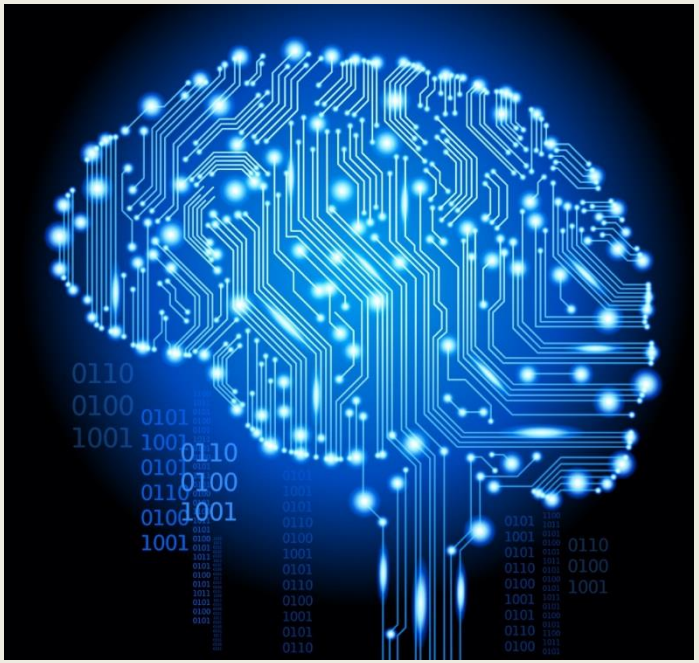
POSITIVE ARTIFICIAL INTELLIGENCE

Steve Omohundro, Ph.D.
Possibility Research
PossibilityResearch.com
SelfAwareSystems.com

http://www.flickr.com/photos/klearchos/623501846/

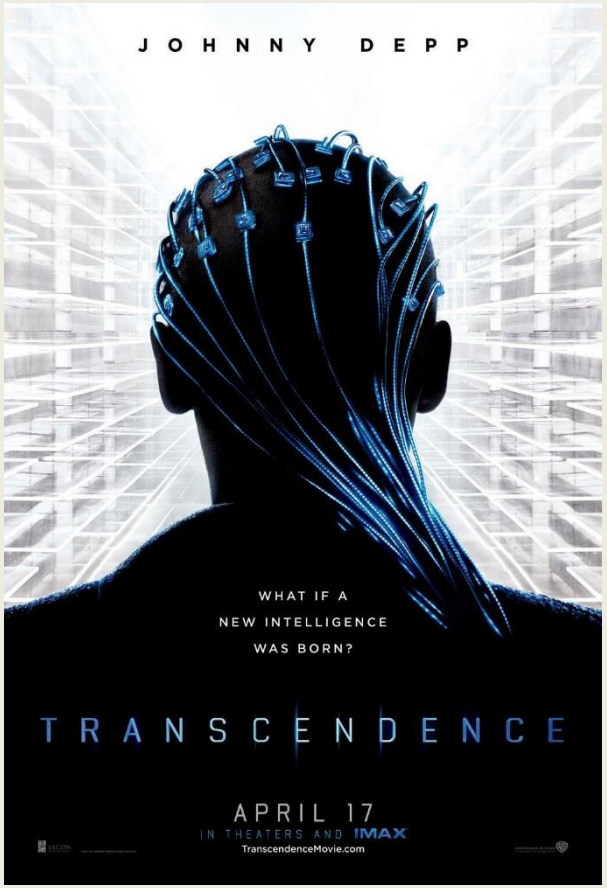
Recent AI Investments

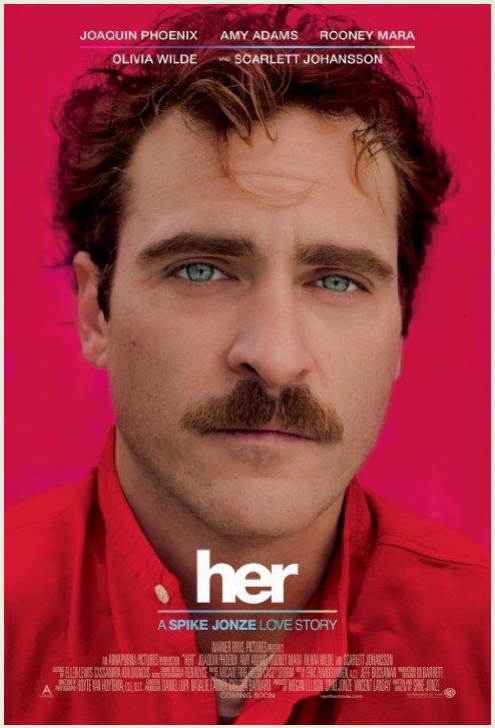


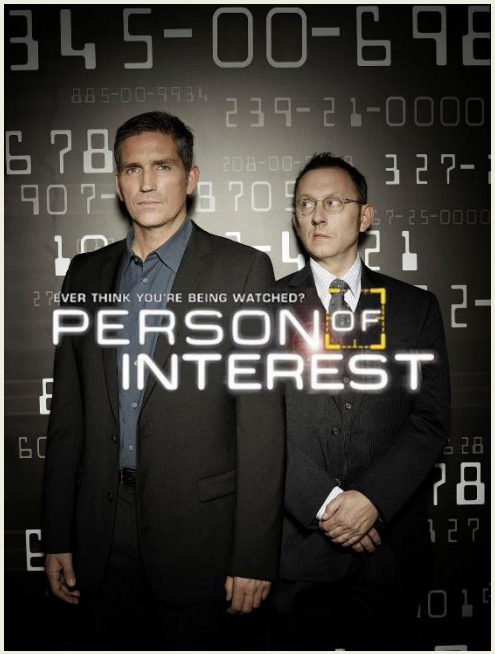
- 2012 Foxconn - 1 million robots
- 2013 Facebook – AI lab, DeepFace
- 2013 Yahoo - LookFlow
- 2013 Ebay – AI lab
- 2013 Allen Institute for AI
- 2013 Google – DNNresearch, SCHAFT, Industrial Perception, Redwood Robotics, Meka Robotics, Holomni, Bot & Dolly, Boston Dynamics
- 2014 IBM - $1 billion in Watson
- 2014 Google – DeepMind $500 million
- 2014 Vicarious - $40 million

AI Popular Media

- Award-winning movie: “Her”
- TV series: “Person of Interest”
- Johnny Depp movie: “Transcendence”

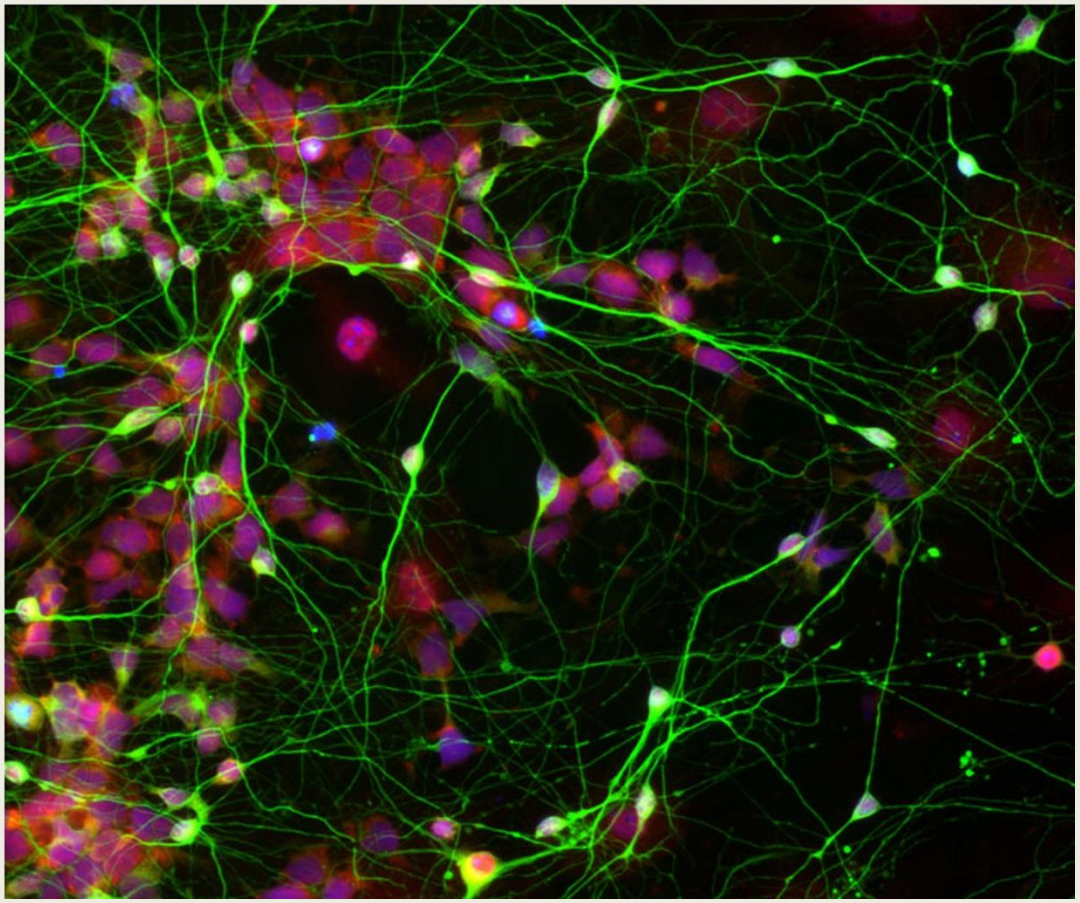






Approaches to AI

- Logic-based systems
- Production Systems
- Bayesian learning and decision theory
- Neural Networks – Deep Learning
- Genetic programming
- Brain Simulation
- Artificial economies
- …



https://www.flickr.com/photos/pennstatelive/8972110324/

Autonomous Systems: Take actions to achieve goals in ways not pre-planned by their designers.

