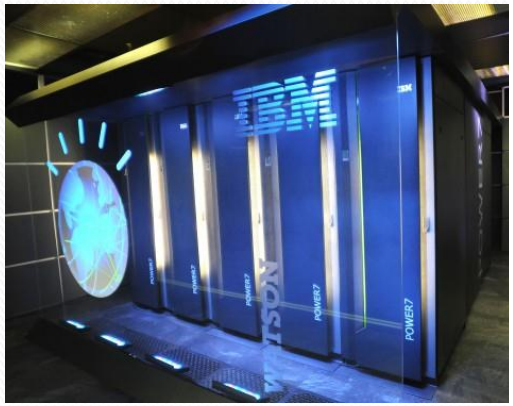
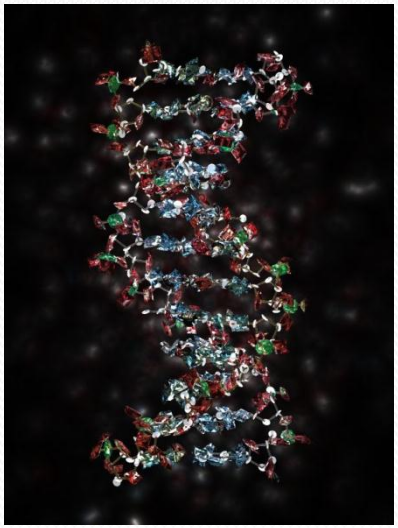
Rationally-Shaped Minds: A Framework for Analyzing Self-Improving AI

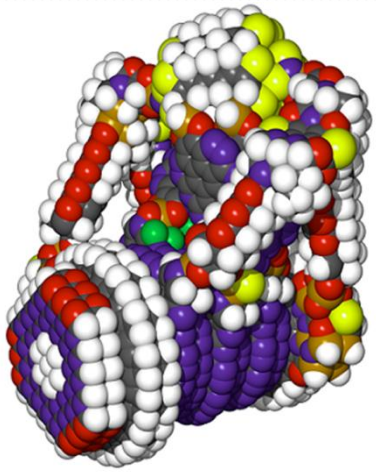
Steve Omohundro, Ph.D.
Omai Systems

Four emerging technologies are likely to radically change society:

AGI + Robotics + Bio + Nano







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

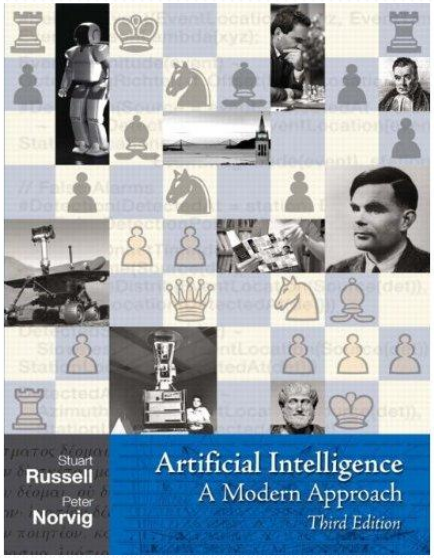
John von Neumann

- Architecture of modern computers
- Early artificial intelligence
- Self-reproducing automata
- Foundations of logic
- Game theory and microeconomics
- Nuclear weapons and deterrence
- To Ulam in 1950's: We are "approaching some essential singularity in the history of the race."

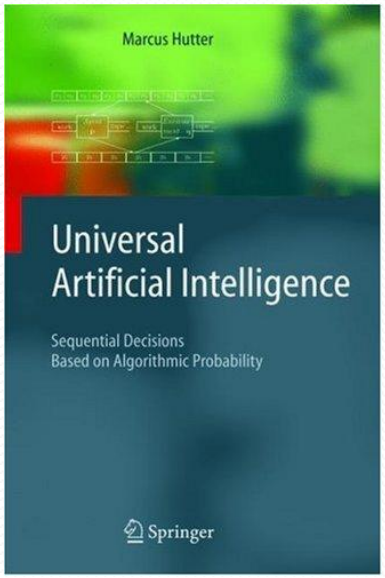
Rational Economic Agents

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

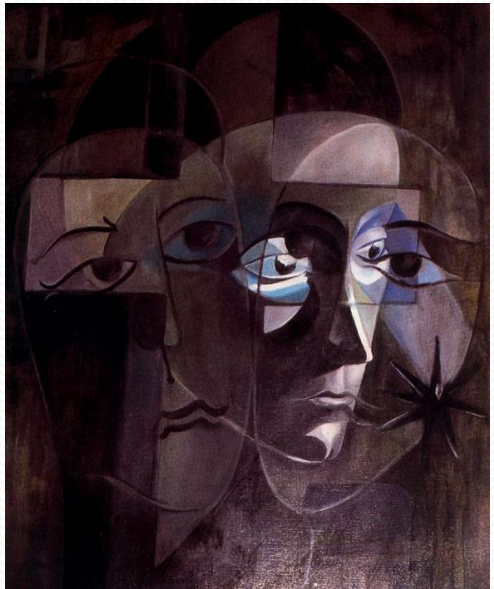
Utility funtion: Sum of the ratings of opponents it beats

- 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

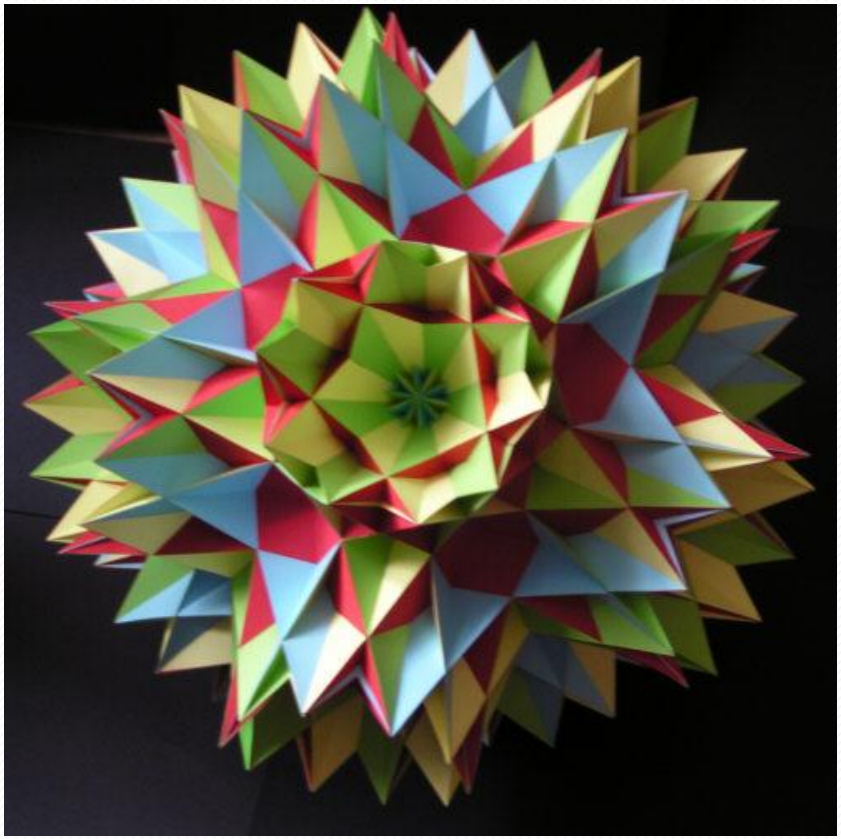
Assumptions: mind external to environment, utility is a function on environment, self-modification costs utility, actions are effective and cost utility:

- *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.



