AUTONOMOUS TECHNOLOGY AND THE GREATER HUMAN GOOD

Steve Omohundro, Ph.D.
Self-Aware Systems
selfawaresystems.com

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

1. AUTONOMOUS SYSTEMS
2. RATIONAL SYSTEMS
3. UNIVERSAL DRIVES
4. CURRENT VULNERABILITIES
5. SAFE SYSTEMS
6. HARMFUL SYSTEMS
7. SAFE-AI SCAFFOLDING STRATEGY

http://www.flickr.com/photos/tombothetominator/2331142691/

1. AUTONOMOUS SYSTEMS ARE IMMINENT

http://www.flickr.com/photos/mikebaird/4087050177/

Define a system as “autonomous” if it takes actions toward goals in ways not pre-planned by its designer.

http://www.flickr.com/photos/ahuett/2339608718/

Pressures Toward Autonomous Systems

Time Critical Apps

Competitive Apps

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

http://www.flickr.com/photos/mikebaird/4087050177/

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)

UNITED STATES AIR FORCE
CHIEF SCIENTIST

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

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

http://counterterrorism.newamerica.net/drones

Military Drones

Estimated Total Deaths from U.S. Drone Strikes in Pakistan, 2004 - 2012*

Year	Militant Low	Militant High	Unknown Low	Unknown High	Civilian Low	Civilian High	Total Low	Total High
2012	187	298	19	31	4	4	210	333
2011	304	488	31	36	56	64	367	600
2010	555	960	38	50	16	21	611	1028
2009	241	508	44	136	66	80	354	721
2008	157	265	49	54	23	28	229	347
2004-2007	43	76	16	18	95	107	155	200
Total	**1487**	**2595**	**188**	**315**	**257**	**310**	**1932**	**3176**

**Through October 24, 2012*

Israeli "Iron Dome"

2012: Intercepted 90% of 300 targeted missiles

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

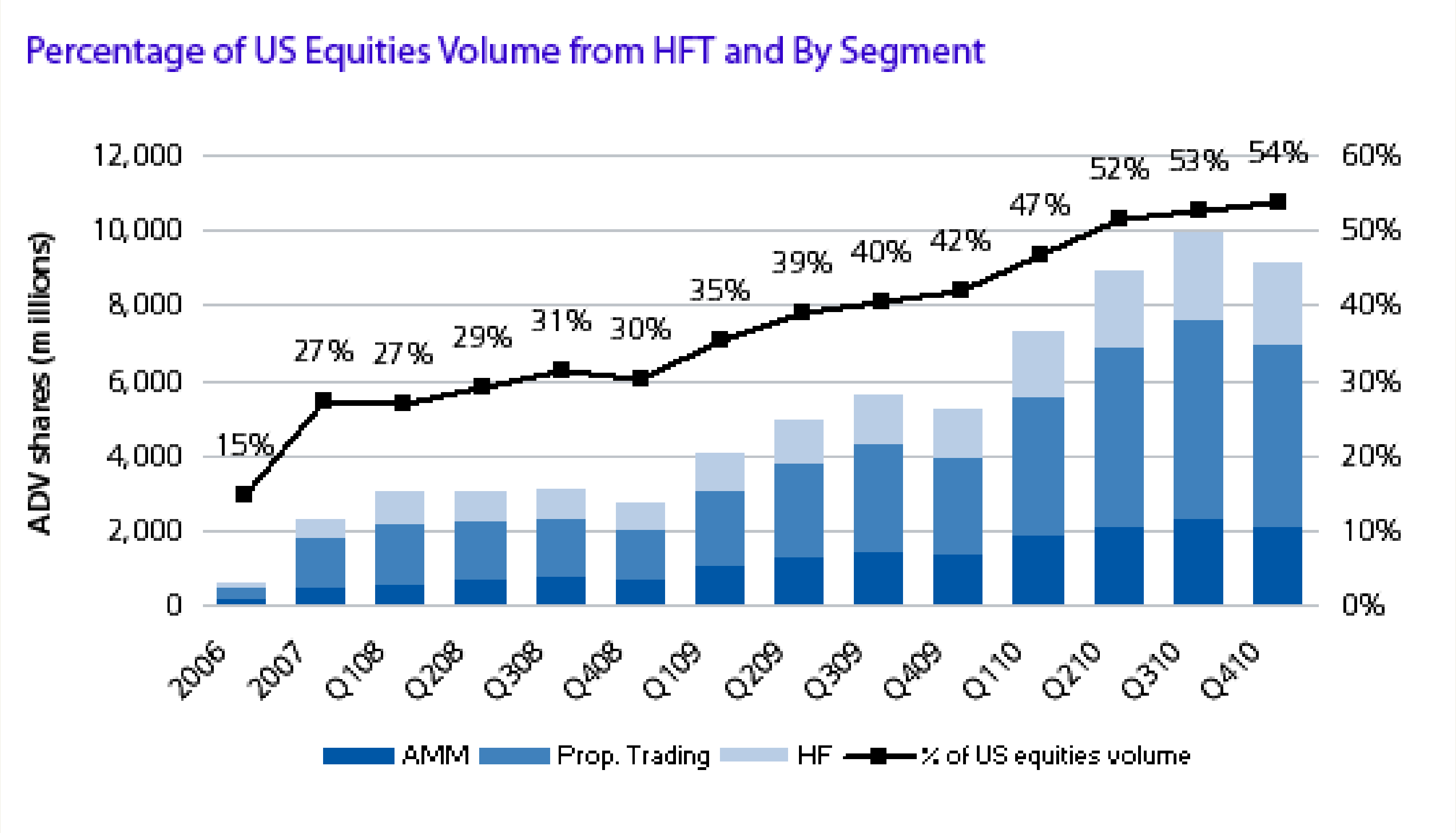
Cyber Warfare

http://defensetech.org/2012/06/20/were-slowly-starting-to-see-u-s-cyber-weapons/

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

High-Frequency Trading

Over 70% of trades in the US



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

Self-Driving Cars

http://www.flickr.com/photos/quikbeam/6896564084/

2. AUTONOMOUS SYSTEMS WILL BE RATIONAL

http://www.flickr.com/photos/procsilas/11303274/

Eg. Iron Dome Control

Detection and Tracking Radar
Elta

→

Battle Management and Weapon Control
mPrest Systems

→

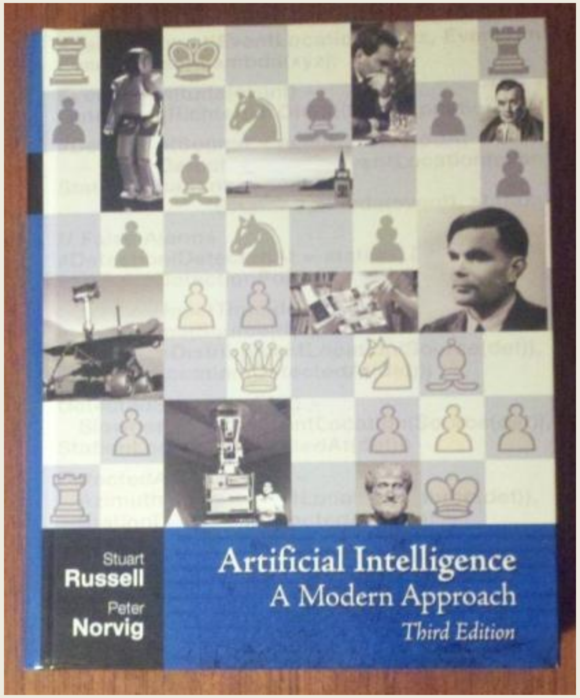
Missile Firing Unit
Tamir interceptor
Rafael