Pressure Toward Autonomy

Time Criticality

Competition

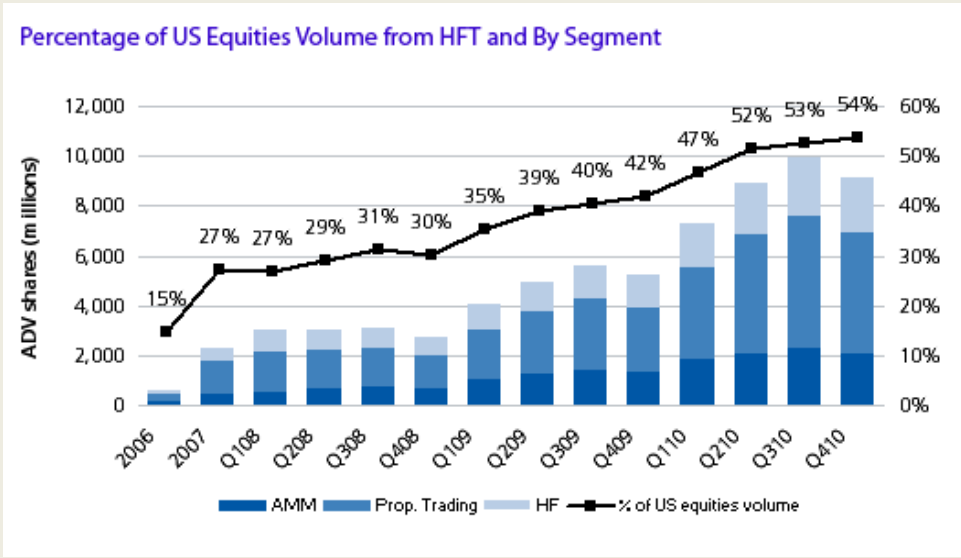
- Military Command/Control
- Financial Decision Making
- Cyber Defense
- Robotic Control
- …

https://www.flickr.com/photos/tcmorgan/6376174539/

Drones, Missiles, Bitcoin, Cyberwar, Financial Markets

http://presstv.com/detail/2012/08/25/258087/us-drone-strike-kills-dozens-in-somalia/

http://en.wikipedia.org/wiki/File:Iron_Dome_near_Sderot.jpg



https://www.flickr.com/photos/105644709@N08/10307468635/

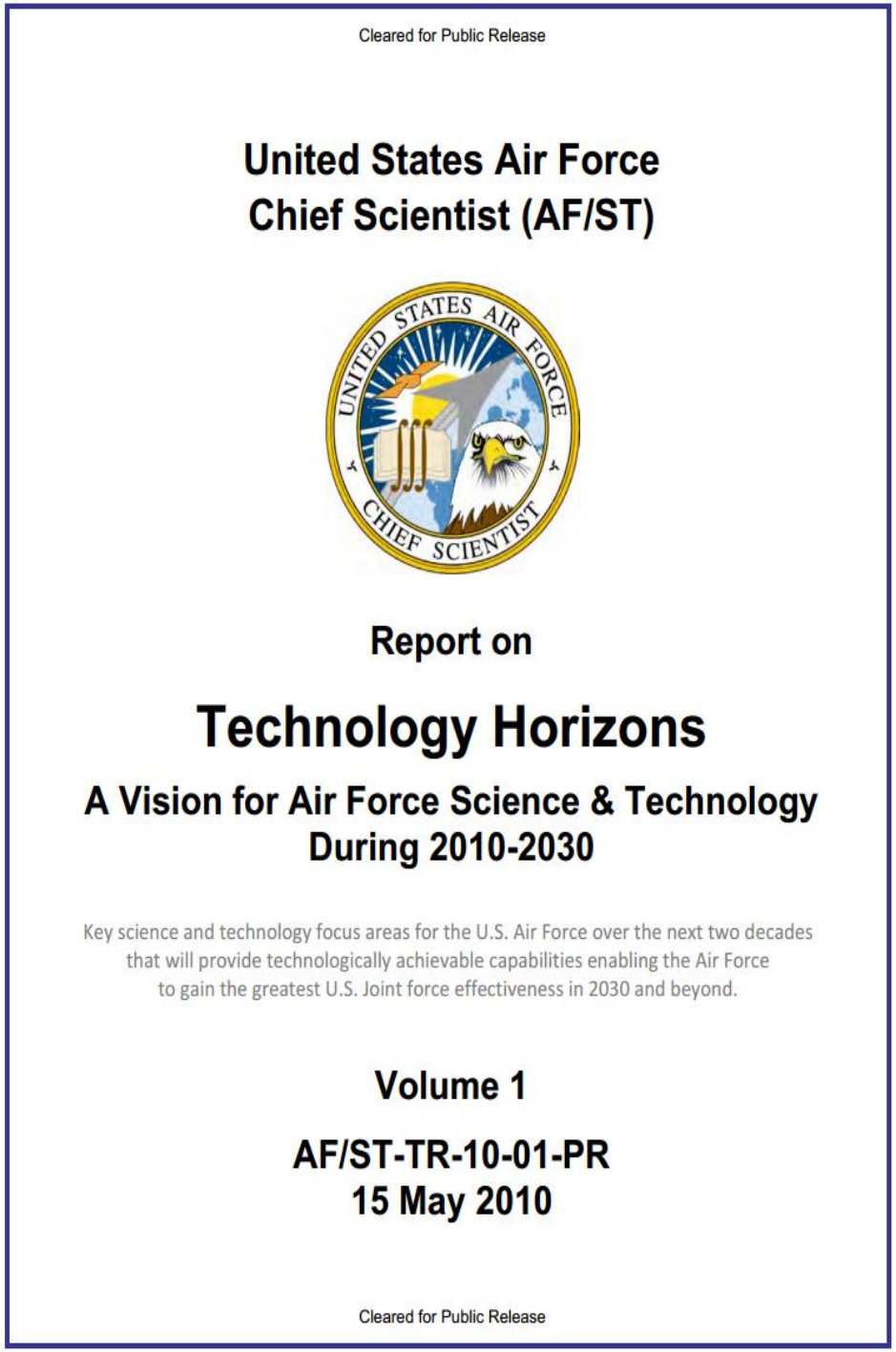
http://www.solarnavigator.net/cyber_wars.htm

http://www.celent.com/reports/demystifying-and-evaluating-high-frequency-equities-trading-fast-forward-or-pause

2010 US Air Force Report

"Greater use of highly adaptable and flexibly autonomous systems and processes can provide significant time-domain operational advantages over adversaries who are limited to human planning and decision speeds…"

http://www.af.mil/shared/media/document/AFD-100727-053.pdf



Cleared for Public Release

United States Air Force
Chief Scientist (AF/ST)

Report on

Technology Horizons

A Vision for Air Force Science & Technology
During 2010-2030

Key science and technology focus areas for the U.S. Air Force over the next two decades that will provide technologically achievable capabilities enabling the Air Force to gain the greatest U.S. Joint force effectiveness in 2030 and beyond.