Drives in any Rational System

- Nothing special about AI here
- Applies to humans, corporations, animals, political groups



Irrationality

Full rationality is too expensive.

So real systems often must be irrational,

but not *arbitrarily* irrational.

- Agents whose goal is to change their goal
- Agents whose goal is to kill themselves
- Agents who want to destroy their resources
- Agents who want to be wrong about the world
- Agents who are sure of the wrong physics

Minds Making Minds

The systems we care about are shaped by other systems for particular purposes:

- Evolution shapes organisms to survive and replicate.
- Economies shape corporations to maximize profit.
- Parents shape children to fit into society.
- AI researchers shape AI systems to behave beneficially.
- Self-improving AI systems shape themselves for goals.

Rationally-Shaped Mind Model

Rational Mind

Shaped Mind

Sensations

Actions

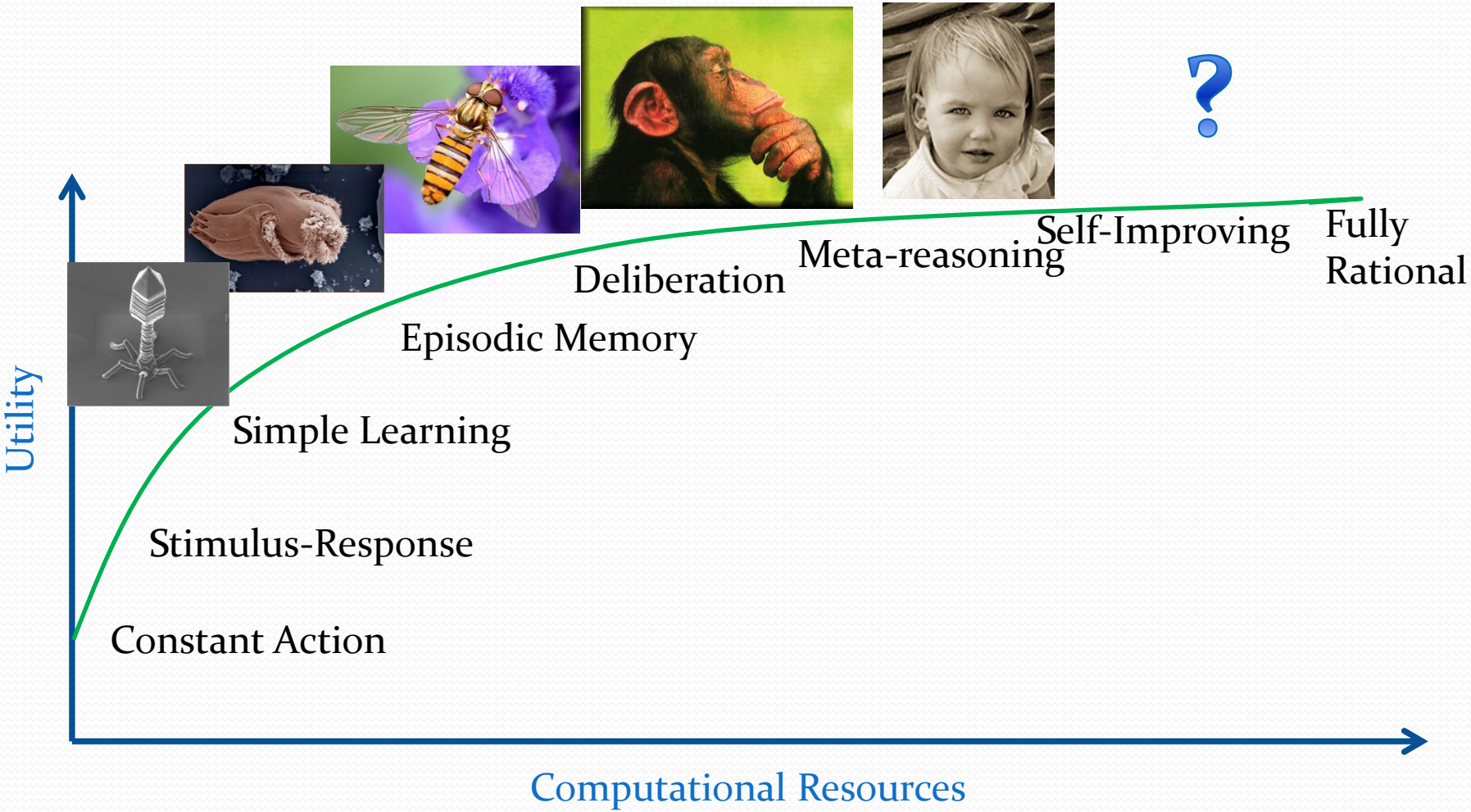
Shaped mind is a finite automata with mental state M_t

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

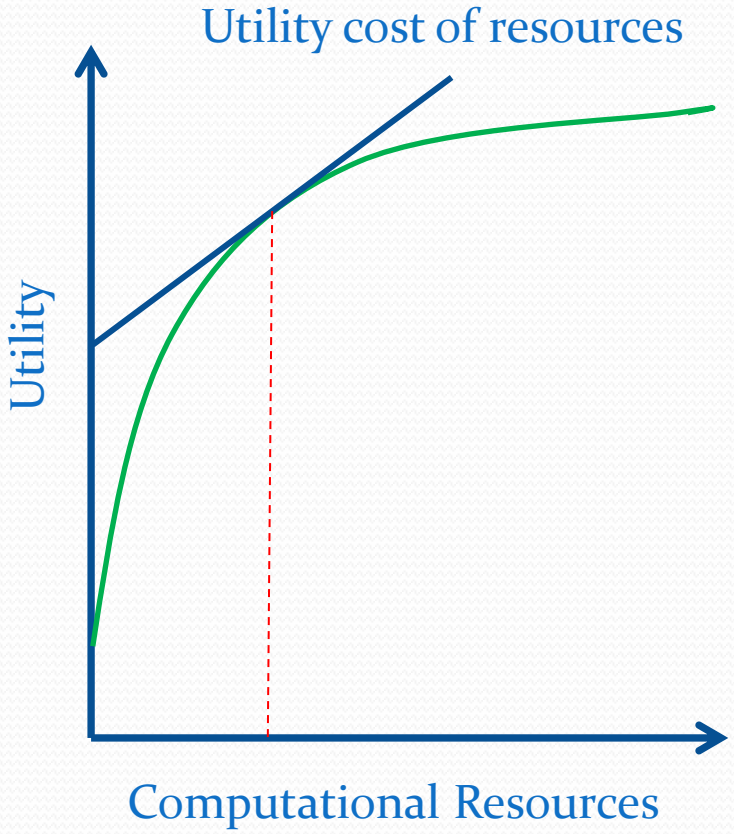
Rational shaper chooses from class C of systems with space/time and other constraints to maximize expected utility:

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

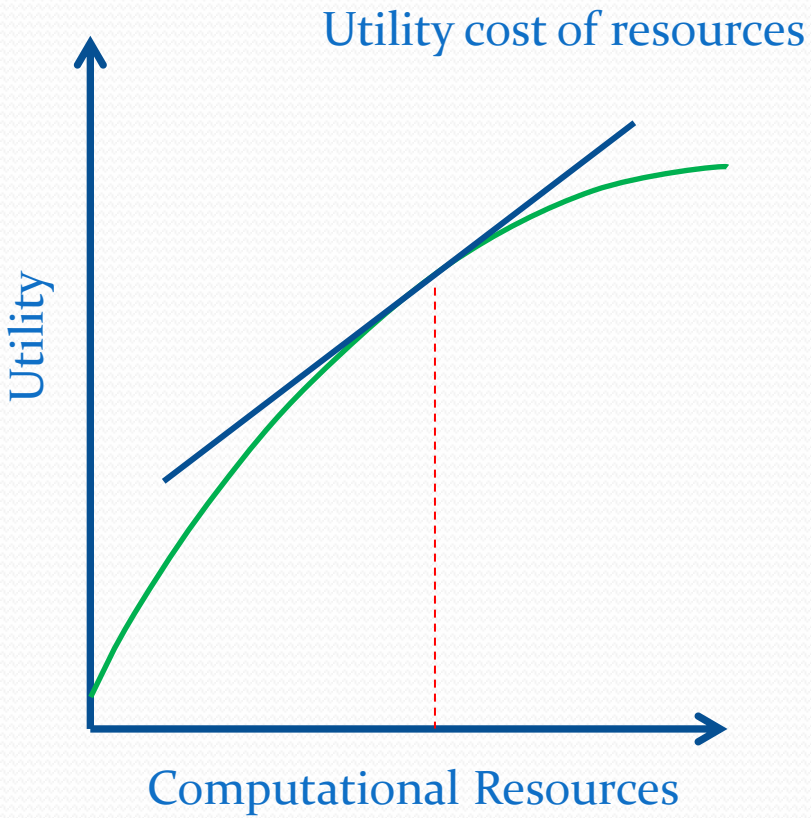
Rationally-Shaped Mind Types



Environment Complexity



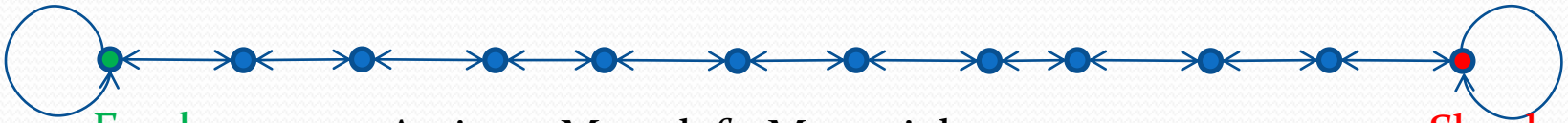
Simple Environment



Complex Environment

We can rank tasks by the time/space complexity
of the simplest rational agent optimal for that task.

