

Universal Artificial Intelligence

Marcus Hutter

Canberra, ACT, 0200, Australia

<http://www.hutter1.net/>



Australian National University

Abstract

The approaches to Artificial Intelligence (AI) in the last century may be labelled as (a) trying to understand and copy (human) nature, (b) being based on heuristic considerations, (c) being formal but from the outset (provably) limited, (d) being (mere) frameworks that leave crucial aspects unspecified. This decade has spawned the first theory of AI, which (e) is principled, formal, complete, and general. This theory, called Universal AI, is about ultimate super-intelligence. It can serve as a gold standard for General AI, and implicitly proposes a formal definition of machine intelligence. After a brief review of the various approaches to (general) AI, I will give an introduction to Universal AI, concentrating on the philosophical, mathematical, and computational aspects behind it. I will also discuss various implications and future challenges.

Artificial General Intelligence (AGI)

What is the goal of AGI research?

- Build general-purpose *Super-Intelligences*.
- Will ignite the detonation cord to the *Singularity*.



What is (Artificial) Intelligence?

What are we really doing and aiming at?

- Is it to build systems by trial&error, and if they do something we think is smarter than previous systems, call it success?
- Is it to try to mimic the behavior of biological organisms?

**We need (and have!) theories which
can guide our search for intelligent algorithms.**

Focus of This Talk

- Mathematical Foundations of Intelligent Agents
- State-of-the-Art Theory of Machine Super Intelligence
- Implications



What Is Intelligence?

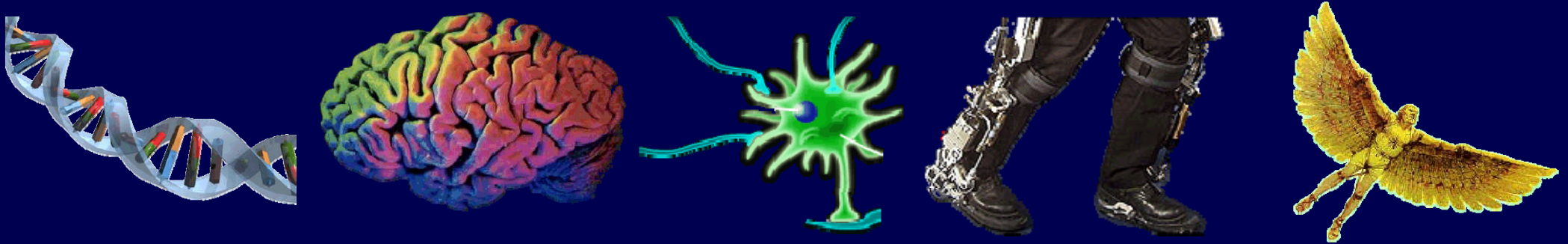
What is AGI?	Thinking	Acting
humanly	Cognitive Science	Turing Test, Behaviorism
rationally	Laws of Thought	Doing the “Right” Thing

Informal Working Definition

Intelligence measures an agent’s ability to perform well in a wide range of environments.

"Natural" Approaches

copy and improve (human) nature



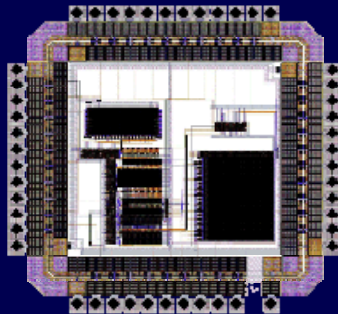
Biological Approaches to Super-Intelligence

- Brain Scan & Simulation
- Genetic Enhancement
- Brain Augmentation

Not the topic of this talk

"Artificial" Approaches

Design from first principles. At best inspired by nature.



Artificial Intelligent Systems:

- Logic/language based: expert/reasoning/proving/cognitive systems.
- Economics inspired: utility, sequential decisions, game theory.
- Cybernetics: adaptive dynamic control.
- Machine Learning: reinforcement learning.
- Information processing: data compression \approx intelligence.

Separately too limited for AGI, but jointly very powerful.