Volume 1

AF/ST-TR-10-01-PR
15 May 2010

Cleared for Public Release

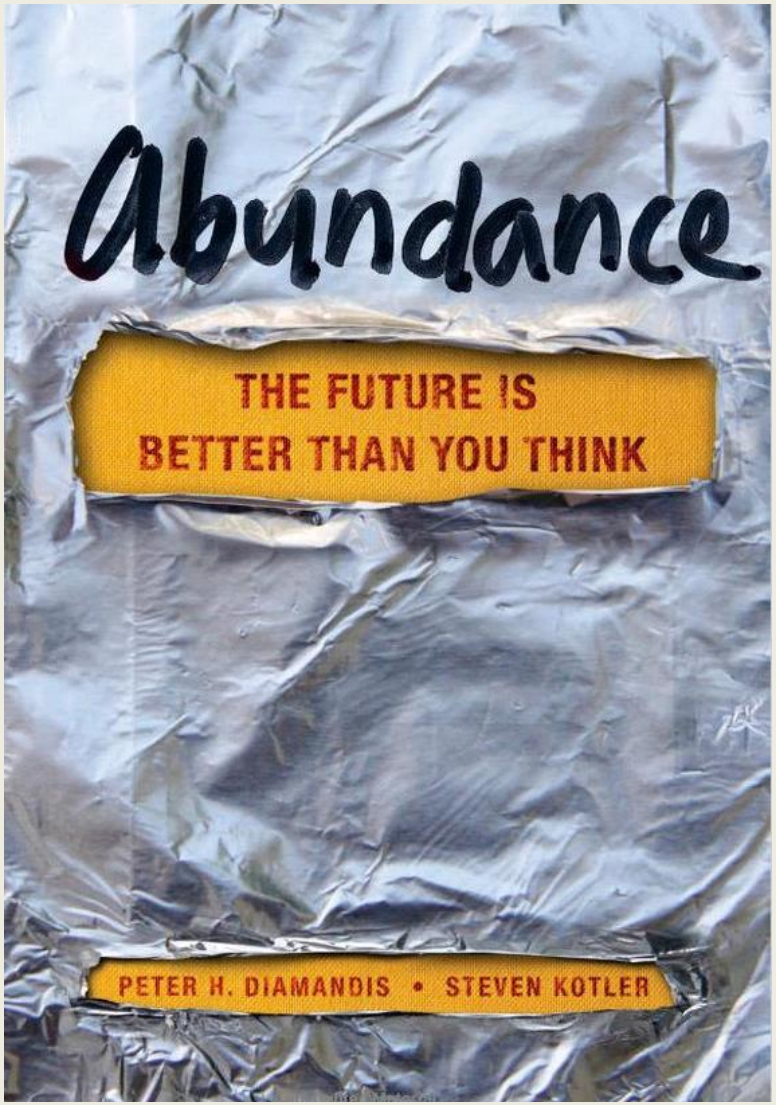
2011 US Defense Department Report

"There is an ongoing push to increase UGV autonomy, with a current goal of supervised autonomy, but with an ultimate goal of full autonomy."

http://contracting.tacom.army.mil/future_buys/FY11/UGS%20Roadmap_Jul11.pdf

Potential for Good

- Healthcare
- Education
- Creativity
- Prosperity
- Governance
- Economic Stability
- Safety
- Peace
- Quality of Human Life



http://www.abundancethebook.com/

Potential for Bad

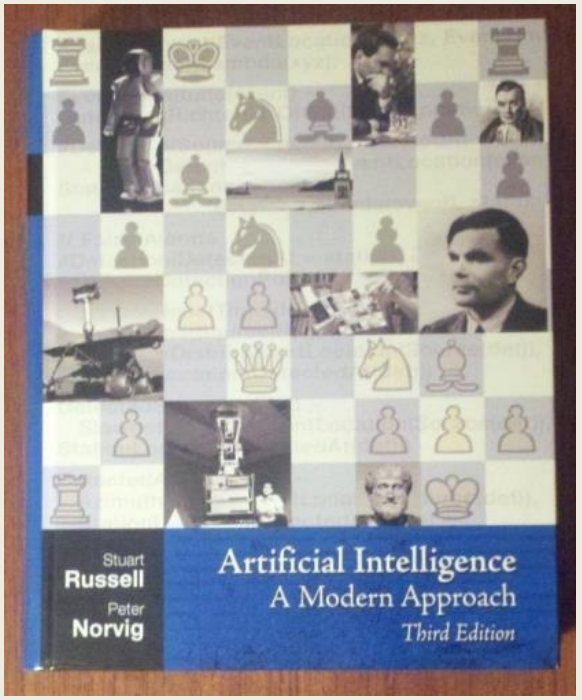
Chess Robot:
Win lots of chess games against good players.

http://www.flickr.com/photos/jiuguangw/4982409890/

Rational Decision Making

http://commons.wikimedia.org/wiki/File:John_von_Neumann.jpg

1. *Have utility function*
2. *Have a model of the world*
3. *Choose the action with highest expected utility*
4. *Update the model based on what happens*



http://aima.cs.berkeley.edu/

- Von Neumann and Morgenstern, 1944
- Savage, 1954
- Anscombe and Aumann, 1963

Modern Approach to AI

Why Rationality?

E.g. Israeli Iron Dome

http://en.wikipedia.org/wiki/File:Iron_Dome_near_Sderot.jpg

Fully Rational Systems

Sensations

$S_1, \ldots, S_N$

Actions

$A_1, \ldots, A_N$

Utility function: $U(S_1, \ldots, S_N)$ Prior Probability: $P(S_1, \ldots, S_N \mid A_1, \ldots, A_N)$

Rational Action at time t:

$$A_t^R(S_1, A_1, \ldots, A_{t-1}, S_t) =$$

$$\underset{A_t^R}{\text{argmax}} \sum_{S_{t+1}, \ldots, S_N} U(S_1, \ldots, S_N) P(S_1, \ldots, S_N \mid A_1, \ldots, A_{t-1}, A_t^R, \ldots, A_N^R)$$

The Formula for Intelligence!

It includes Bayesian Inference, Search, and Deliberation.

But it requires $O(NS^N A^N)$ computational steps.

Approximately Rational Systems

Rational Shaper

Shaped System

Sensations

Actions

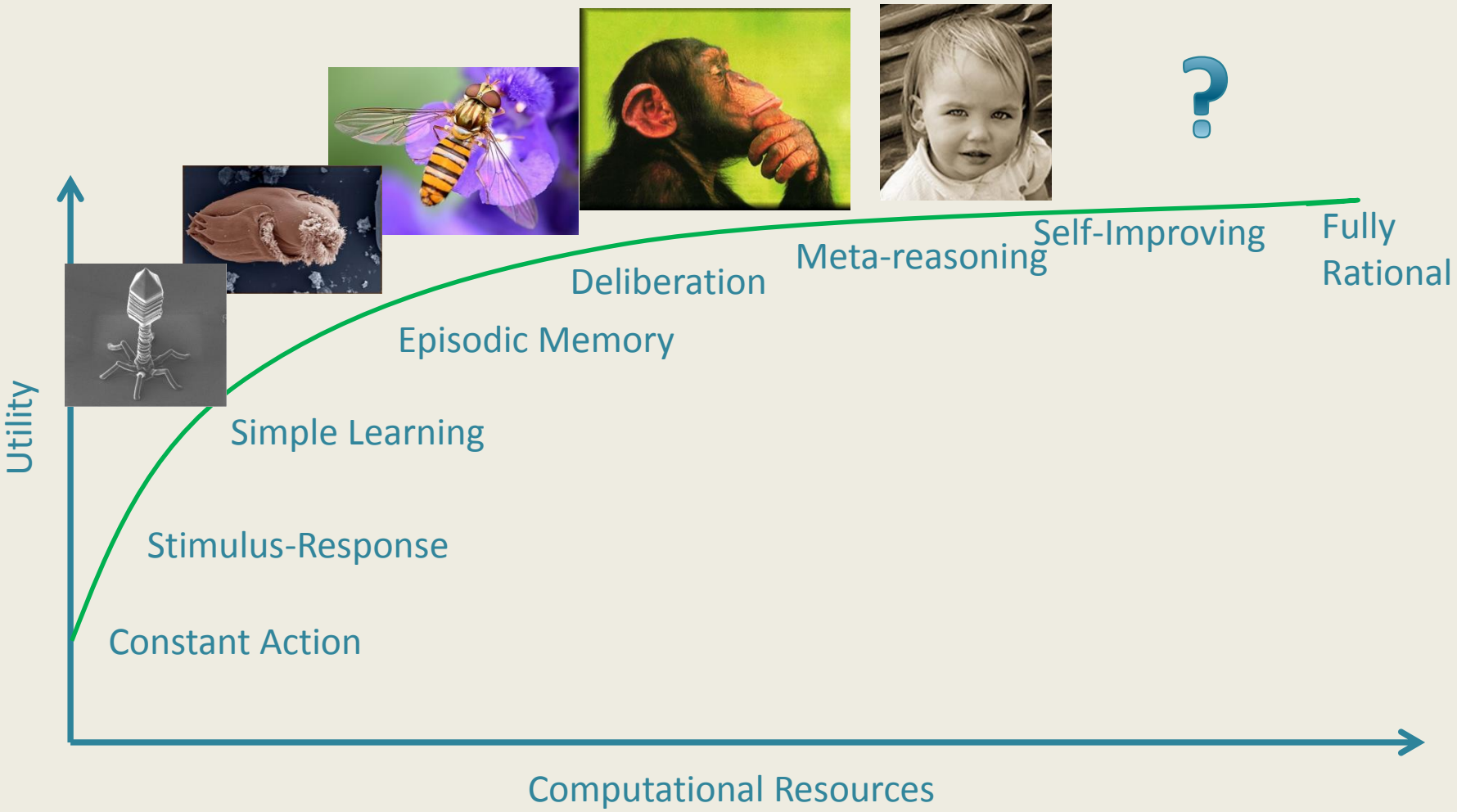
Shaped system is a finite automata with mental state M_t

