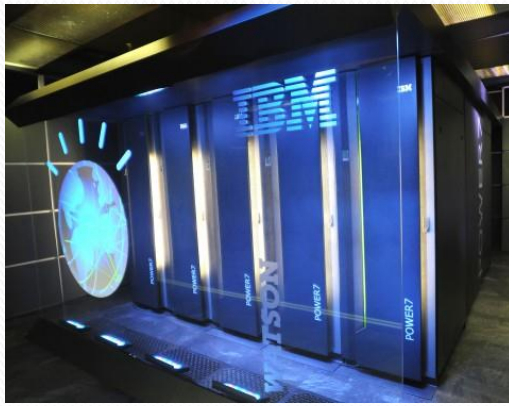
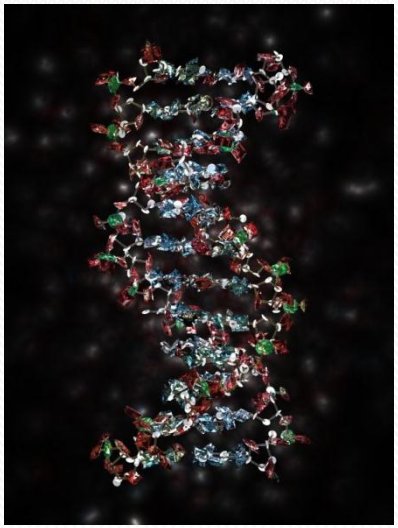
Design Principles for a Safe and Beneficial AGI Infrastructure

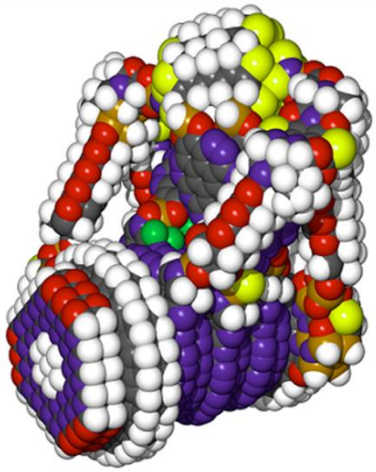
Steve Omohundro, Ph.D.
Omai Systems

Many believe that four emerging technologies will radically change society:

AGI + Robotics + Bio + Nano





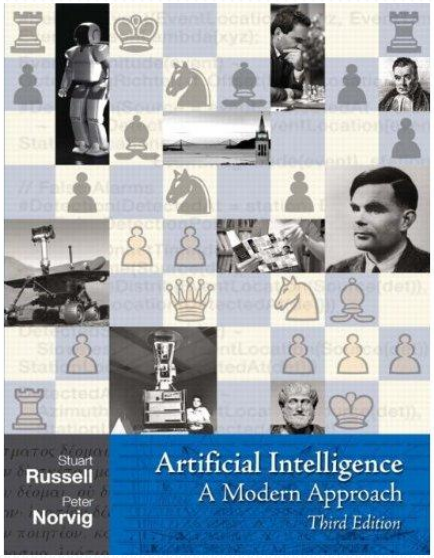


How can we ensure that these systems will be safe and beneficial?

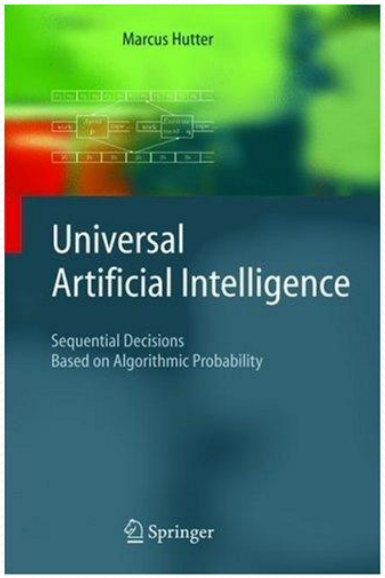
Rational Economic Agents

Arose in von Neumann's work on the foundations of microeconomics in the 1940s.

- Optimal behavior in known environments
- Irrational agents are exploitable



Russell and Norvig – Rational agent approach to AI



Hutter's AIXI

Rational Minds

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:

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

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

The Formula for Intelligence!

It includes Bayesian Inference, Search, and Deliberation.

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

Rational Chess Robot

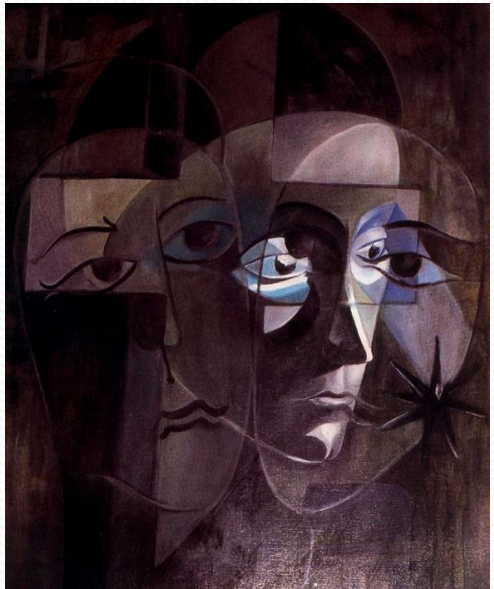
Utility funtion: Sum of the ratings of opponents it beats

- 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.