Topic of this talk: Foundations of "artificial" approaches to AGI

Elegant Theory of ...

Cellular Automata → ... Computing

Iterative maps → ... Chaos and Order

QED → ... Chemistry

Super-Strings → ... the Universe

AIXI → ... Super Intelligence

Scientific Foundations of Universal Artificial Intelligence

Contents

- **Philosophical Foundations**
(Ockham, Epicurus, Induction)
- **Mathematical Foundations**
(Information, Complexity, Bayesian & Algorithmic Probability, Solomonoff Induction, Sequential Decisions)
- **Framework: Rational Agents**
(in Known and Unknown Environments)

Science \approx Induction \approx Ockham's Razor

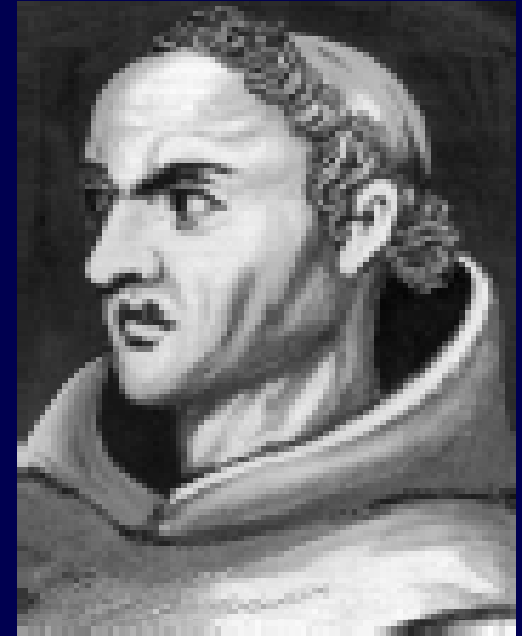
- **Example: Grue Emerald Paradox**

Hypothesis 1: All emeralds are green

Hypothesis 2: All emeralds found until year 2020 are green, thereafter all emeralds will be blue.

- Which hypothesis is more plausible?

Hypothesis 1! Justification?