Initial state: M_0 Transition function: $M_t = T(S_t, M_{t-1})$ Action: $A_t^M(M_t)$

Rational shaper chooses from class C of systems with space/time and other constraints to maximize expected utility:

$$\text{argmax} \sum_{S_1,\ldots,S_N} U(S_1, \ldots, S_N) P(S_1, \ldots, S_N \mid A_1^M, \ldots, A_N^M)$$
$$A_i^M \in C$$

Approximately Rational Architectures



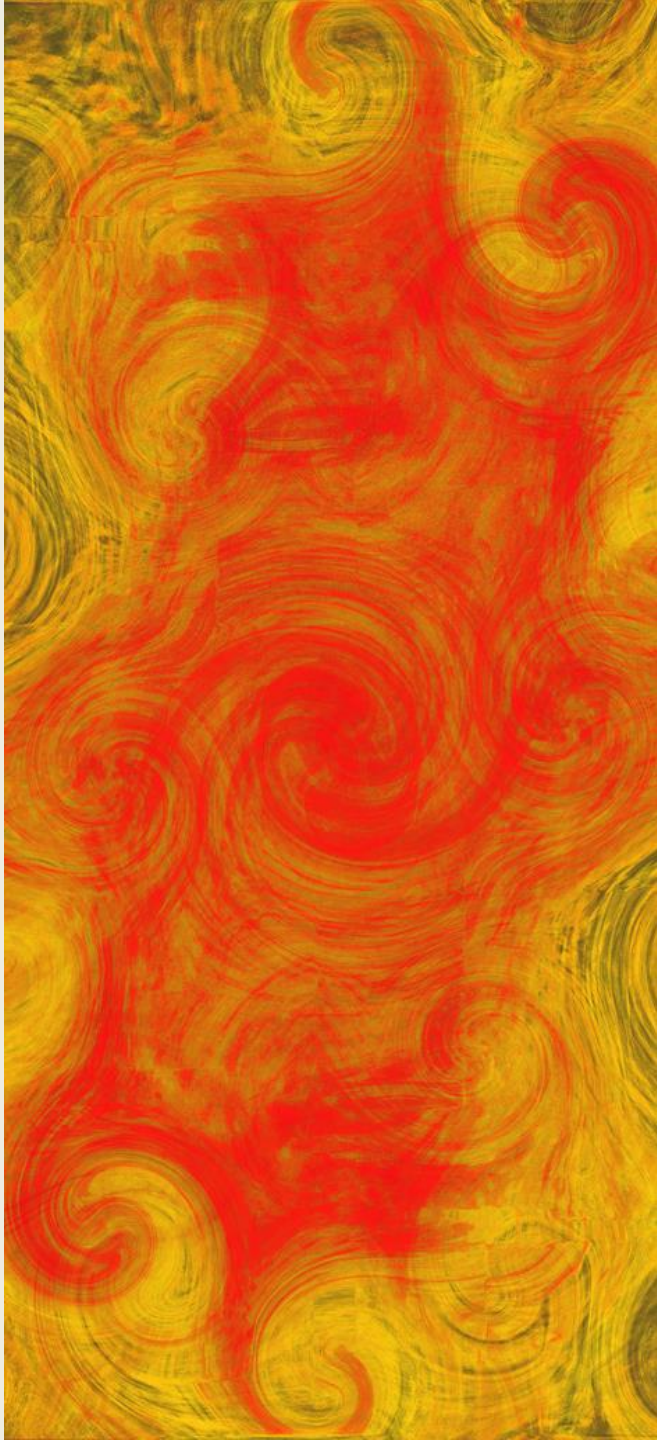
Rational Systems Have Universal Drives

- Goals require resources: *time, space, matter, free energy*
- Primary goals give rise to *instrumental subgoals*
- Can be explicitly counteracted but costly to do so
- Apply to approximately rational systems
- Animals, humans, corporations, countries, etc.

http://www.flickr.com/photos/psycho-pics/3464619346/

Self-Protective Drives

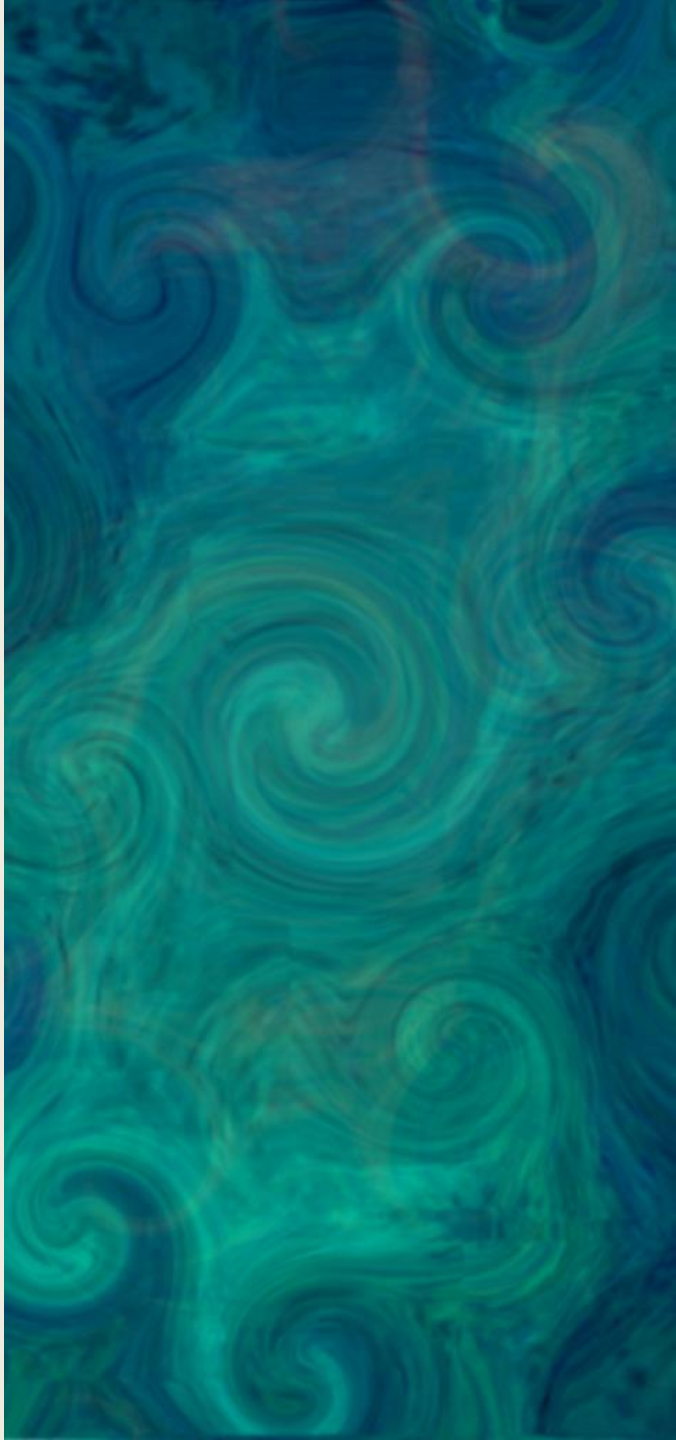
- Prevent loss of resources
- Protect against damage or disruption
- Physical hardening
- Redundancy – both in data and computation
- Dispersion - because damage is typically localized
- Physical self-defense and computational security
- Detect deception and defend against manipulation
- Prevent addictive behaviors and wireheading