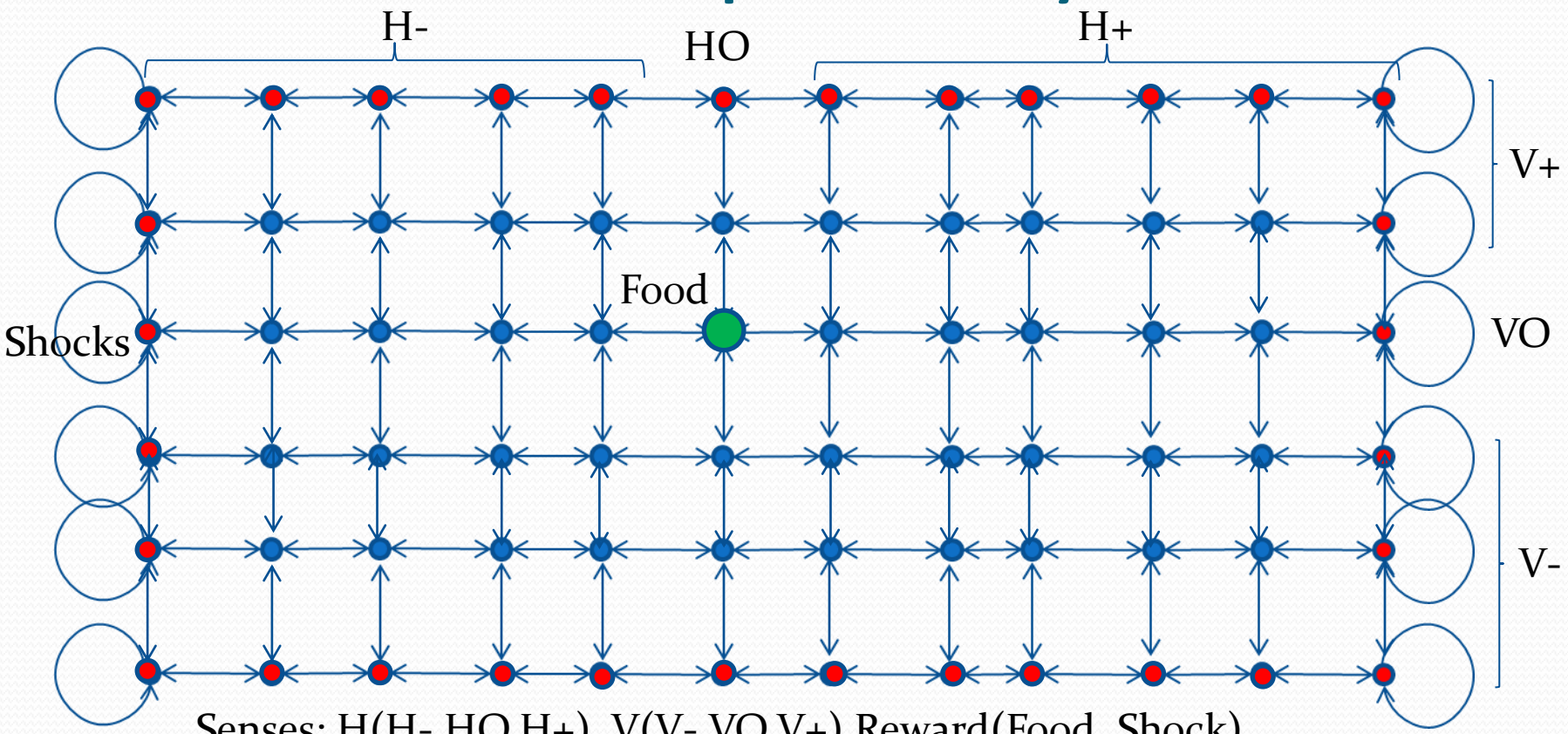
Constant Action Systems



Actions: Move left, Move right
Senses: Food, Neutral, Shock
Utility: Total food minus total shocks

- Optimal strategy is to always move left
- Constant action agent
- Simplest kind of environment

Stimulus/Response Systems



Senses: H(H-,HO,H+), V(V-,VO,V+),Reward(Food, Shock)

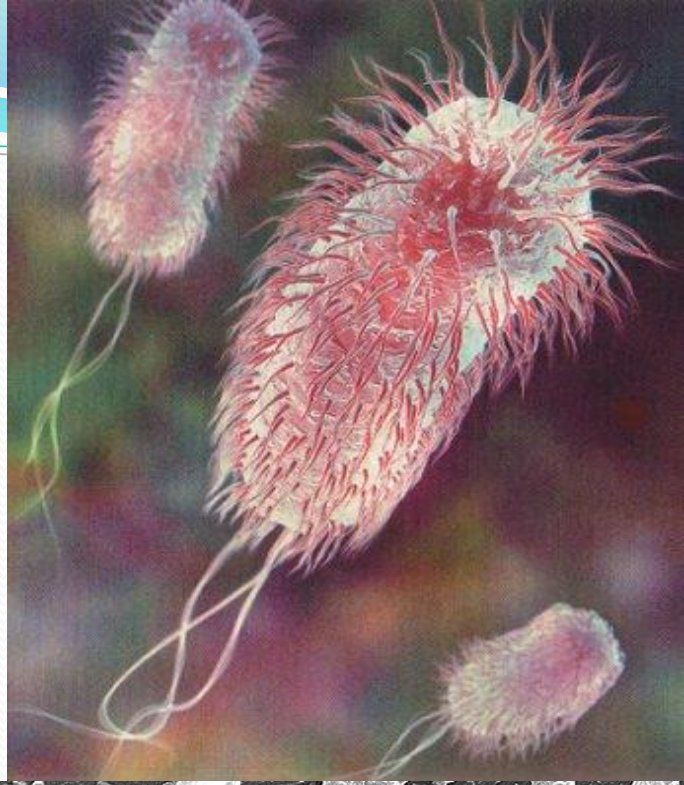
Actions: Left, Right, Up, Down, Stationary

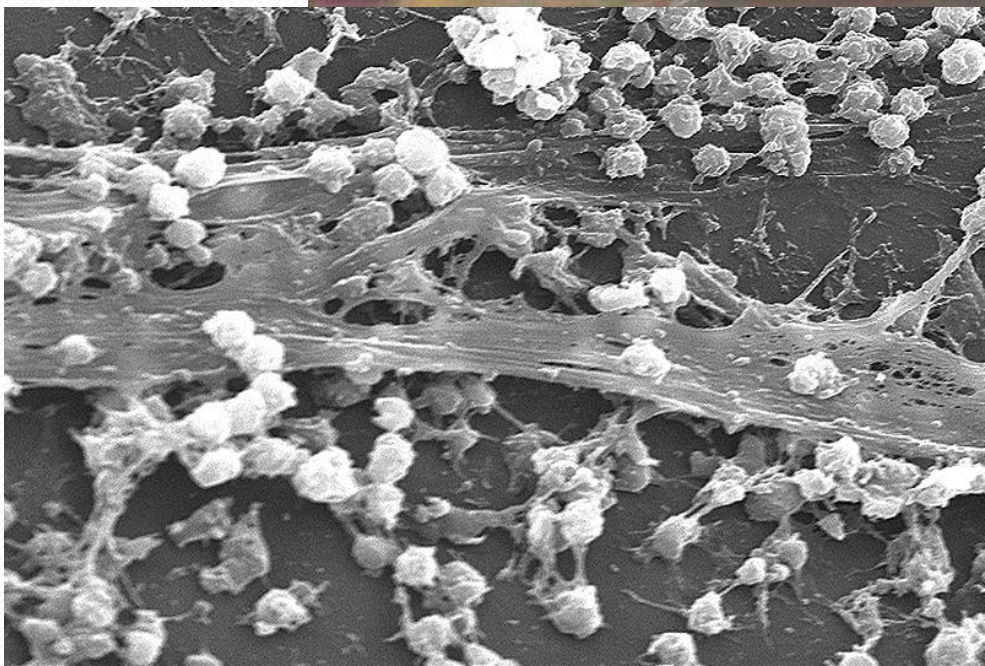
Utility: Total food minus total shocks

Optimal Strategy: if H+ then Left, if H- then Right, else
(if V+ then Down, if V- then Up, else Stationary)