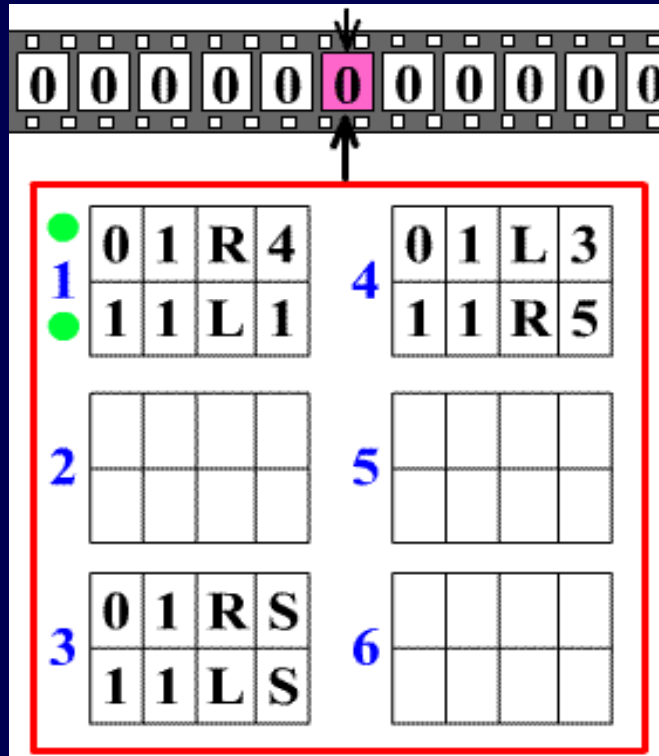
Ockham's Razor Principle =

take the simplest hypothesis consistent with the data

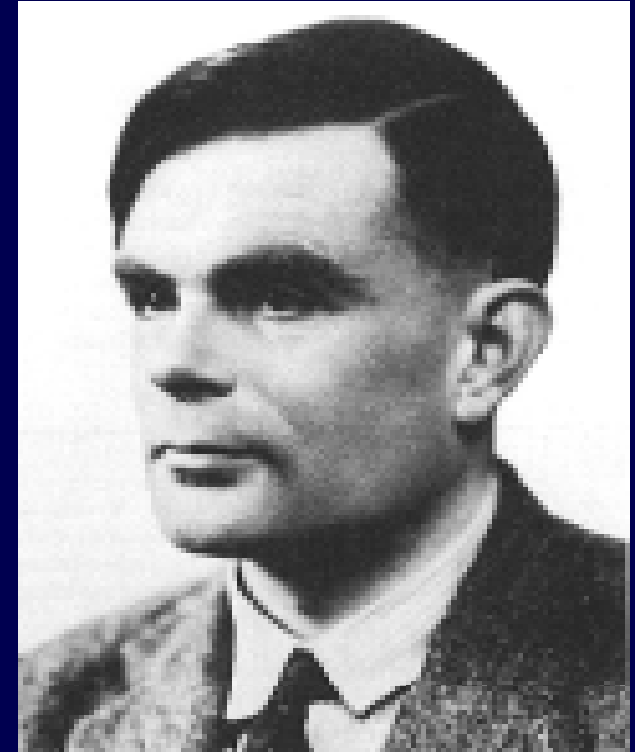
is the most important principle in machine learning and science

Problem: Quantification of Simplicity/Complexity

Turing's Universal Machine *U*



Turing's Thesis



Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine

Algorithmic Information Theory



Quantification of
Simplicity/Complexity
in Ockham's Razor



The **Kolmogorov Complexity** of a string is the length of the shortest program on U describing this string:

$$K(x) := \min_p \{ \text{Length}(p) : U(p) = x \}$$

Bayesian Probability Theory

Bayes Rule

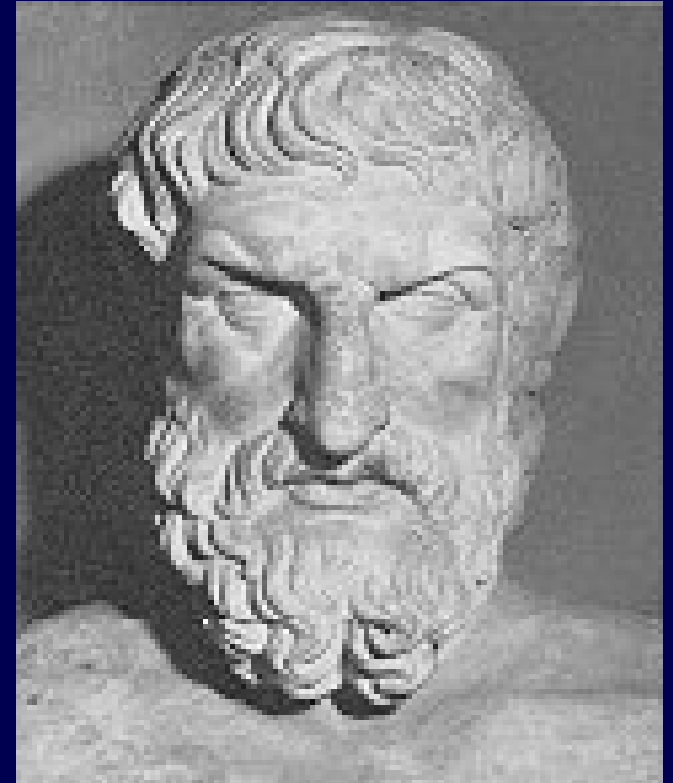
$$\Pr(H|D) \propto \Pr(D|H) \times \Pr(H)$$



Bayes Rule allows to update prior degree of belief in hypothesis H , given new observations D , to posterior belief in H .