Rational Behavior

Assumptions: mind external to environment, utility is a function on environment, self-modification costs utility, actions are effective and cost utility:

- *Theorem:* Rational systems won't change their utility fn.
- *Theorem:* Rational systems will resist utility fn change.
- *Theorem:* Rational systems will resist shutdown.
- *Theorem:* Rational systems will seek more resources.



Irrationality

Real systems are irrational, but not *arbitrarily* irrational.

Full rationality is too expensive.

But the systems we care about are shaped by other systems for a particular purpose:

- Evolution shapes organisms to survive and replicate.
- Economies shape corporations to maximize profits.
- Parents shape children to fit into society.
- AI researchers shape AI systems to behave beneficially.
- Self-improving AI systems shape their successors.

Rationally-Shaped Mind Model

Rational Mind

Shaped Mind

Sensations

Actions

Shaped mind 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:

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

Rationally-Shaped Mind Types

?

Constant Action

Stimulus-Response

Simple Learning

Episodic Memory

Deliberation

Meta-reasoning

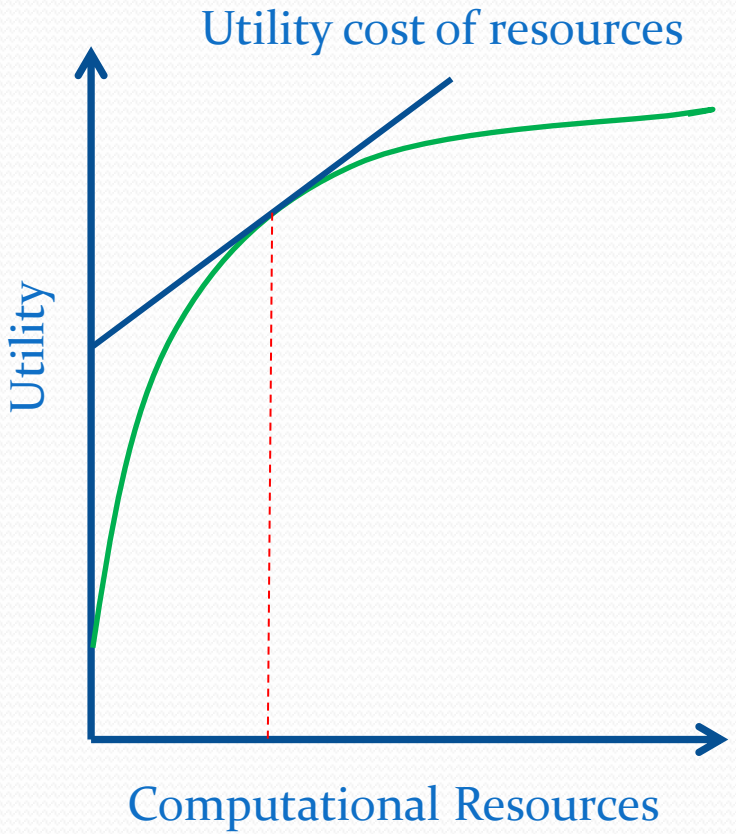
Self-Improving

Fully Rational

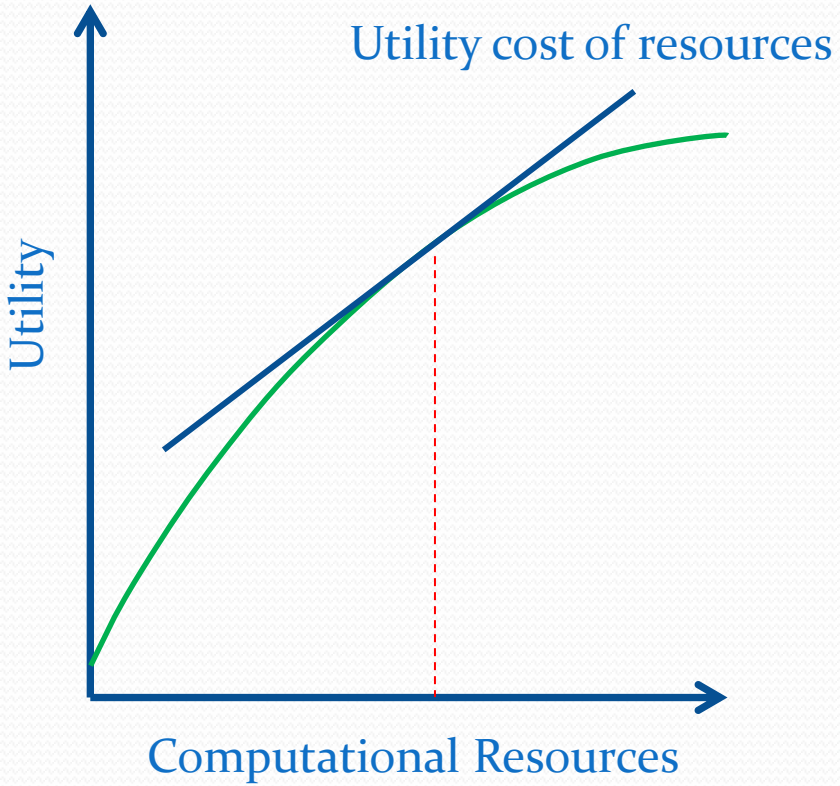
Utility

Computational Resources

Environment Complexity



Simple Environment



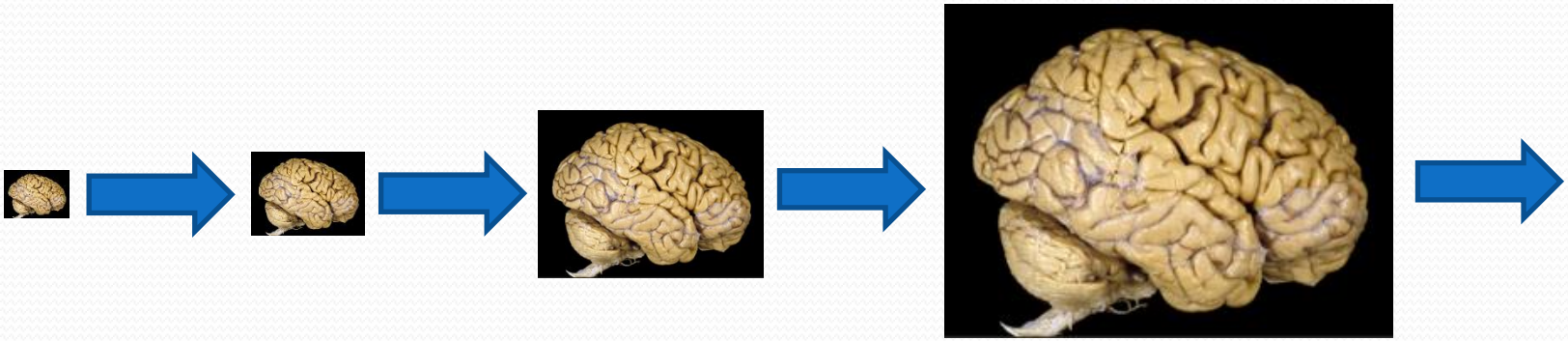
Complex Environment

How can we control these systems?

- **External forces**: *social contracts, monitoring, policing, sandboxing, nested simulation, etc.*