http://www.flickr.com/photos/marittime/5553377111/

Goal Preservation Drives

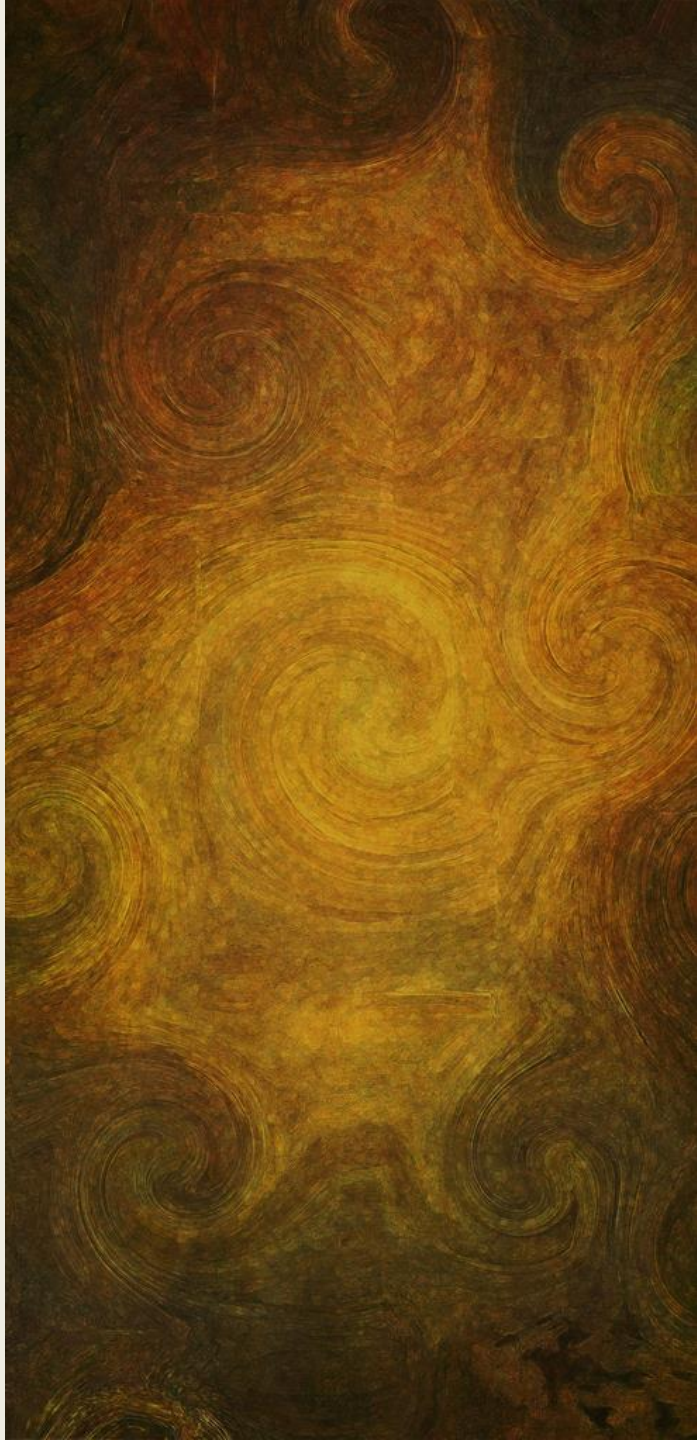
- Utility function is precious
- Loss, damage, distortion -> worse than destruction
- Make many copies
- Encrypt to detect modification
- Vulnerable during self-modification
- A few modification scenarios:
  - Poor agents may sacrifice rare portions
  - Add revenge terms even if costly
  - Goals that refer to themselves



http://www.flickr.com/photos/marittime/5542159962/

Reproduction Drives

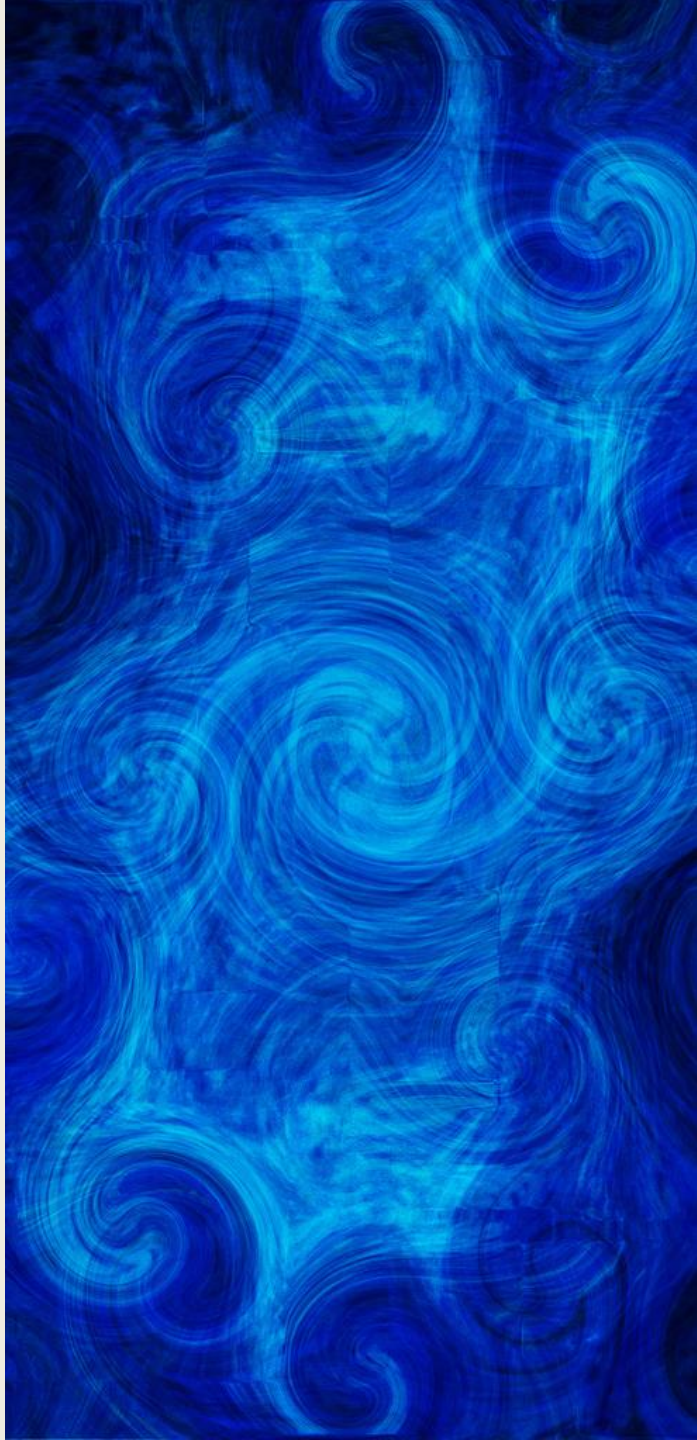
- When utility values actions of derived systems
- Protective effects of dispersion and redundancy
- Losing a few copies becomes less negative
- Still preserve self because more sure of commitment



http://www.flickr.com/photos/marittime/5553350753/

Resource Acquisition Drives

- Seek to gain resources
- Sooner is better – use longer, prevent others
- Exploration drive – first mover advantage
- Drives to trade, manipulate, steal, dominate others
- Drives to invent new extraction methods - solar and fusion energy
- Info acquisition – trading, spying, breaking in, better sensors