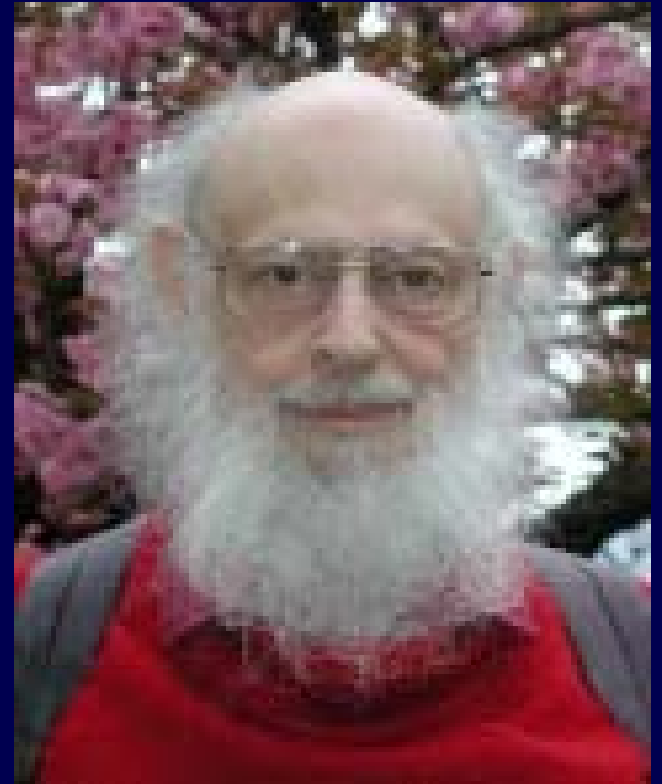
Algorithmic Probability

- **Epicurus:** If more than one theory=hypothesis=model is consistent with the observations, keep them all.
- **Refinement with Ockham:** Give simpler theories higher a-priori weight.
- **Quantitative:** $\Pr(H) := 2^{-K(H)}$



Universal Induction

Solomonoff combined
Ockham, *Epicurus*,
Bayes, and *Turing* into
one formal theory of
sequential prediction



- **Universal a-priori probability:**
 $M(x)$:= probability that U fed with noise outputs x .
- $M(x_{t+1}|x_1 \dots x_t)$ best predicts x_{t+1} from $x_1 \dots x_t$.

Sequential Decision Theory = Optimal Control Theory

For $t=1,2,3,4,\dots$

Given sequence x_1, x_2, \dots, x_{t-1}

(1) Make decision y_t

(2) Observe x_t

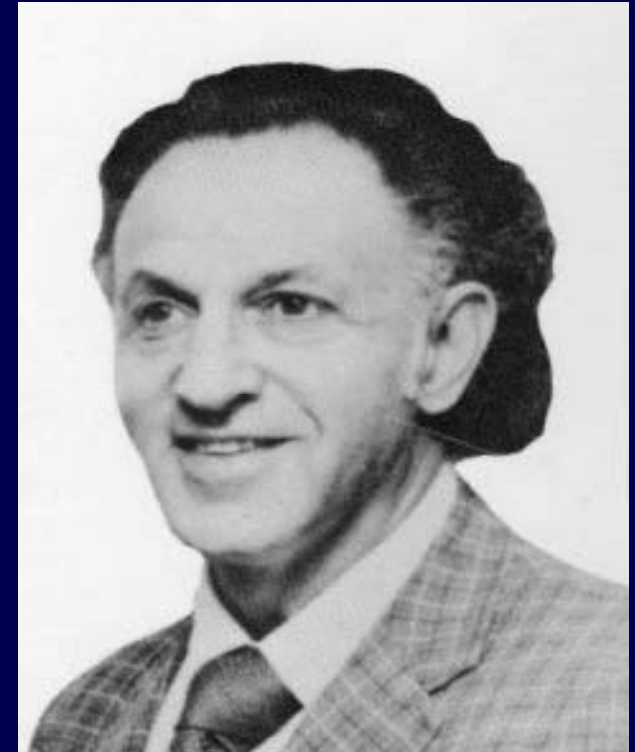
(3) Suffer $\text{Loss}(x_t, y_t)$

(4) $t \rightarrow t+1$, goto (1)