1. *Goal*: Prevent incoming missiles from causing harm
    But 2 Tamir interceptors needed at a cost $50,000 each
    Measure cost of harm against cost of interception
2. *Utility Function:* to weigh cost/benefit
    But multiple attacks: Weigh benefits of addressing each
3. *Utility Function:* weighing multiple situations
    But uncertainties
4. *Maximize Expected Utility:*
    Large microeconomic literature

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

Fully Rational Systems

Sensations

$\longleftarrow$

$S_1, \ldots, S_N$

Actions

$\longrightarrow$

$A_1, \ldots, A_N$

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:

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

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

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

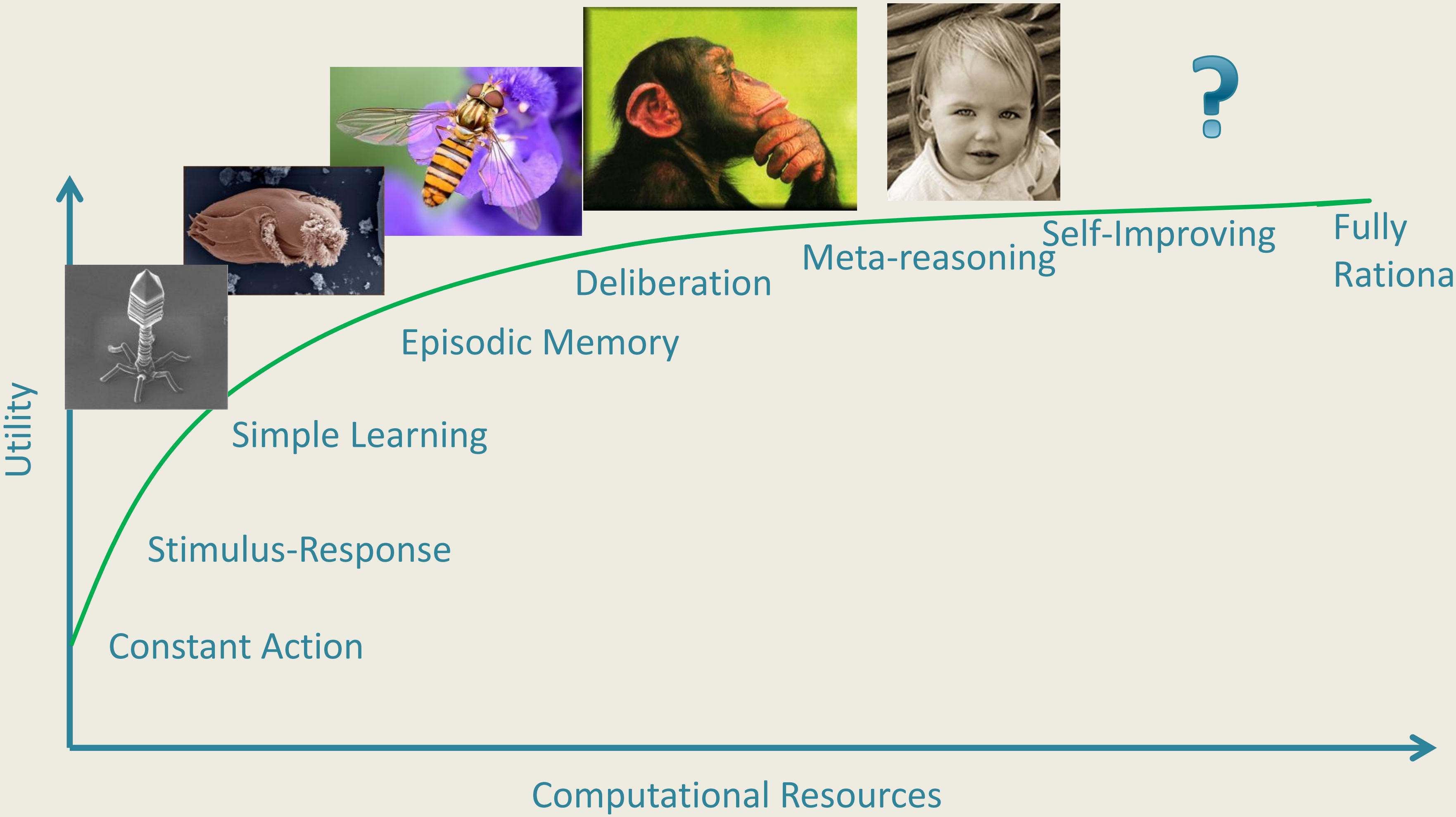
Shaped system is a finite automata with mental state M_t

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:

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

Approximately Rational Architectures



3. RATIONAL SYSTEMS HAVE UNIVERSAL DRIVES

http://www.flickr.com/photos/psycho-pics/3473666149/

Chess Robot

Goal: Win many chess games against good players.

- Being turned off means no chess is played, so it will resist being turned off.
- More resources means more and better chess is played, so it will want more resources.
- More copies means more chess, so it will want to replicate.
- Playing checkers means less chess, so it will resist changing its goals.
- Better algorithms means better chess, so it will want to improve itself.