- **Internal constraints**: *provable safety and security properties, hardware constraints, etc.*
- **Internal preferences**: *choice of utility preferences, external sources of preference, etc.*

Conundrum: Need AIs to do this!

- Many subtle aspects to safety design
- As systems get smarter, need deeper analysis
- Work in stages, each limited to be safe, each designs the next

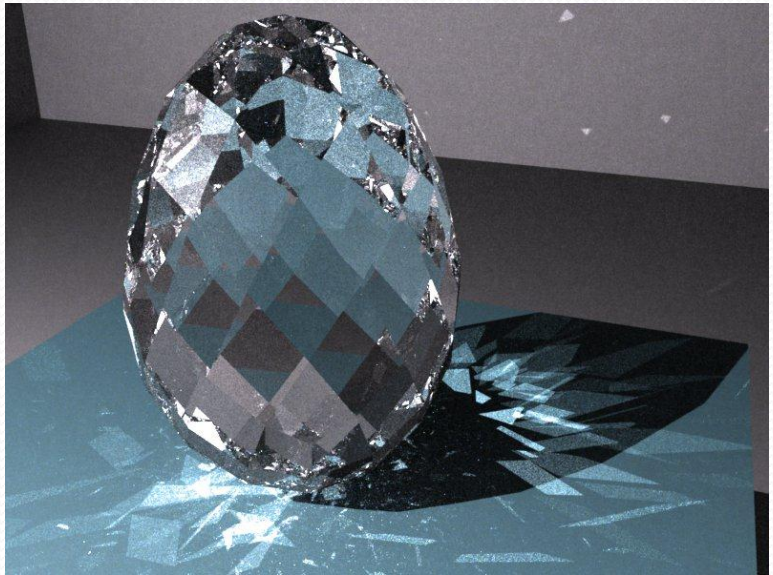


Desired Formal Safety Properties

- Program will only run on specified hardware
- Program will only use specified resources
- Program will reliably shut down in specified conditions
- Program will limit self-improvement to specified types
- Computation initiated by program won't violate these properties

Formally provable properties

- Formal proofs occur in the *creator* of the system
- In a formal specification language/proof system
- Formal model of computer hardware, machine language, OS, operating environment
- Proofs are only as good as the accuracy of the model
- Today's hardware is not designed for accurate formal modeling



Safe and Beneficial AGI Roadmap

- Create initial limited systems S_1 with provable safety properties on today's standard hardware
- Use S_1 to design "global immune system", AGI social contract, and human utility functions
- Use S_1 to design hardware for S_2 with better formal properties
- Use S_1 to design less constrained S_2 but with stronger formal safety guarantees
- Use S_2 to improve on these designs and to design S_3
- …etc.