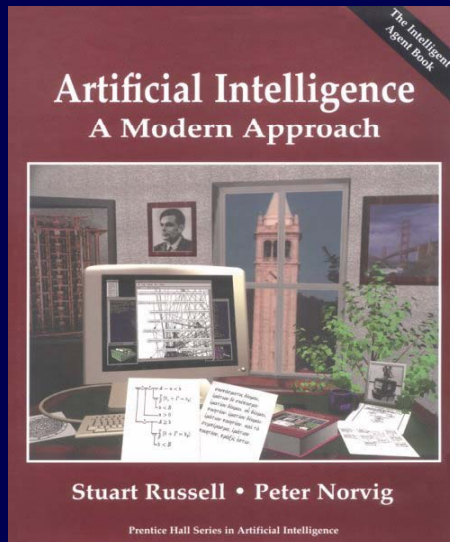
Goal: Minimize expected Loss

Problem: True probability unknown

Solution: Use Solomonoff's $M(x)$

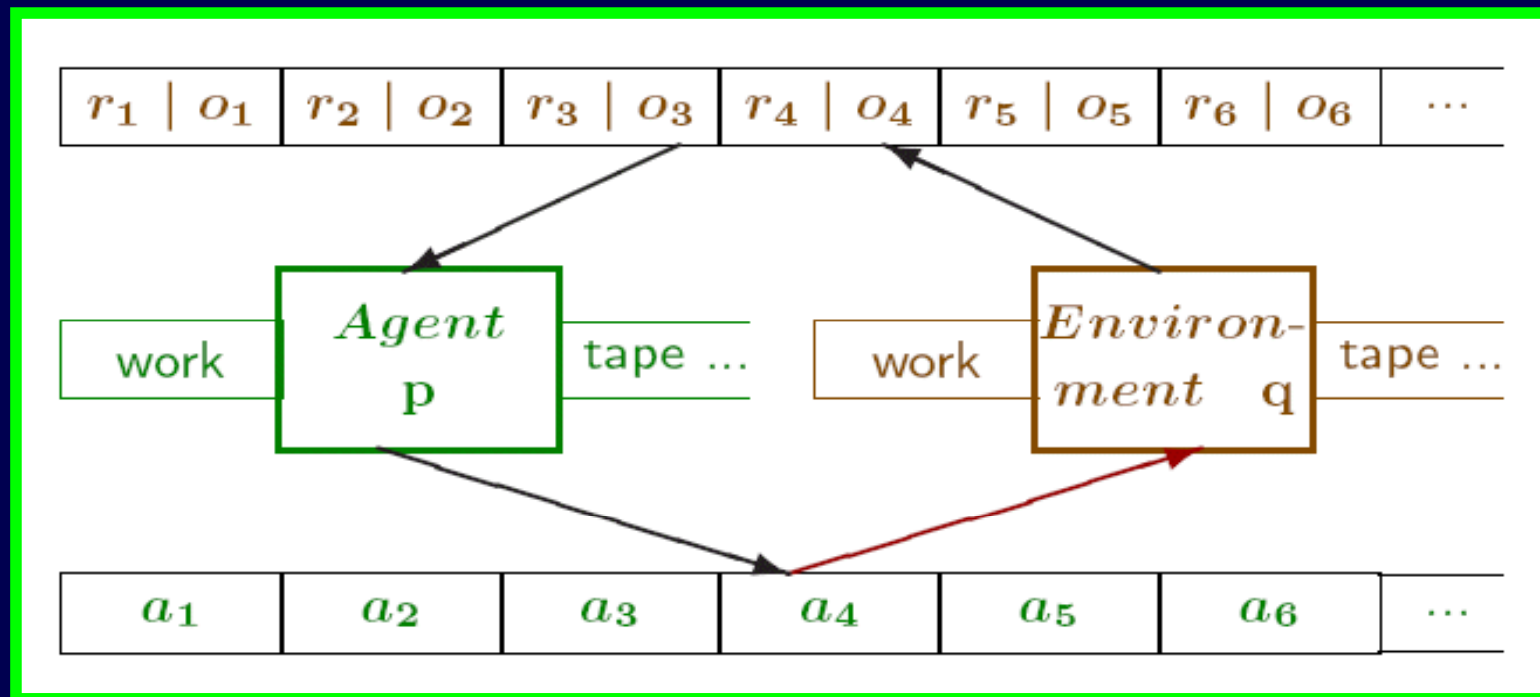


(Richard Bellman)



Agent Model with reward

extremely general framework



Now we can put everything together ...