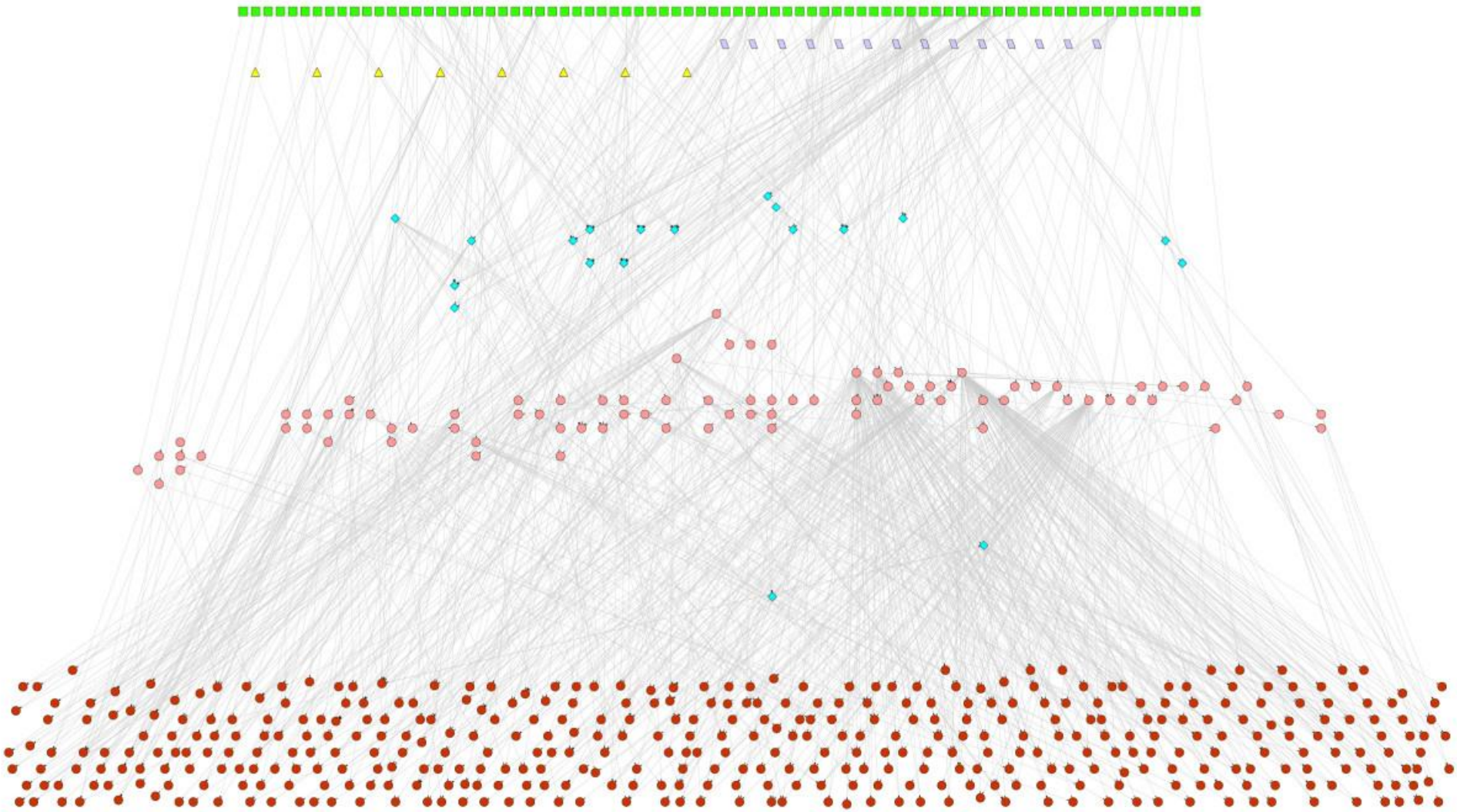
E. Coli

- Billions in your gut
- Only 4,377 genes
- Survives frozen soils
- Stomach acid -> Zen state
- Senses optimal gut location





E. Coli's Brain



External metabolites green, Stimuli yellow, Enzyme genes brown, TFs pink

Simple Learning Systems

Two-armed bandit

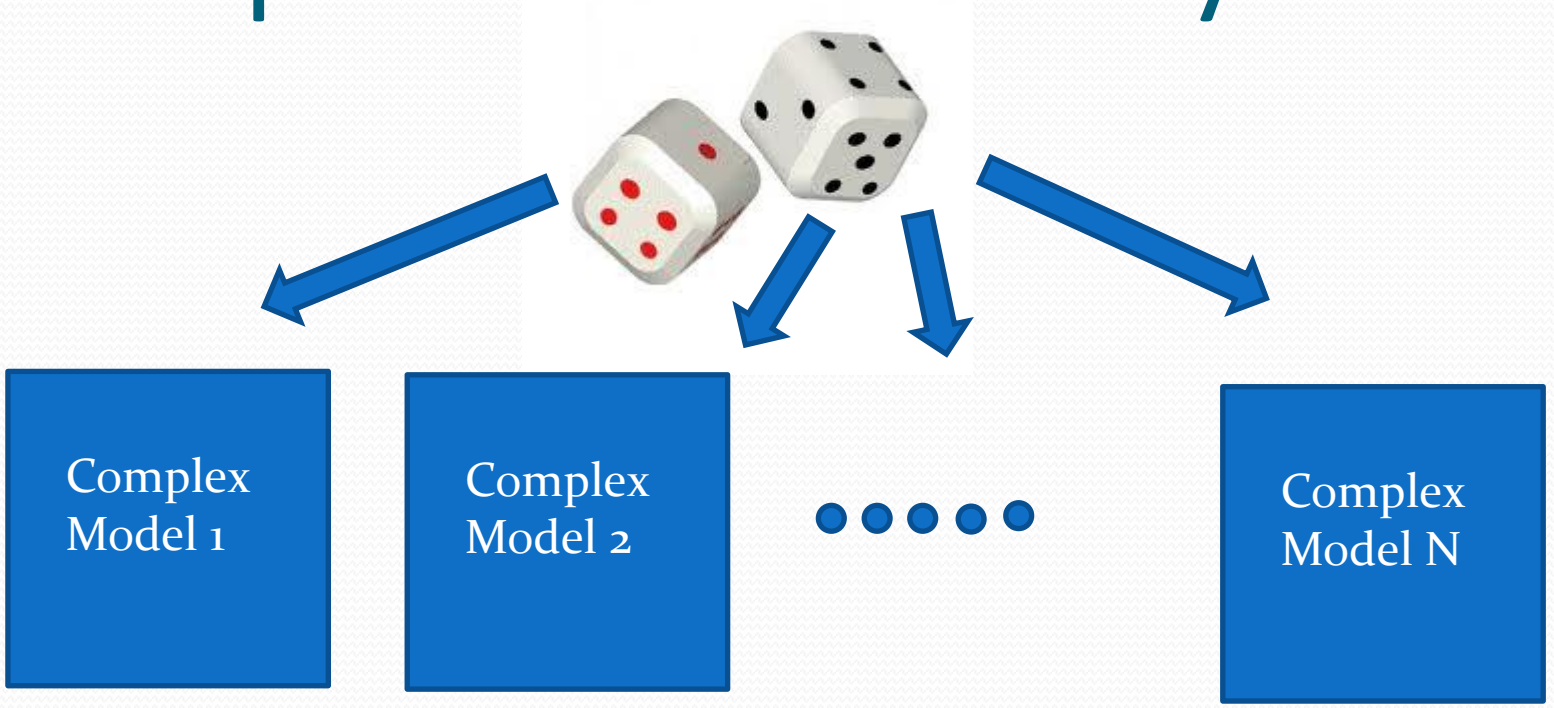
payoff prob. p_1

payoff prob. p_2

Actions: Pull 1, Pull 2
Senses: Payoff, No Payoff
Utility: Total payoff
Prior: Uniform over p_1 and p_2
Optimal strategy: Gittins indices
- Exploration phase: Estimate p_1 and p_2
- Exploitation phase: Only pull the one believed better

The shaper isn't certain about the world so the shaped mind has to learn what's best and then act on it.