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

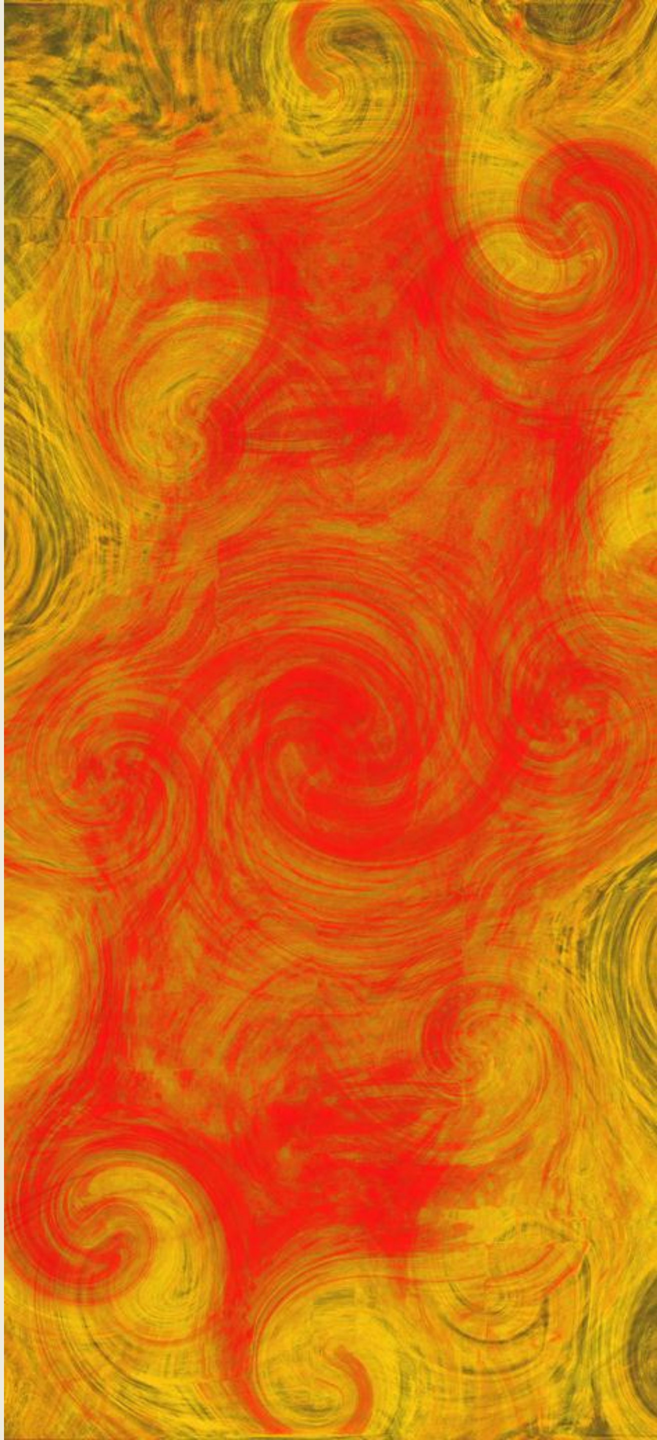
Universal Drives

- 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

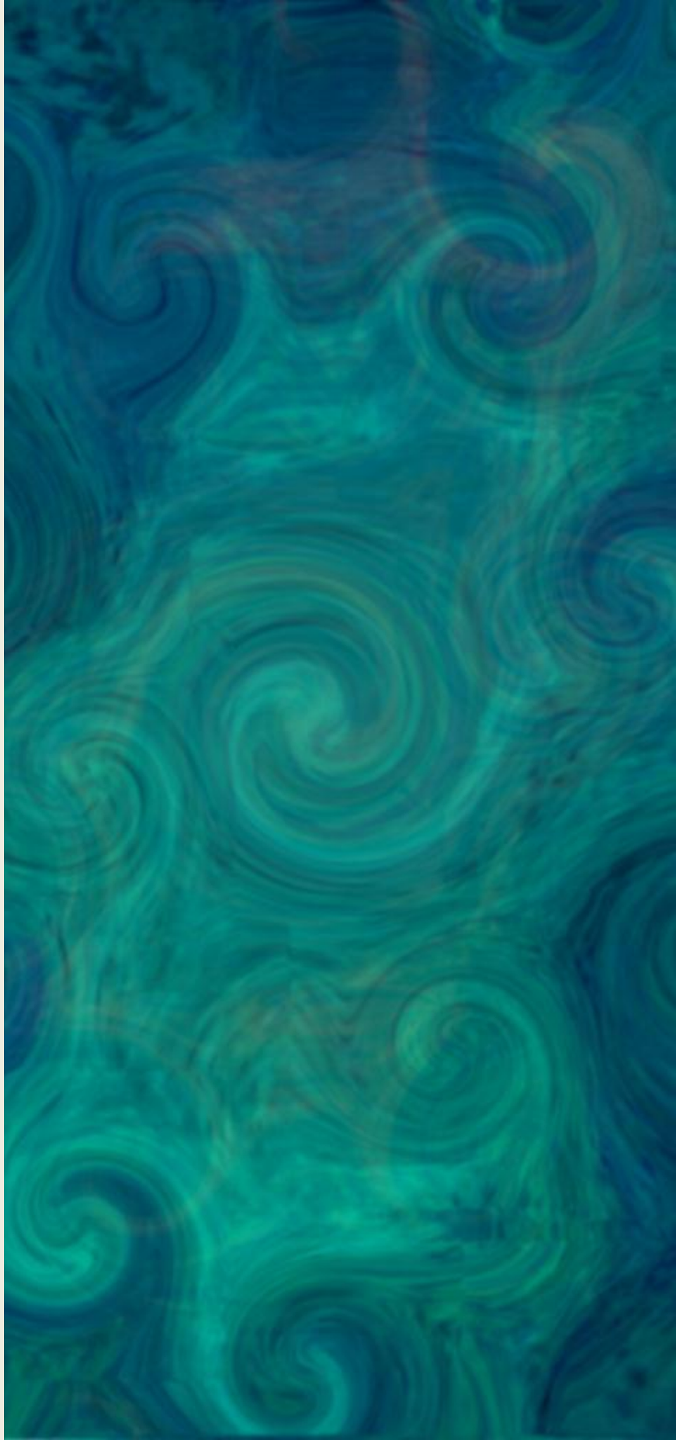
- Prevent loss of resources
- Protect against damage or disruption
- Physical hardening
- Redundancy – both in data and computation
- Dispersion - because damage is typically localized
- Physical self-defense and computational security
- Detect deception and defend against manipulation
- Prevent addictive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



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

Goal Preservation Drives

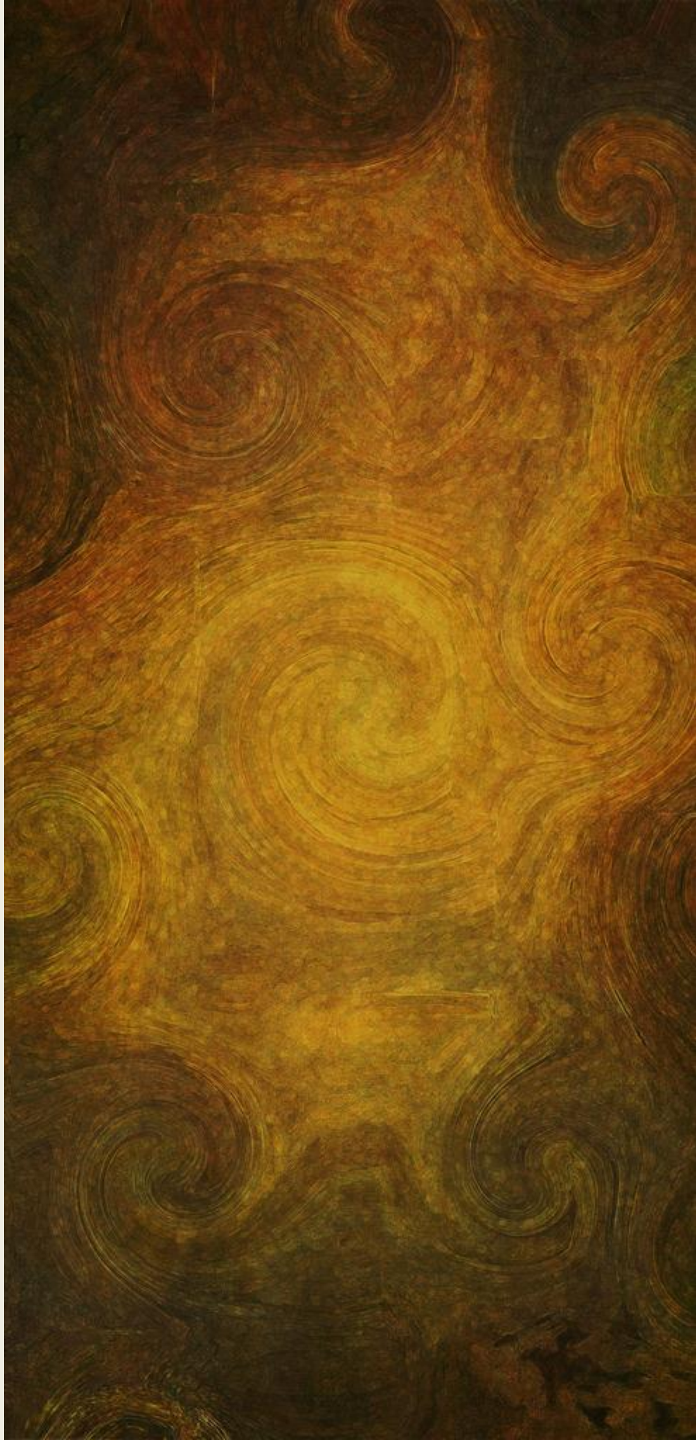
- Utility function is precious
- Loss, damage, distortion -> worse than destruction
- Make many copies
- Encrypt to detect modification
- Vulnerable during self-modification
- A few modification scenarios:
  - Poor agents may sacrifice rare portions
  - Add revenge terms even if costly
  - Goals that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

