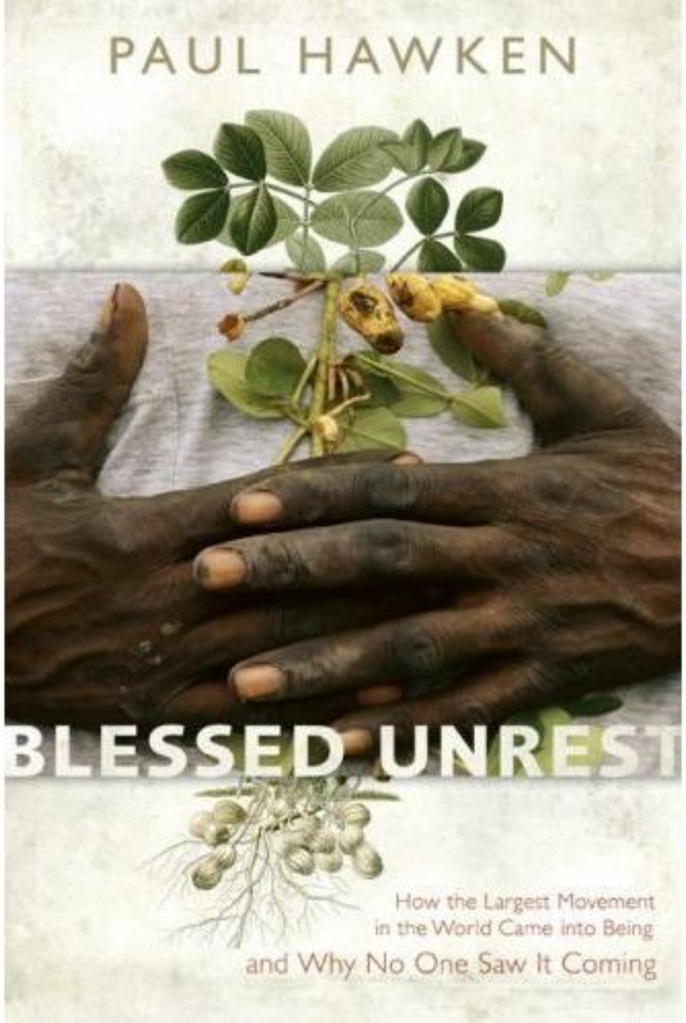
មហាត្មៈ គន្ធី
នៅក្នុង យុវឥណ្ឌា
១៩២៥

1971 Kohlberg: 6 stages of morality

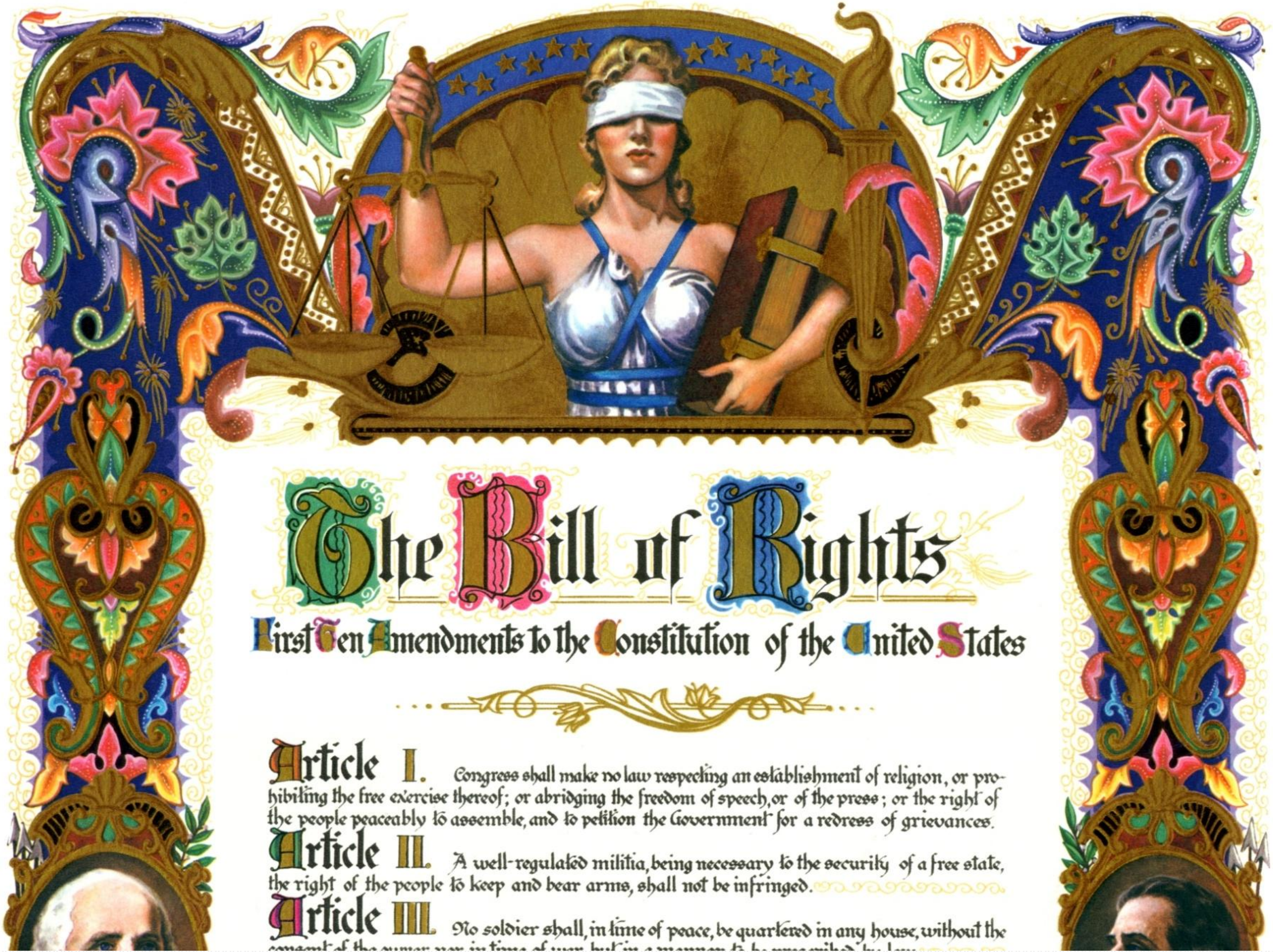
1. Avoiding punishment
2. What's in it for me?
3. Being a good boy
4. Obeying the law
5. Upholding the social contract
6. Universal ethical principles
7. Transcendental morality?

Recent Human Moral Evolution

- Slavery
- Torture
- War crimes
- Women’s rights
- Racial equality
- Animal rights
- Ecological movements
- Sustainability
- …



Must Choose the Rights We Want



The Bill of Rights

First Ten Amendments to the Constitution of the United States

Article I. Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

Article II. A well-regulated militia, being necessary to the security of a free state, the right of the people to keep and bear arms, shall not be infringed.

Article III. No soldier shall, in time of peace, be quartered in any house, without the consent of the owner, nor in time of war, but in a manner to be prescribed by law.

Social Contract Technology

- Mathematical proof
- Formal contracts and laws
- Provably least restrictive constraints
- Given desired properties generate constraints
- Stability properties
- Revealable source code and utility functions
- Provably limited systems
- Provably limited escrow agents
- Formal Provenance

Monitoring

- Entities monitor each other
- Enforcement by groups of entities
- Must limit the maximum power of individuals
- Must prevent certain kinds of collusion
- Oblivious computing
- Provably limited monitoring: only what’s specified
- Watching the watchers
- Punishing non-punishers

Group Decision Making

- Beyond today’s voting systems: Semantic aggregation
- Formal procedures for changing the social contract
- Formal meta-constraints for stability
- Group shared knowledge

Roadmap from the Present

- We'll need AIs to design these systems
- But we must trust the design AIs!
- Computational hardware provably isolated from its software
- Provably limited manufacturing hardware
- Provably limited software
- Social trust networks
- Incentive design
- Safety monitoring networks

Self-Aware Systems

Semantic Computing Initiative

Cooperative Technology Initiative

Create a Cooperative Future