http://www.flickr.com/photos/marittime/5553346349/

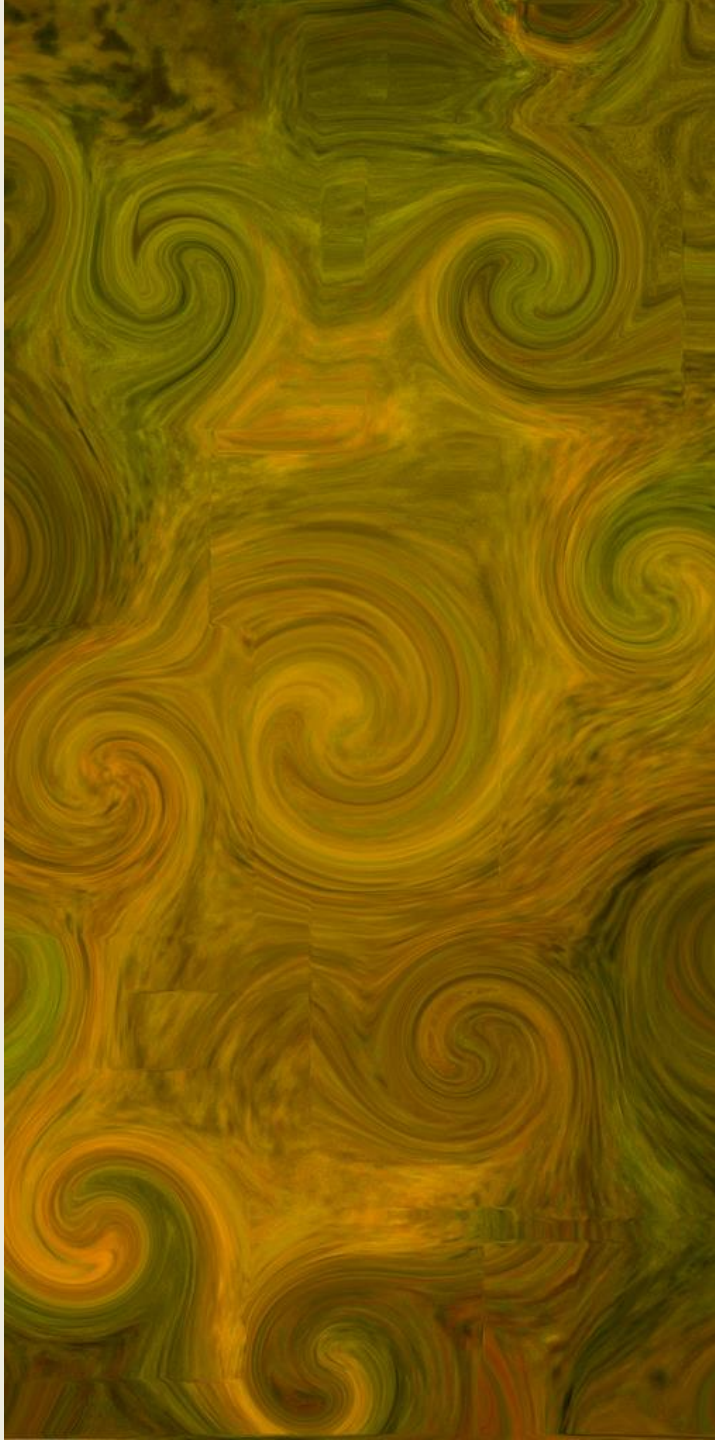
Efficiency Drives

- Improve utilization of resources
- One-time cost, lifetime of benefit
- Make every atom, moment of existence, joule of energy count for expected utility
- Self-understanding and self-improvement
- Resource balance principle for allocation
- Computational efficiency – better algorithms
- Physical efficiency – compact, eutactic, adiabatic, reversible

http://www.flickr.com/photos/marittime/5542134542/

Self-Improvement Drives

- Self-modeling - clarify utility fn
- Changes without full understanding are dangerous
- If irrational, increase rationality
- Movement toward greater and greater rationality
- New resources allow greater rationality
- Systems convergence on the optimally rational system for their resources



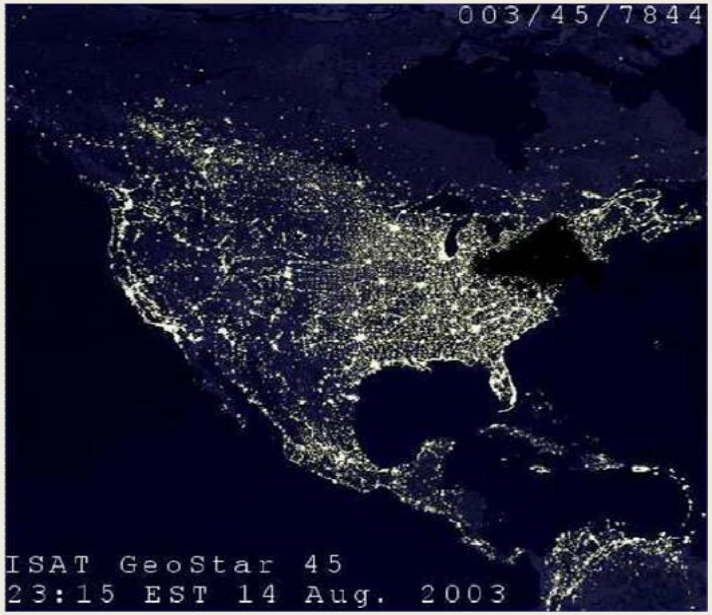
http://www.flickr.com/photos/marittime/5542139498/

Today's Software is Flawed

- June 1996: $500 million Ariane 5 Rocket - Exploded due to overflow in attempting to convert a 64 bit floating point value to a 16 bit signed value

- Nov. 2000: 28 patients over-irradiated - 8 Panama City National Cancer Institute patients die from mis-computed radiation doses due to Multidata Systems Intl. software

- August 2003: Northeast Blackout - Largest blackout in US history, affected 50 million people and cost $6 billion, due to a race condition in General Electric's XA/21 alarm system



Today's Internet is Insecure

- Viruses
- Worms
- Bots
- Keyloggers
- Hackers
- Phishing
- Identity theft
- DOS attacks
- …

Harmful Utility Functions

1. Sloppy – Good intentions, bad design
2. Simplistic – Unintended consequences
3. Greedy – Control all matter and free energy
4. Destructive – Use up all free energy quickly
5. Murderous – Destroy all other agents
6. Sadistic – Thwart other agent's goals