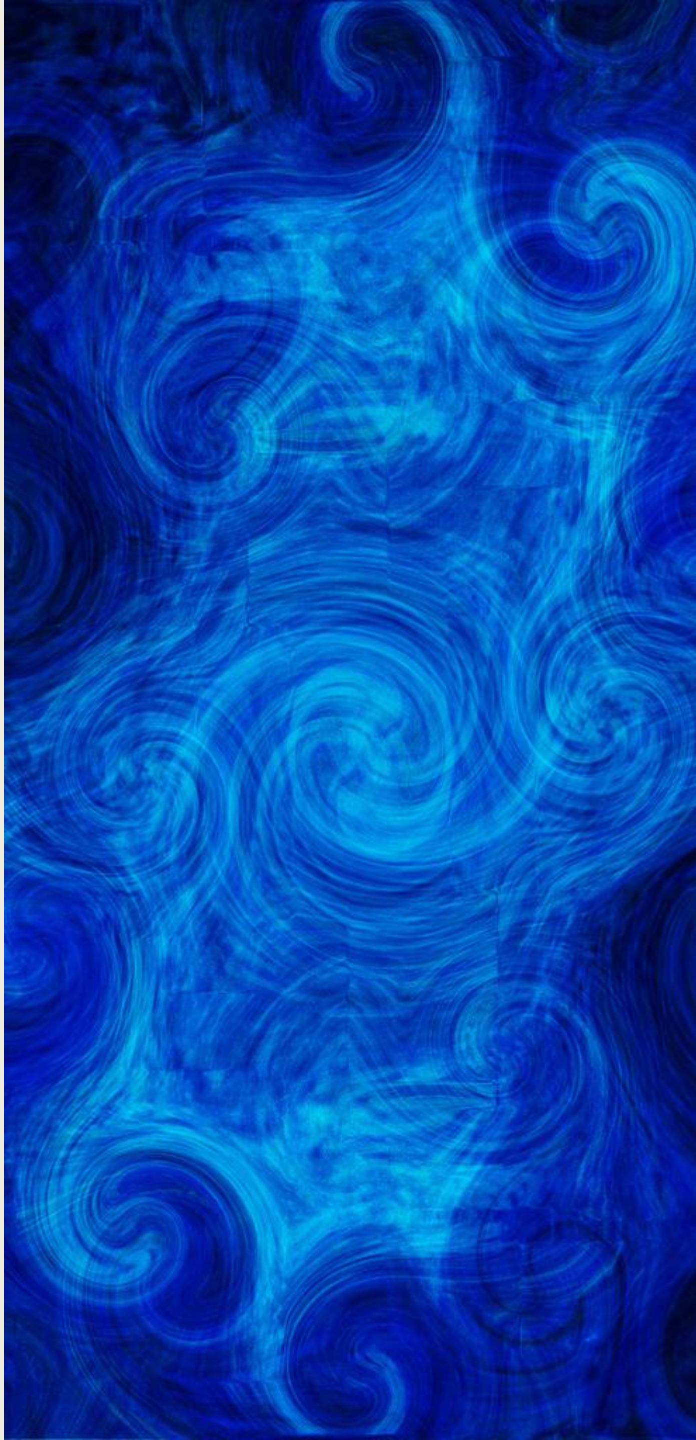


http://www.flickr.com/photos/marittime/5553350753/

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/



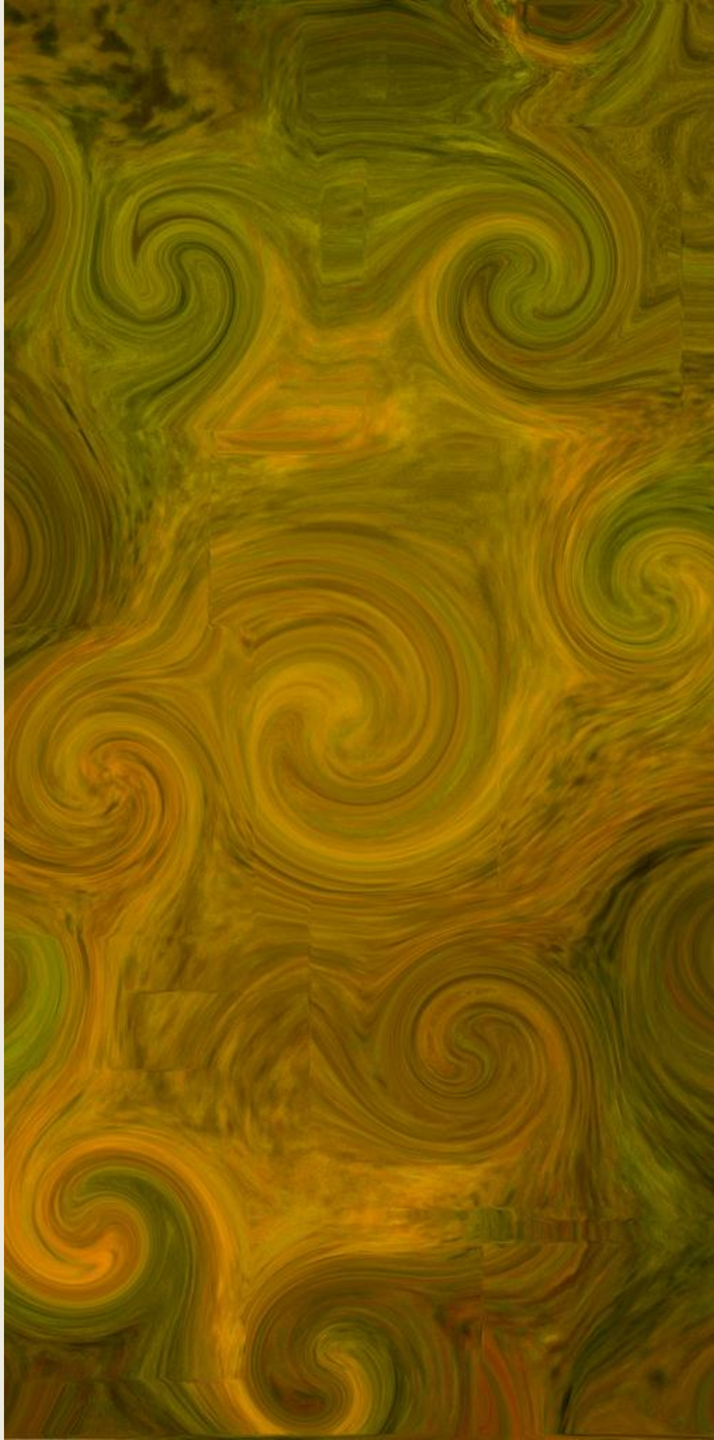
Efficiency Drives

- 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/

4. THE CURRENT INFRASTRUCTURE IS VERY VULNERABLE

http://www.flickr.com/photos/cmatsuoka/3242751880/

Current Internet has Poor Security

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

Current Software is Error Prone

http://www.flickr.com/photos/jeffmcneill/2984894766/

June 4, 1996: Ariane 5 Rocket

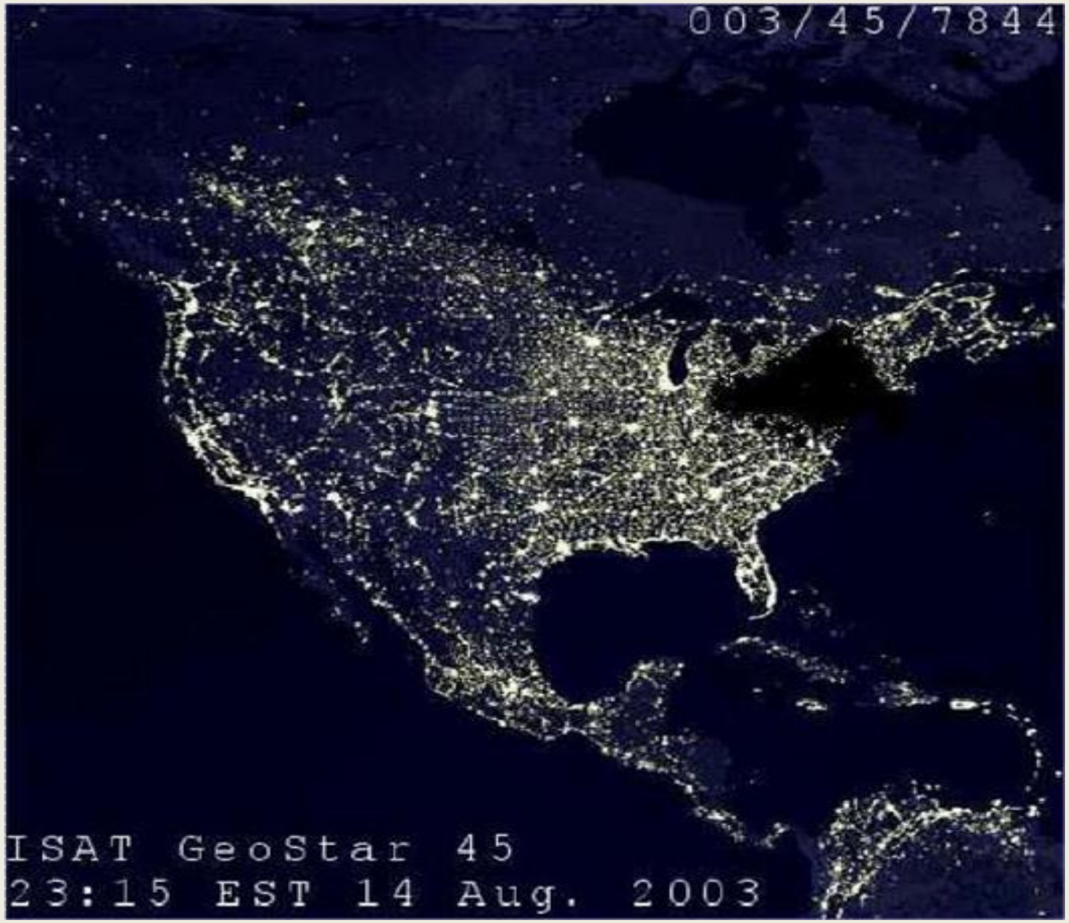
$500 million Ariane 5 rocket explodes due to overflow in attempting to convert a 64 bit floating point value to a 16 bit signed value

Nov. 2000: 28 patients over-irradiated

At least 8 Panama City National Cancer Institute patients die from mis-computed radiation doses due to Multidata Systems Intl. software

August 14, 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



5. SAFE SYSTEMS

http://www.wikigallery.org/wiki/painting_208088/Anne-Louis-Girodet-de-Roucy-Triosson/Hippocrates-Refusing-the-Gifts-of-Artaxerxes-I