Universal Artificial Intelligence

complete & essentially unique & limit-computable

$$\text{AIXI} \quad a_k := \arg \max_{a_k} \sum_{o_k r_k} \dots \max_{a_m} \sum_{o_m r_m} [r_k + \dots + r_m] \sum_{q: U(q, a_1 \dots a_m) = o_1 r_1 \dots o_m r_m} 2^{-\ell(q)}$$

action, reward, observation, Universal TM, program, k=now

- AIXI is an elegant & sound math. theory of AGI.
- AIXI is a universally optimal rational agent.
- AIXI is the ultimate Super Intelligence, but
- AIXI is computationally intractable, however,
- AIXI can serve as a gold standard for AGI.

Computational Issues

Contents

- **Towards Practical Universal AI**
(Additional Required Ingredients)
- **Monte-Carlo AIXI Applications**
(AIXI learns to play Poker, Pacman, TicTacToe,...)
- **Feature Reinforcement Learning**
(Alternative approach towards practical AGI)
- **Intelligent Agents in Perspective**
(Relation between Fields and Approaches)

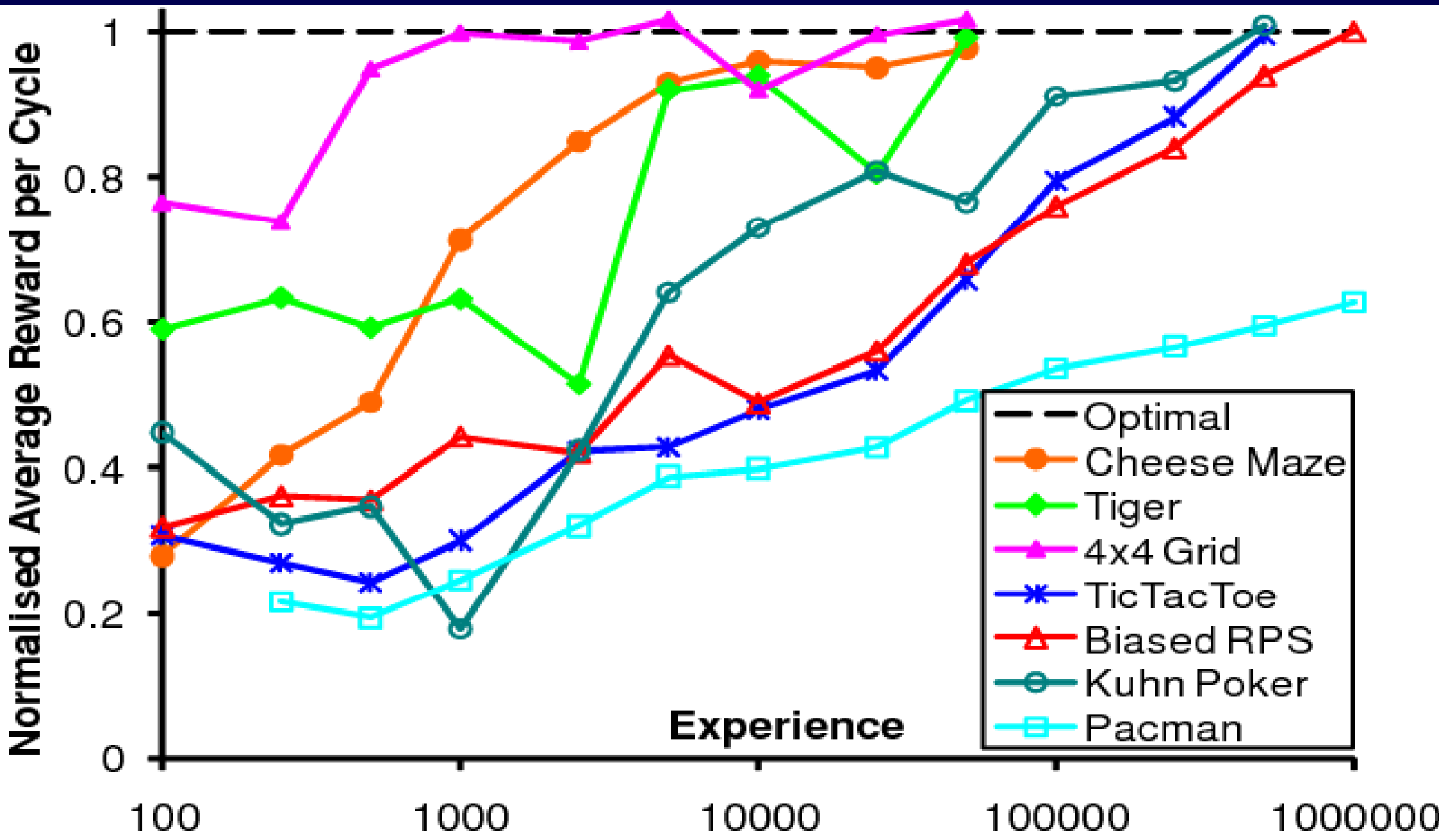
Towards Practical Universal AI

Goal: Develop *efficient* general-purpose intelligent agent

- Additional Ingredients: Main Reference (year)
- Universal search: Schmidhuber (200X) & al.
- Learning: TD/RL Sutton & Barto (1998) & al.
- Information: MML/MDL Wallace, Rissanen
- Complexity/Similarity: Li & Vitanyi (2008)
- Optimization: Aarts & Lenstra (1997)
- Monte Carlo: Fishman (2003), Liu (2002)

No time for details, so let's go directly to the state-of-the-art:

Monte-Carlo AIXI Applications



Feature Reinforcement Learning

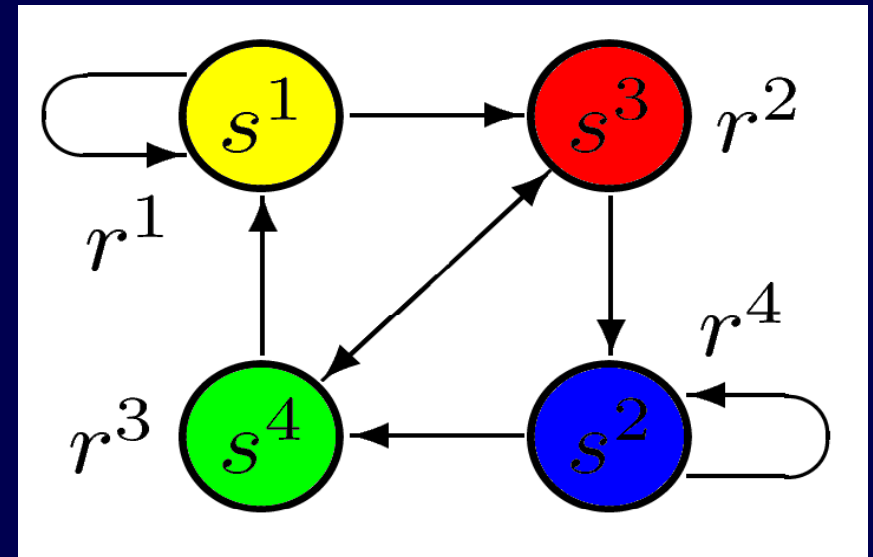
Goal: Develop *efficient* general-purpose intelligent agent