Simple Deliberative Systems

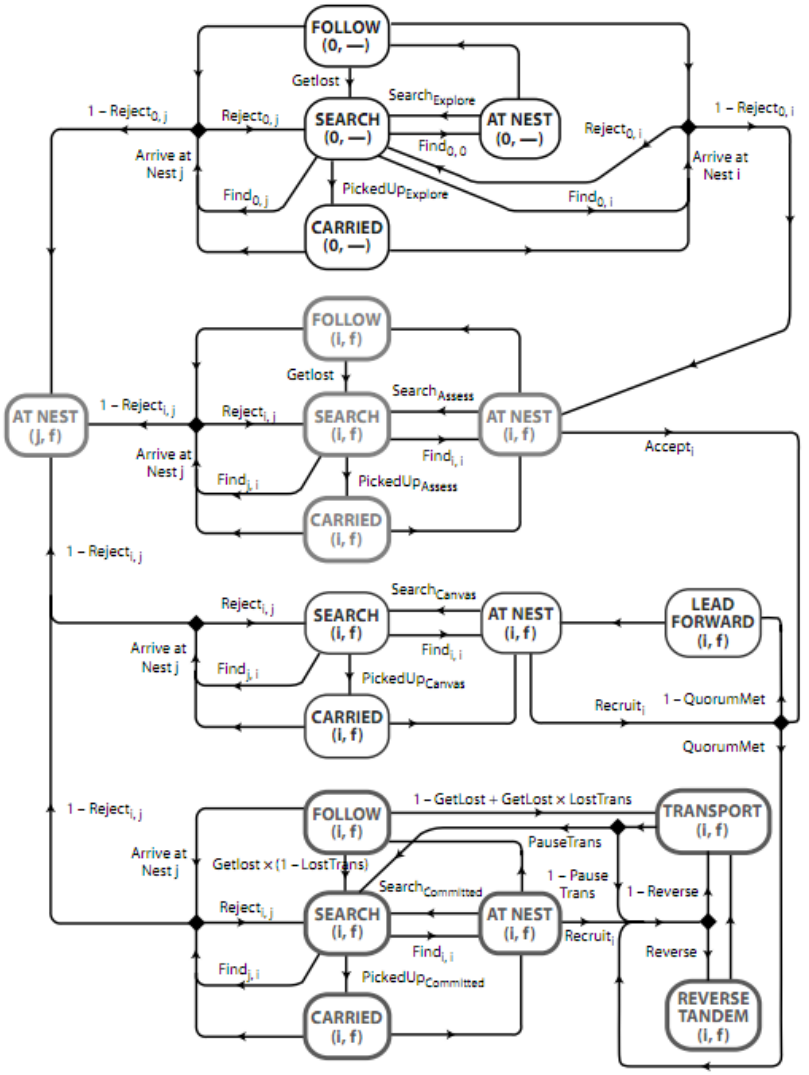


Mixture of complex models
Shaper stores compessed models to be unfolded when type is known
Best representation is formal specification with deliberative search
Pushes toward Kolmogorov complexity of component models

Ant emigration – Pratt 2005

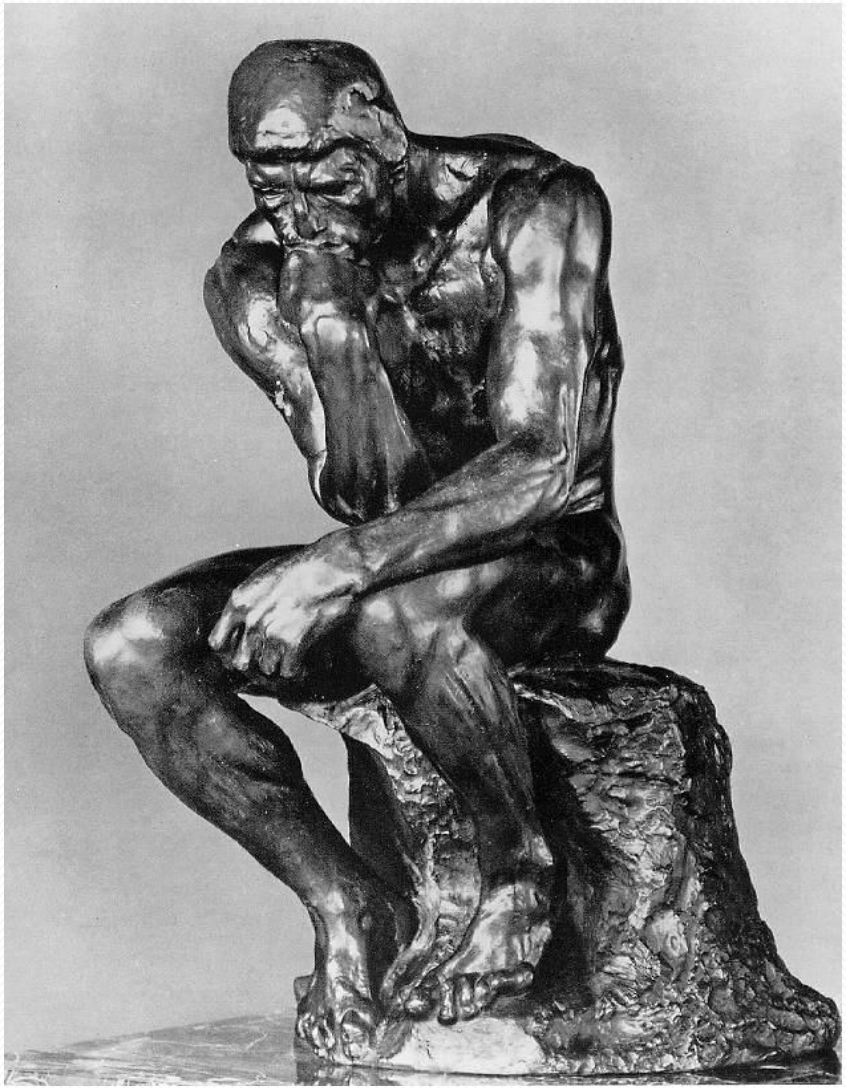
Temnothorax albipennis

- Individuals explore nests
- Once found, evalutates
- If good, leads another there
- Population reaches quorum
- Commits, and transports