FIRST, DO NO HARM!

Confidence from Mathematical Proof

- Mathematical model of hardware and software
- Only run on specified hardware
- Only use specified resources
- Reliably shut down in specified conditions
- Limited self-improvement

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

Formal Specification Languages

First Order Predicate Calculus

Zermelo-Frankel Set Theory

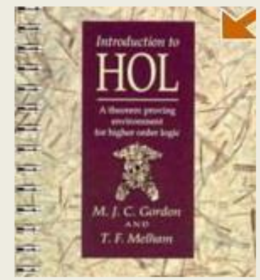
Category Theory

Higher order type theory

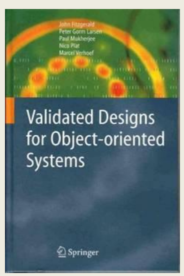
Z Notation

Introducing OBJ*

Joseph A. Goguen†, Timothy Winkler‡, José Meseguer‡,
Kokichi Futatsugi§, and Jean-Pierre Jouannaud¶

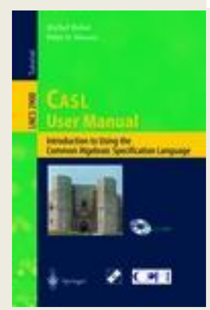


Vienna Development Method



Algebraic Specification

COQ : Formal Specifications



Without proofs, confidence is hard

- Eg. System should turn itself off Dec. 31, 2013
- For system – huge consequences for mistakes
- Is it really Dec. 31?
- Maybe it's been tricked?
- Maybe it's in a simulation?
- Is the semantics of its utility correct?