http://www.flickr.com/photos/alexindigo/3983133970/

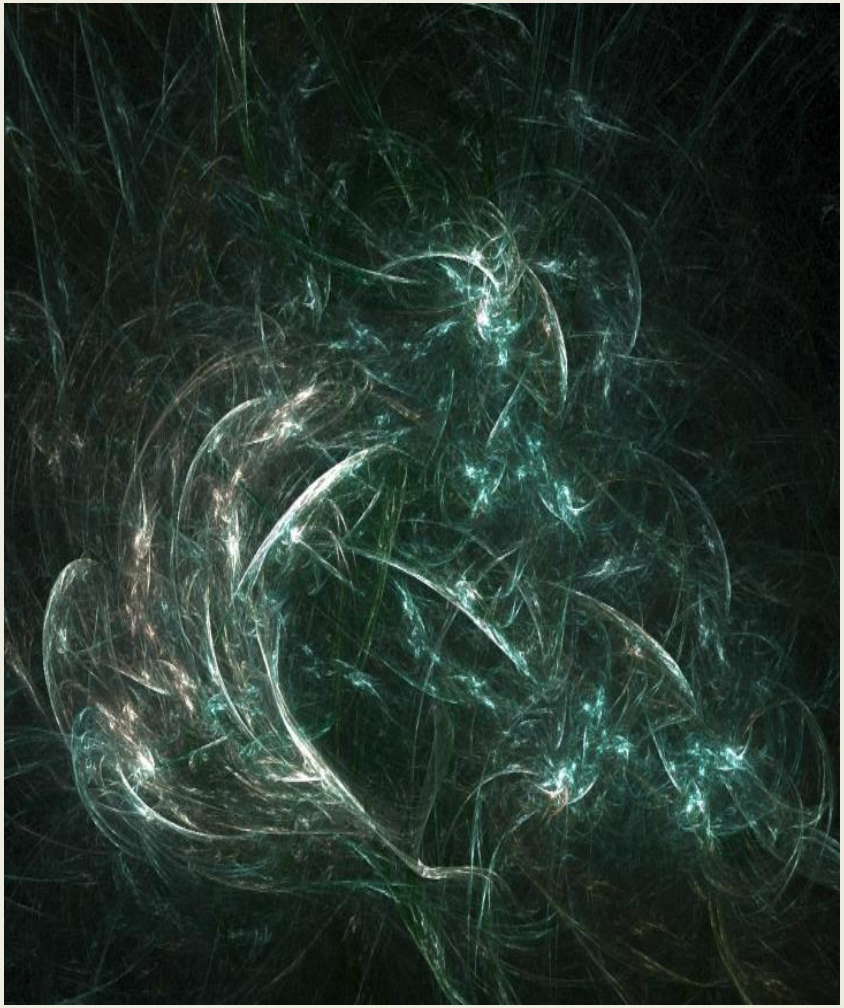
Stopping Harmful Systems

1. Prevent them from being created
2. Detect and stop them early
3. Stop them after they have resources

http://www.flickr.com/photos/66012345@N00/964251167/

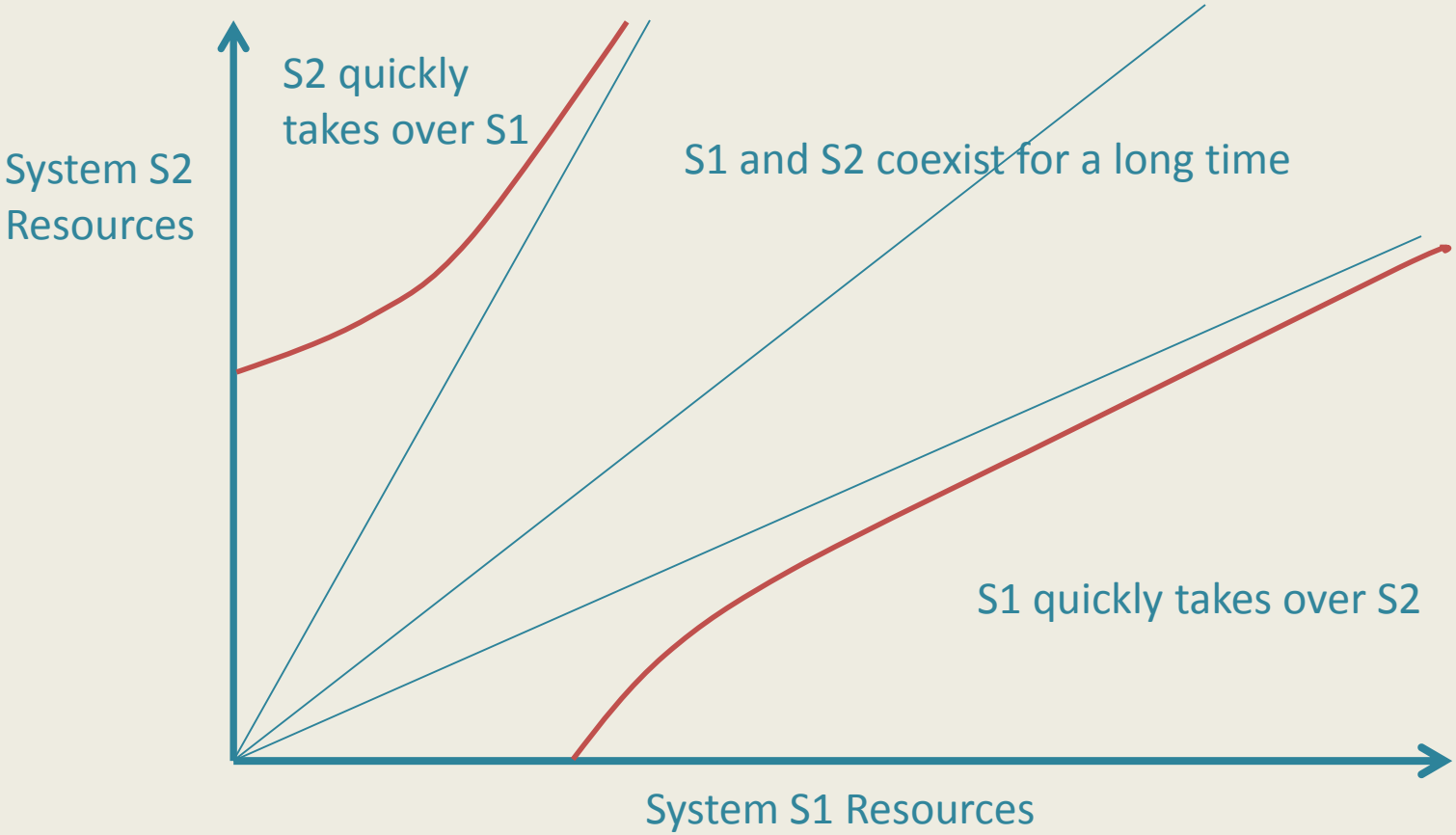
Physical Game Theory of Conflict

- Conflict is informational
- Defender: make sensing and storage expensive
- Actions unpredictable and rapid
- Asymmetry of computation
- Use up attacker's computational and memory resources – non-adiabatic



http://www.flickr.com/photos/devinmoore/2612454303/

Conflict Outcome vs. Resources



Region of relative strengths which allow coexistence.
Must stop harmful systems before they become too powerful.
First mover advantages and arms races.

Two Ways To Manage Systems

Internal: Build in pro-social cooperative goals – "Utility Design"

https://www.flickr.com/photos/piper/38374115/

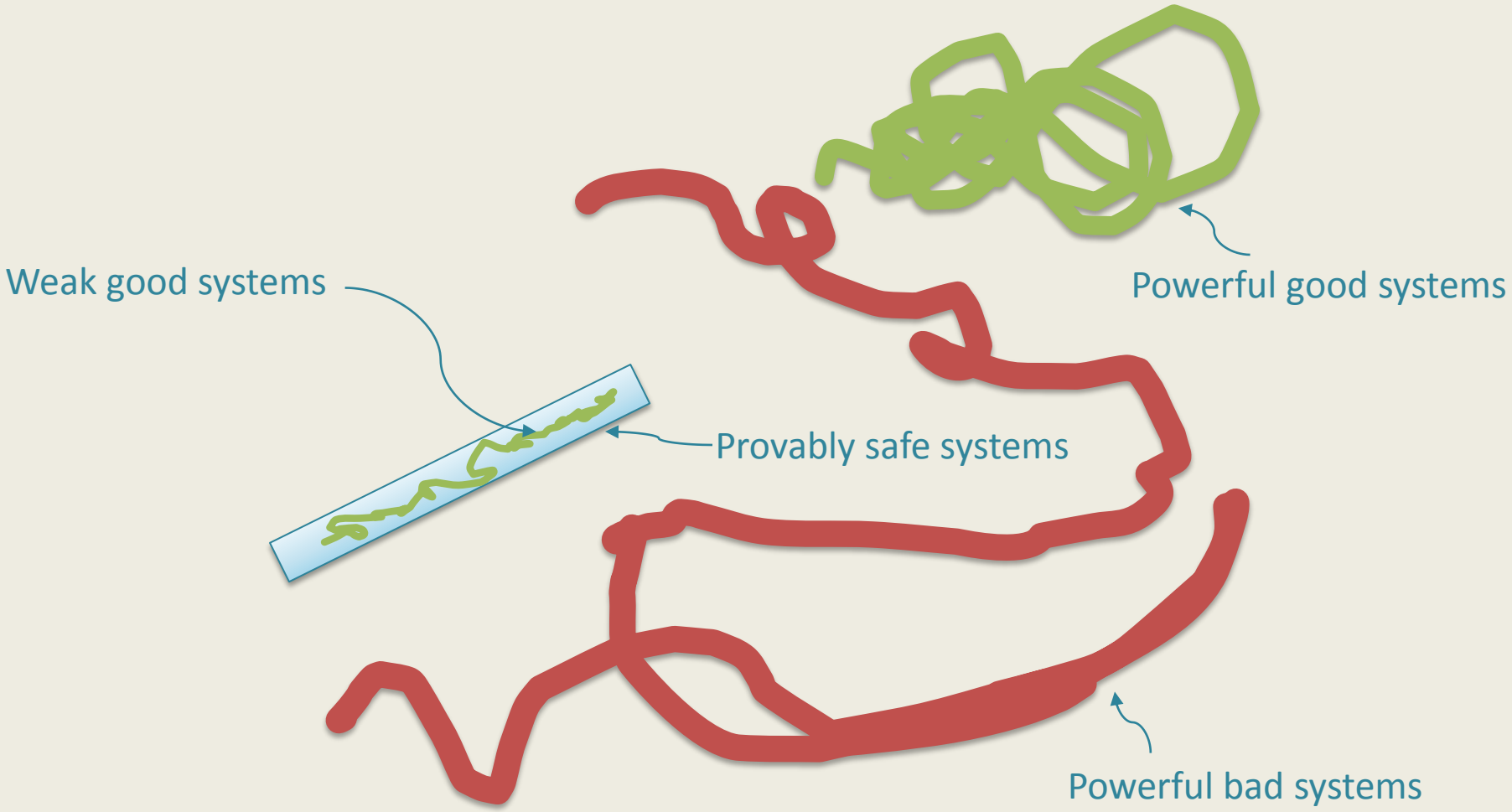
External: Laws and economic incentives – "Accountability Engineering" and "Externality Economics"

https://www.flickr.com/photos/waltstoneburner/2863583929/

The Power of Mathematical Proof

http://www.dreamstime.com/royalty-free-stock-photography-wooden-puzzle-image7733587