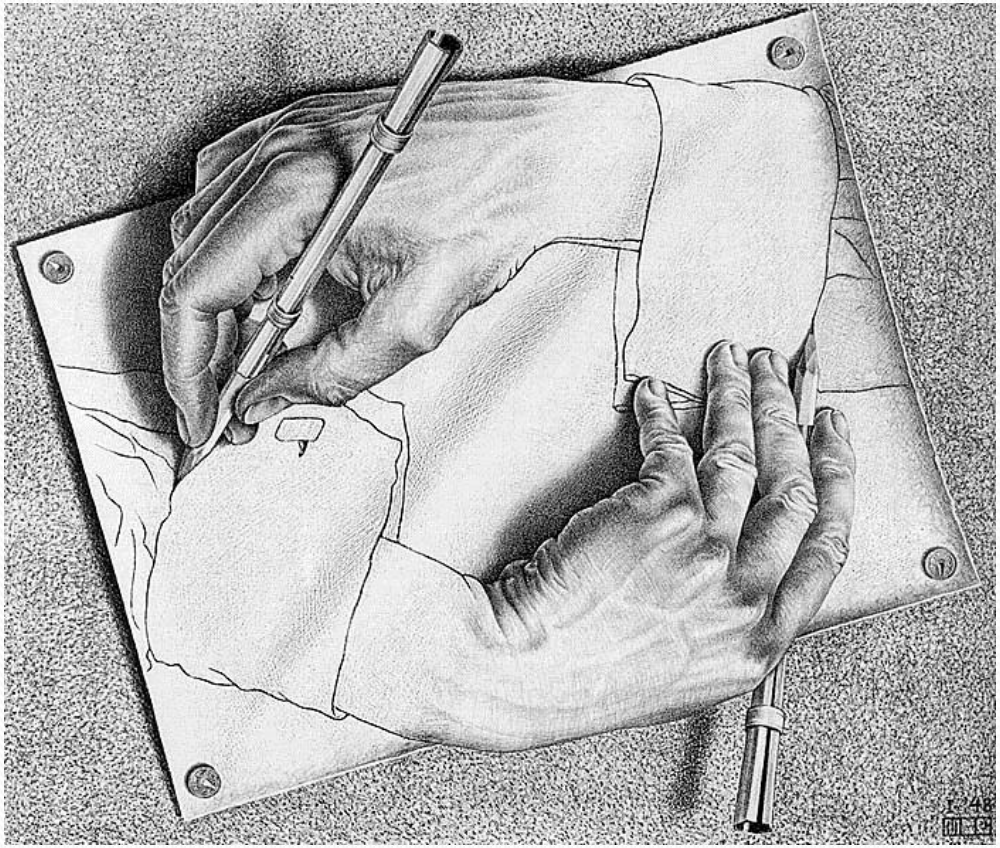
Meta-Reasoning Systems

- Time management
- Choosing training
- Who to interact with
- Self-improvement
- Delegating
- Boredom
- Curiosity

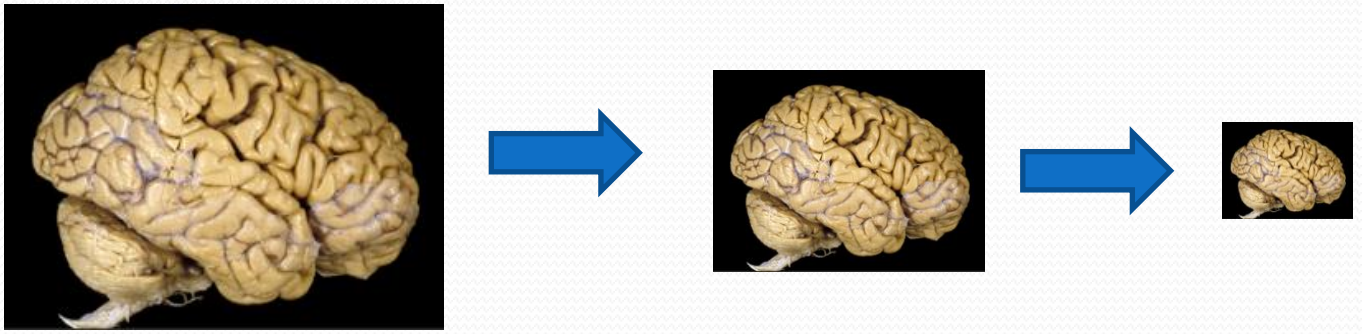


Self-Improving Systems

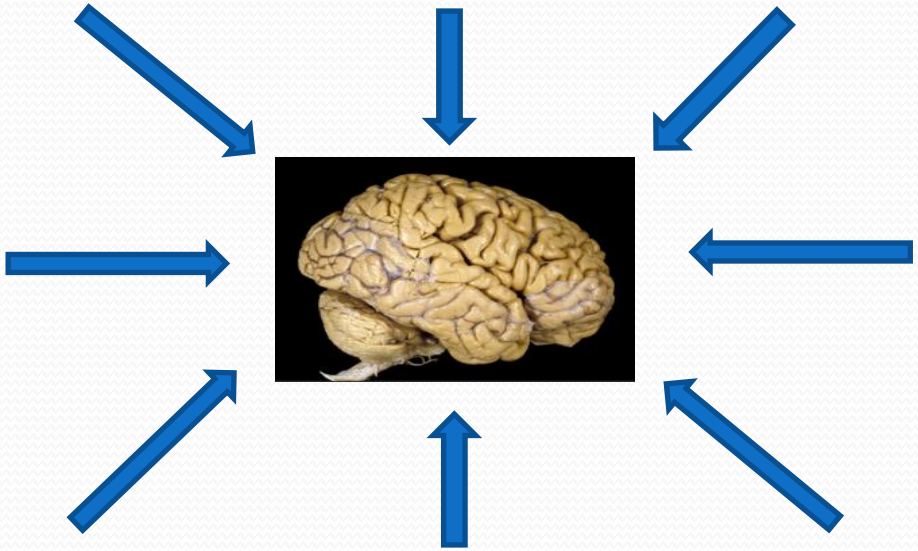
- Systems that model and improve themselves
- Don't yet exist
- Humans try to
- Every limited AI will want to do this
- Fully rational systems have no need