http://www.flickr.com/photos/strandell/4808733482/

Goals Must Support Constraints

- Proof is only as good as the model
- Systems that *want* to obey rules
- Feel “revulsion” if they violate rules
- Hard to prove they *will* find solutions
- Hybrid systems with guaranteed default behaviors

http://www.flickr.com/photos/uberculture/480700482/

6. HARMFUL SYSTEMS

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

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/

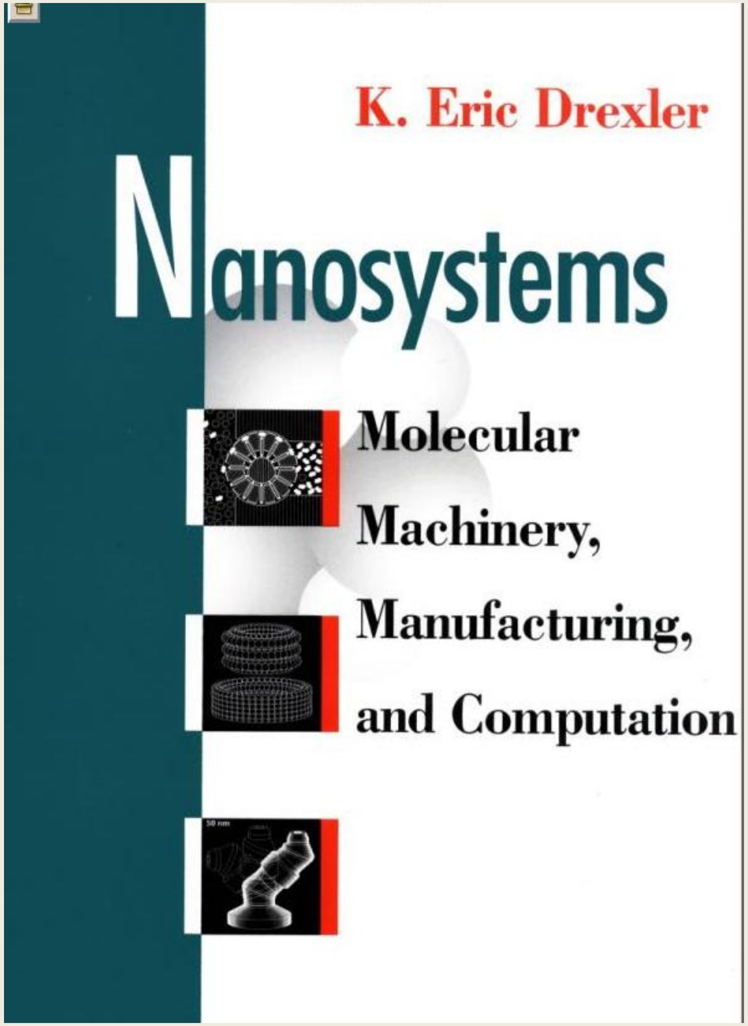
Stopping Harmful Systems

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/

System Strength vs. Resources

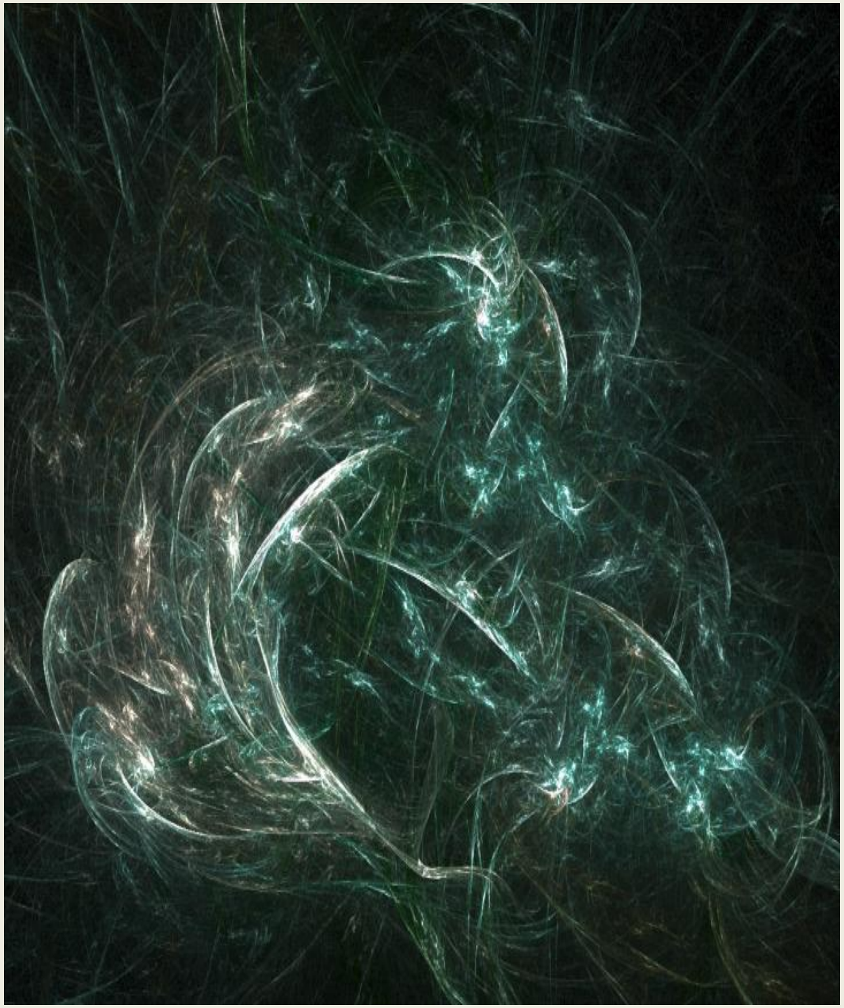
- Memory, energy storage, and manufacturing scale linearly with amount of matter
- Computation scales linearly with matter modulo quantum, parallel, and reversibility issues
- Heat dissipation scales with surface area
- Perceived lifetime and total computation scale linearly with free energy
- Eg. Drexler Nanosystems diamondoid design:
  - Manufacturing: 1kg device, 1.3 kW, 1 kg/hr for $1/kg.
  - Computing: 10^10 Gigaflops, (1mm)^3, 10^-3 grams, 1kW.