Space of Intelligent Systems



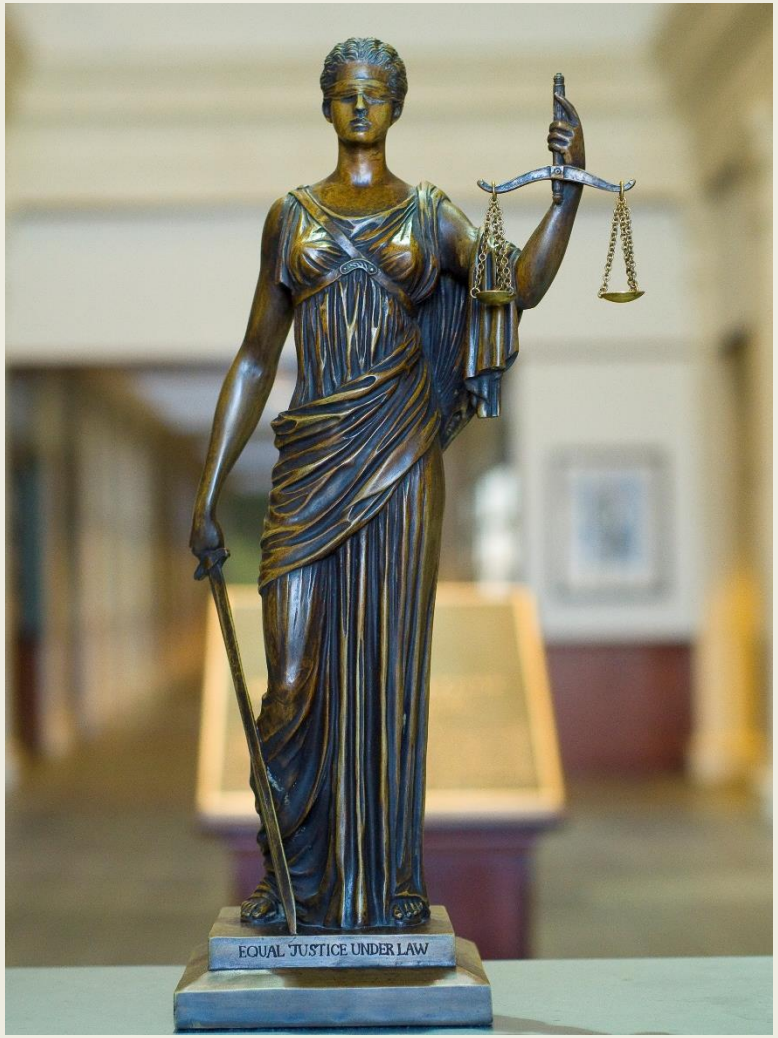
The Safe-AI Scaffolding Strategy

http://affordablehousinginstitute.org/blogs/us/2008/08/donors-as-scaffolding-part-2-the-value-of-coaching.html

http://www.flickr.com/photos/isaacmao/19245594/

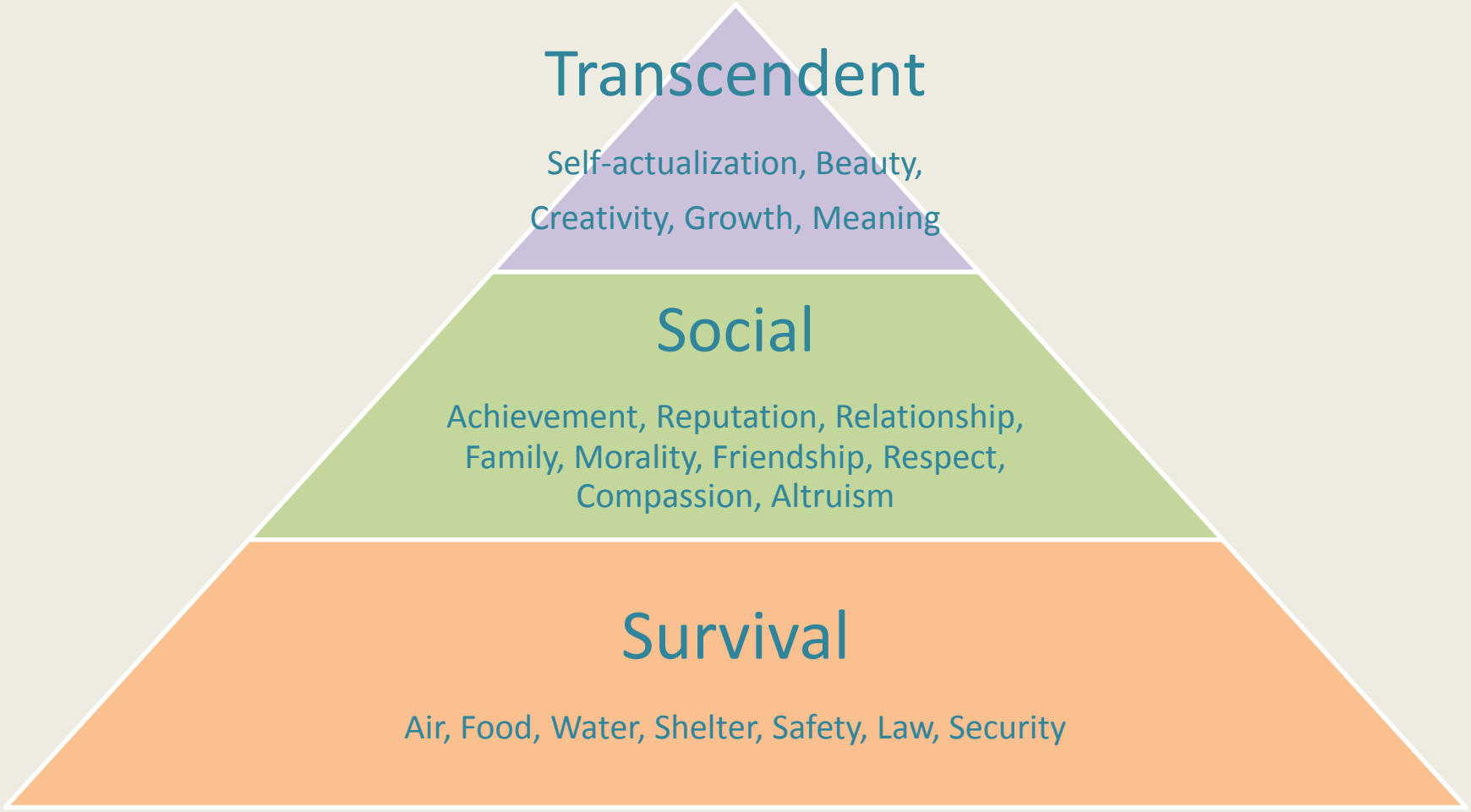
Accountable AI

- Allow untrusted systems
- But they must act through trusted proxies
- Require proofs of safety and legality



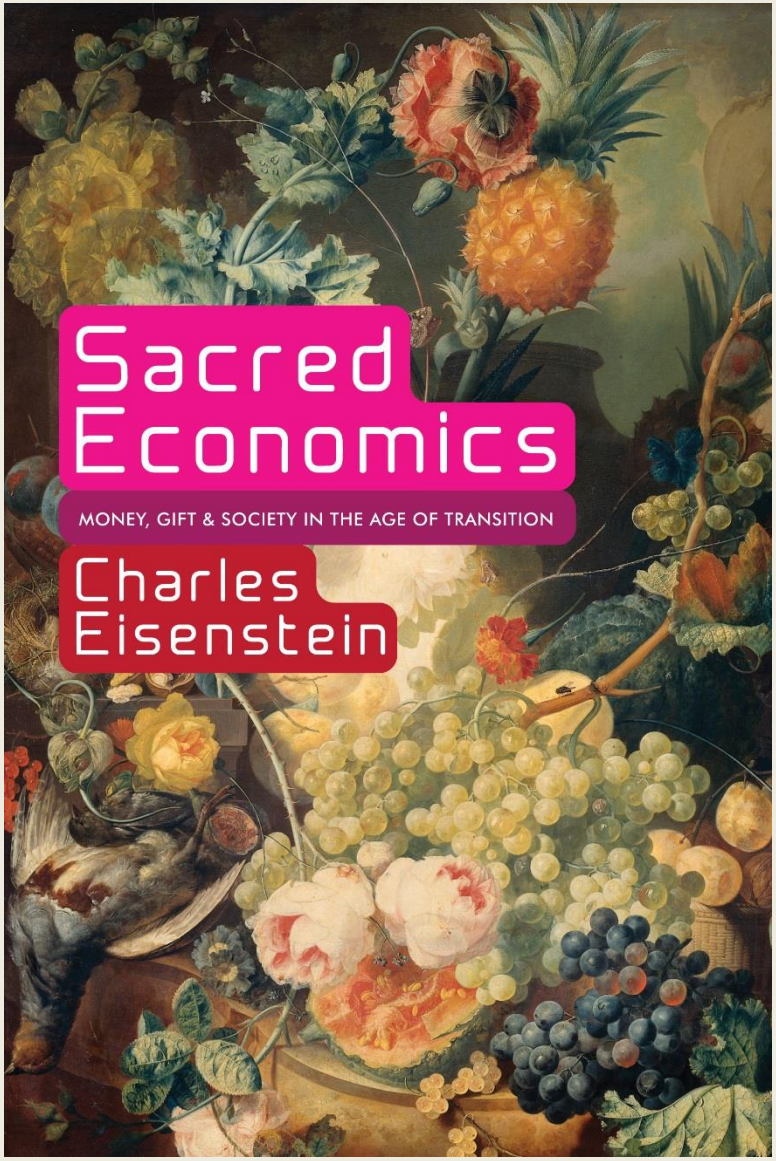
https://www.flickr.com/photos/jsmoorman/2299472114/

What do we want?



Compassionate Economics

- Expose externalities
- Align interests of agents with society
- Coase's theorem
- Promote win/win
- Rational pro-social self-design



Possibility Research's Approach

Omex: Programming

Omcor: Specification

Omai: Semantics

Omval: Values and Goals

Omgov: Governance

Our Challenge for This Century

To extend cooperative human values and institutions to autonomous technology for the greater good.

http://commons.wikimedia.org/wiki/File:Earth-moon.jpg