AUTONOMOUS TECHNOLOGY AND THE GREATER HUMAN GOOD

Steve Omohundro, Ph.D.
Self-Aware Systems
selfawareshystems.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/
“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...”
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.”

Military Drones

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Militant Low</th>
<th>Militant High</th>
<th>Unknown Low</th>
<th>Unknown High</th>
<th>Civilian Low</th>
<th>Civilian High</th>
<th>Total Low</th>
<th>Total High</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>187</td>
<td>298</td>
<td>19</td>
<td>31</td>
<td>4</td>
<td>4</td>
<td>210</td>
<td>333</td>
</tr>
<tr>
<td>2011</td>
<td>304</td>
<td>488</td>
<td>31</td>
<td>36</td>
<td>56</td>
<td>64</td>
<td>367</td>
<td>600</td>
</tr>
<tr>
<td>2010</td>
<td>665</td>
<td>960</td>
<td>38</td>
<td>60</td>
<td>16</td>
<td>21</td>
<td>611</td>
<td>1028</td>
</tr>
<tr>
<td>2009</td>
<td>241</td>
<td>508</td>
<td>44</td>
<td>135</td>
<td>66</td>
<td>80</td>
<td>354</td>
<td>721</td>
</tr>
<tr>
<td>2008</td>
<td>157</td>
<td>265</td>
<td>49</td>
<td>64</td>
<td>23</td>
<td>28</td>
<td>229</td>
<td>347</td>
</tr>
<tr>
<td>2004-2007</td>
<td>43</td>
<td>76</td>
<td>16</td>
<td>18</td>
<td>95</td>
<td>107</td>
<td>155</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>1487</td>
<td>2595</td>
<td>188</td>
<td>315</td>
<td>257</td>
<td>310</td>
<td>1932</td>
<td>3176</td>
</tr>
</tbody>
</table>

*Through October 24, 2012


http://counterterrorism.newamerica.net/drones
Israeli
“Iron Dome”

2012: Intercepted 90% of 300 targeted missiles

Cyber Warfare

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

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

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

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

Modern Approach to AI
Fully Rational Systems

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 \( t \):
\[
A_t^R (S_1, A_1, \ldots, A_{t-1}, S_t) = \arg\max_{A_t^R} \sum_{S_{t+1}, \ldots, S_N} U(S_1, \ldots, S_N) P(S_1, \ldots, S_N \mid A_1, \ldots, A_{t-1}, A_t^R, \ldots, A_N^R)
\]

**The Formula for Intelligence!**

*It includes Bayesian Inference, Search, and Deliberation.*

But it requires \( O(N S^N A^N) \) computational steps.
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:

$$\argmax_{A^M \in C} \sum_{S_1, \ldots, S_N} U(S_1, \ldots, S_N) P(S_1, \ldots, S_N | A_1^M, \ldots, A_N^M)$$
Approximately Rational Architectures

- Computational Resources
- Utility
- Constant Action
- Simple Learning
- Episodic Memory
- Deliberation
- Meta-reasoning
- Self-Improving
- Fully Rational

- Stimulus-Response

Diagram: Improvement of architectures with increasing computational resources and utility.
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/maritime/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
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
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
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/maritime/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/maritime/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
- ...

Flame virus 'much bigger than Stuxnet'

San Francisco Chronicle

CGAWKER

Apple’s Worst Security Breach: 114,000 iPad Owners Exposed
Current Software is Error Prone

http://www.flickr.com/photos/jeffmcneill/298494766/
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

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
Formal Specification Languages

First Order Predicate Calculus

Zermelo-Frankel Set Theory

Category Theory

Higher order type theory

Z Notation

Vienna Development Method

Algebraic Specification

Introducing OBJ*
Joseph A. Goguen¹, Timothy Winkler², José Meseguer², Kokichi Futatsugi³, and Jean-Pierre Jouannaud⁴

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?
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
6. HARMFUL SYSTEMS

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, $(1\text{mm})^3$, $10^{-3}$ 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/
Conflicts Outcome vs. Resources

- S2 quickly takes over S1
- S1 and S2 coexist for a long time
- S1 quickly takes over S2

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://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
Human Values and Institutions

• Beyond Safety to Flourishing!

• Positive Psychology - 1998

• Maslow 2.0: Prosocial, Creativity, Contribution Needs

• Universal Declaration of Human Rights

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