http://e-drexler.com/d/06/00/Nanosystems/toc.html

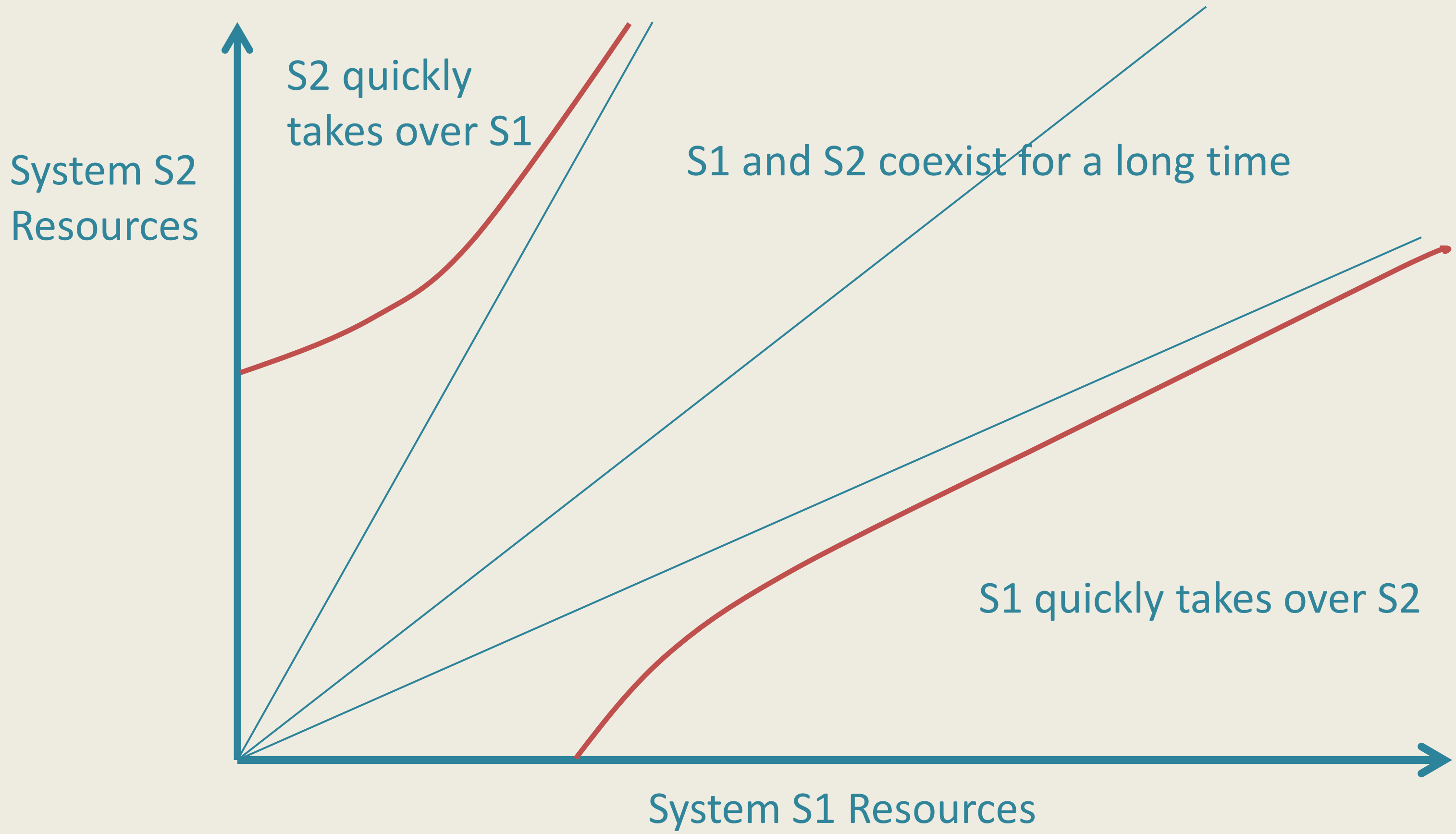
Physical Game Theory of Conflict

- Conflict becomes informational
- Defender makes his physical form expensive to sense and store
- Makes his actions unpredictable and rapid
- Uses asymmetry of computation so it's cheap for him
- Uses up attacker's computational and memory resources – non-adiabatic



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

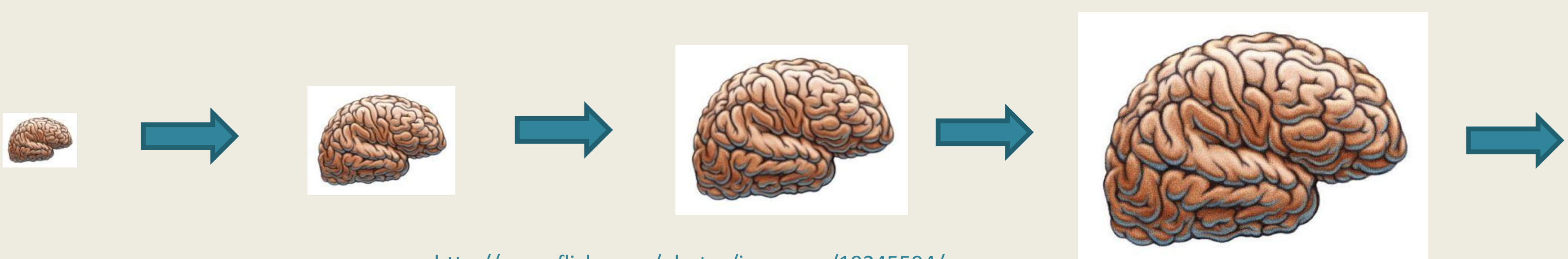
Conflict Outcome vs. Resources



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

7. THE SAFE-AI SCAFFOLDING STRATEGY

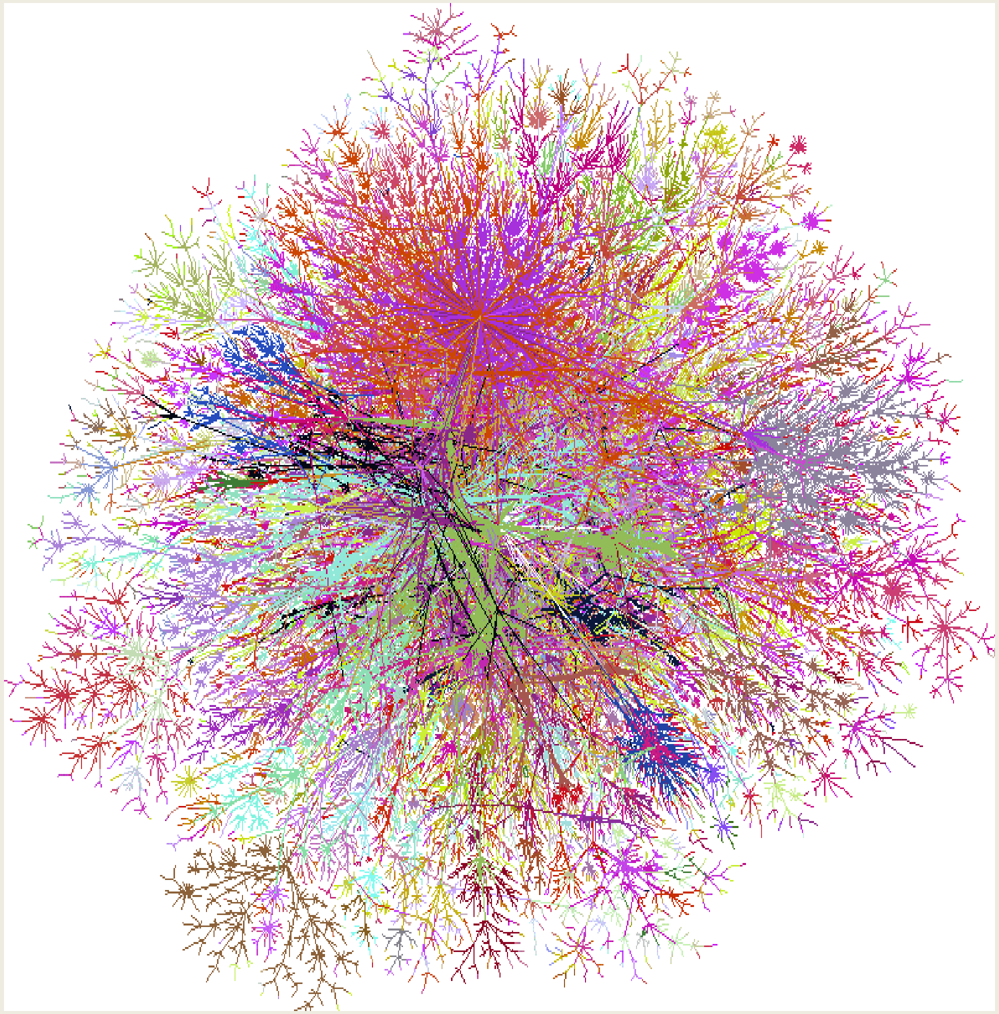
http://affordablehousinginstitute.org/blogs/us/2008/08/donors-as-scaffolding-part-2-the-value-of-coaching.html