Rationally-Shaped-Shaped Systems

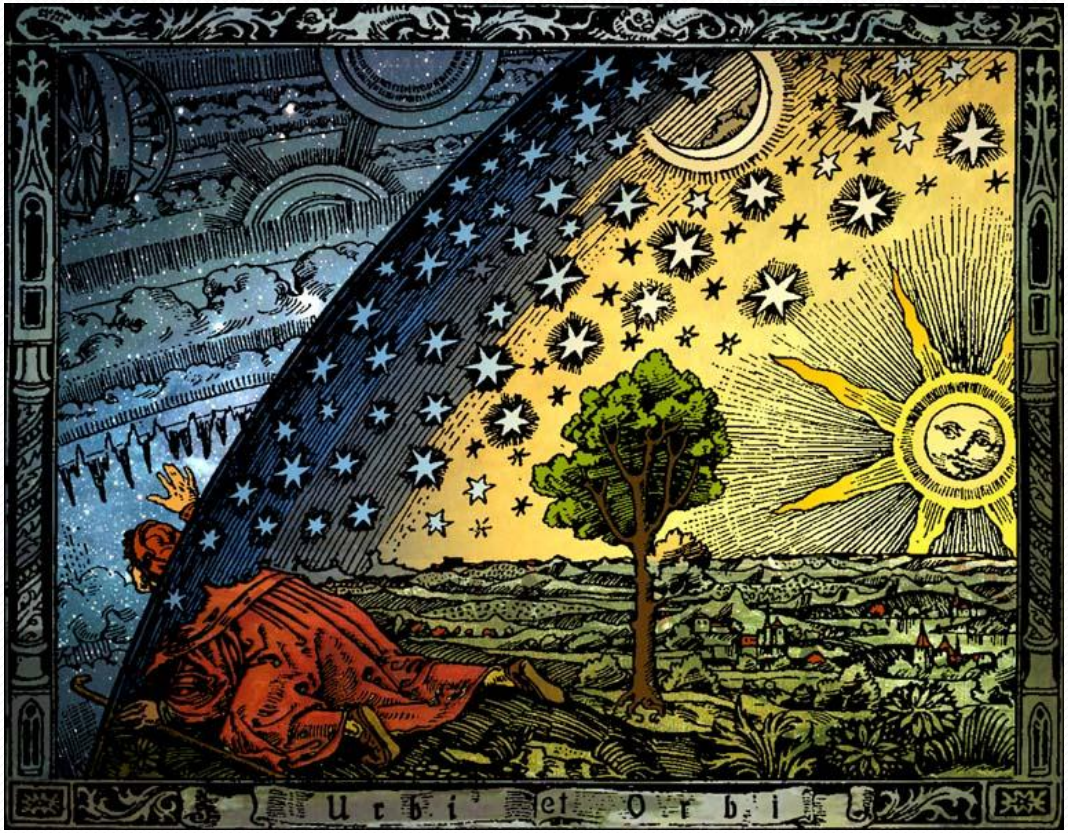


Conjecture: Self-improving minds converge on good approximations to the optimal rational mind for their computational resources.



Extension: Expanding Minds

- Allow actions that increase computational resources.
- Systems which plan for expansion
- Eg. Human genome is only 23000 genes



Rationally-Shaped Systems

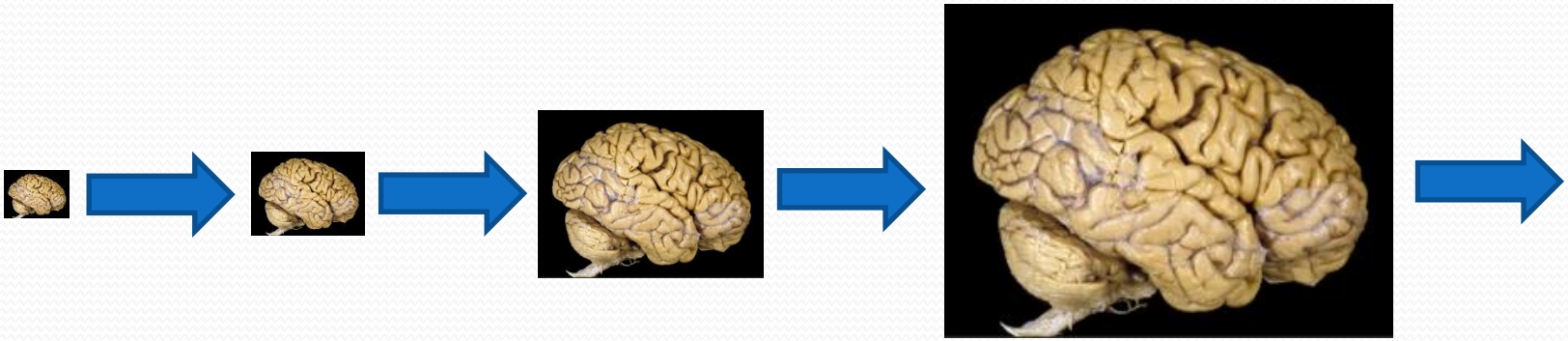
- Exhibit same drives as fully rational systems
- But we can impose desired constraints
- Create safety by controlled irrationality

How can we shape these systems?

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

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

Formal Specification Languages

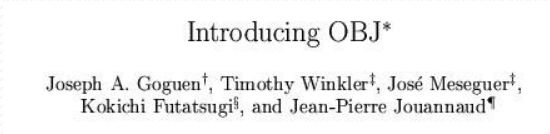
First Order Predicate Calculus

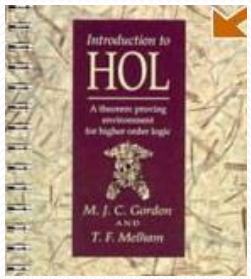
Zermelo-Frankel Set Theory

Category Theory

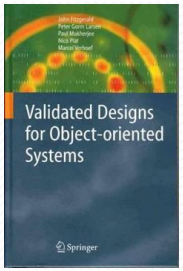
Higher order type theory

Z Notation





Vienna Development Method

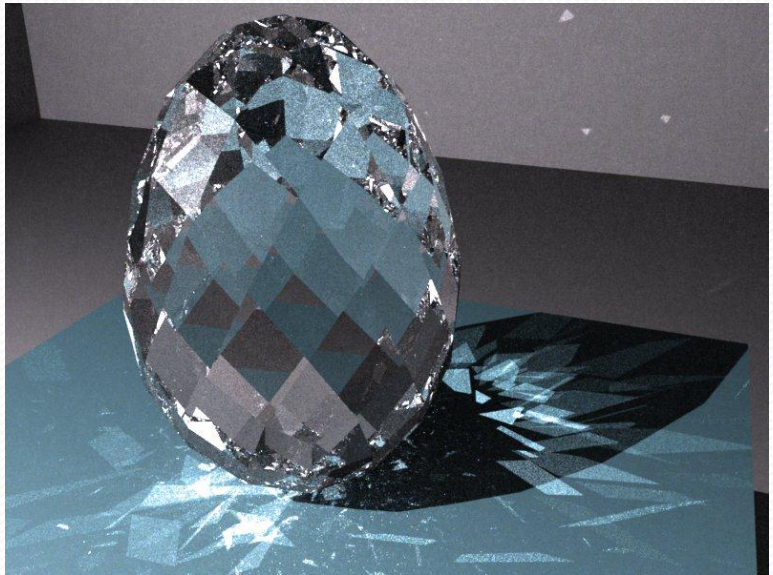


Algebraic Specification

COQ : Formal Specifications

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

- 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.

Huge Benefits for Humanity

- Curing disease and aging
- Modeling and solving social problems
- Knowledge synthesis and access
- Automated discovery and invention
- Provably correct and secure software
- Self-driving cars
- Robot control
- Automated manufacturing
- Improved energy utilization
- …