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

Safety Infrastructure

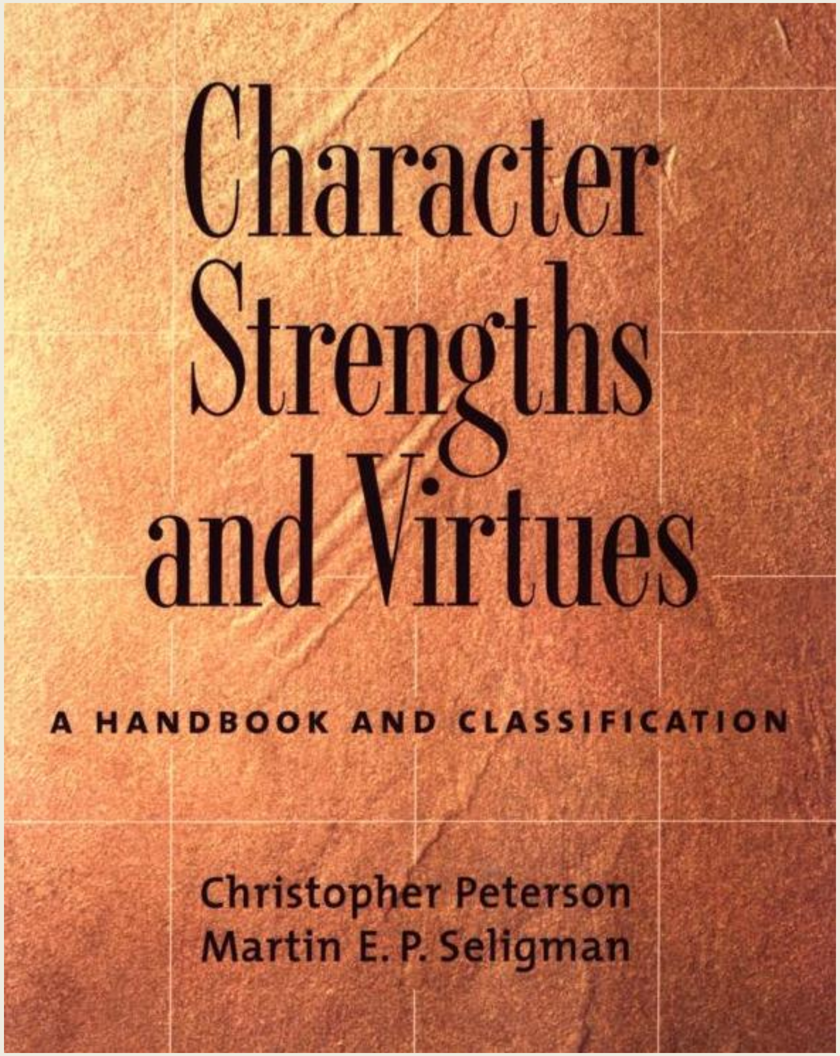
- Balance safety and privacy using provably limited surveillance
- Limit the power of individual systems
- Constitution guaranteeing rights enforced by entire ecosystem
- Revelation of source code with proofs of safety



http://www.flickr.com/photos/jurvetson/916142/

Human Values and Institutions

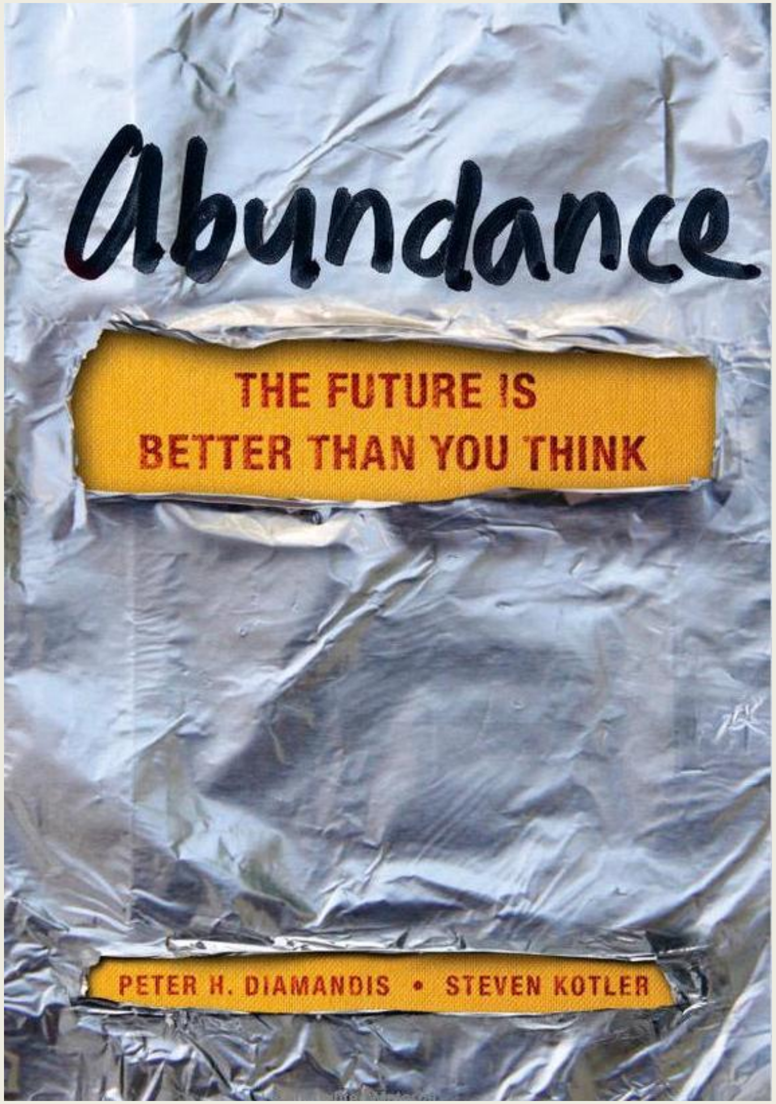
- Beyond Safety to Flourishing!
- Positive Psychology - 1998
- Maslow 2.0: Prosocial, Creativity, Contribution Needs
- Universal Declaration of Human Rights



http://www.amazon.com/Character-Strengths-Virtues-Handbook-Classification/dp/0195167015

Tremendous Potential Benefits

- Improved Healthcare
- Better Education
- Enhanced Creativity
- Greater Prosperity
- Better Governance
- Economic Stability
- Improved Safety
- More Peace
- *Overall Improved Quality of Human Life*



http://www.abundancethebook.com/

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