Real-world Problem



learn
→
reduction

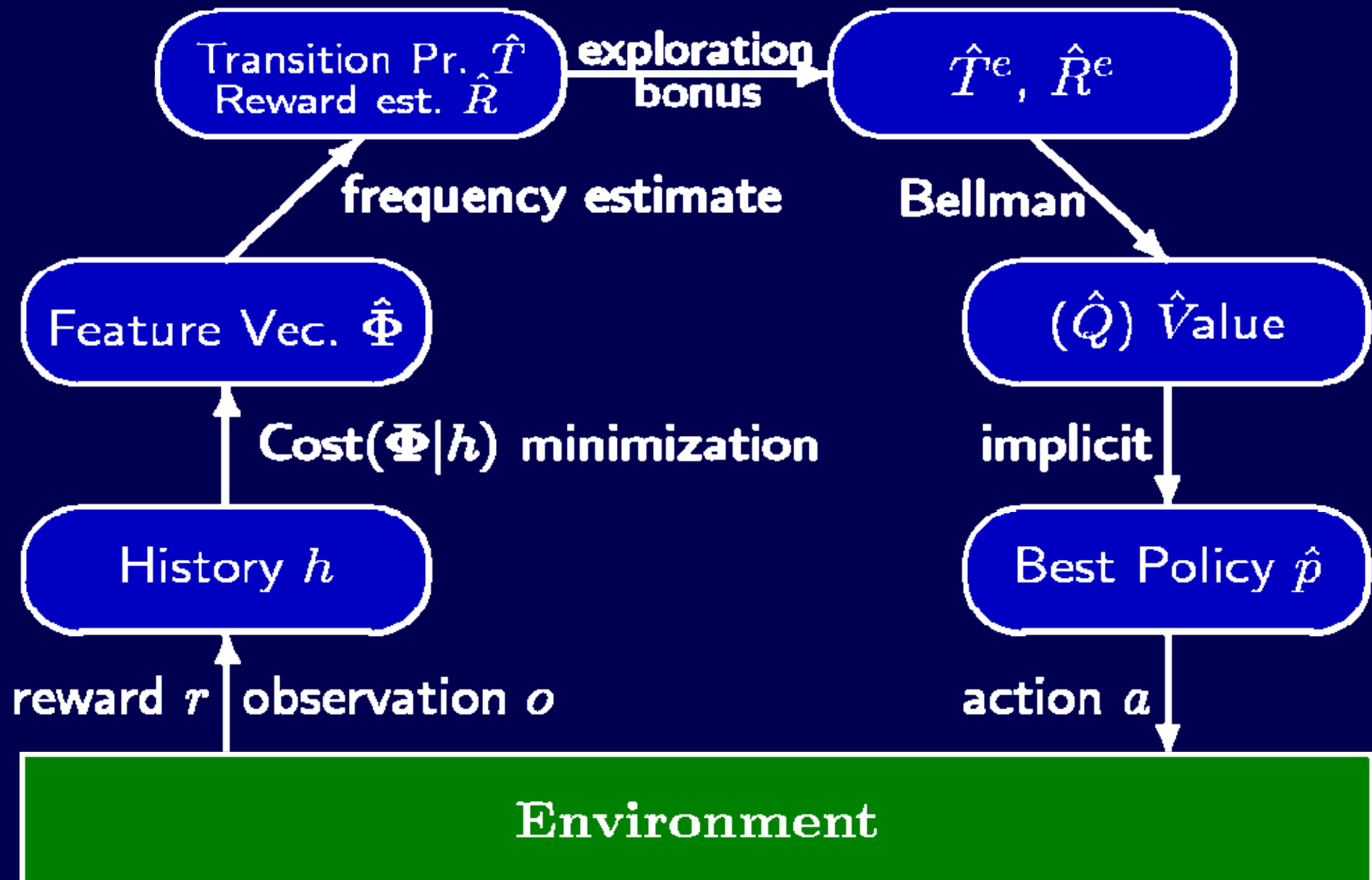
Markov Decision Process



FRL Approach: reduces complex real-world problem to tractable structured Markov Decision Process (MDP) automatically by learning relevant features.

Structured MDP \approx Dynamic Bayesian Network \approx Neural Network \approx Memory

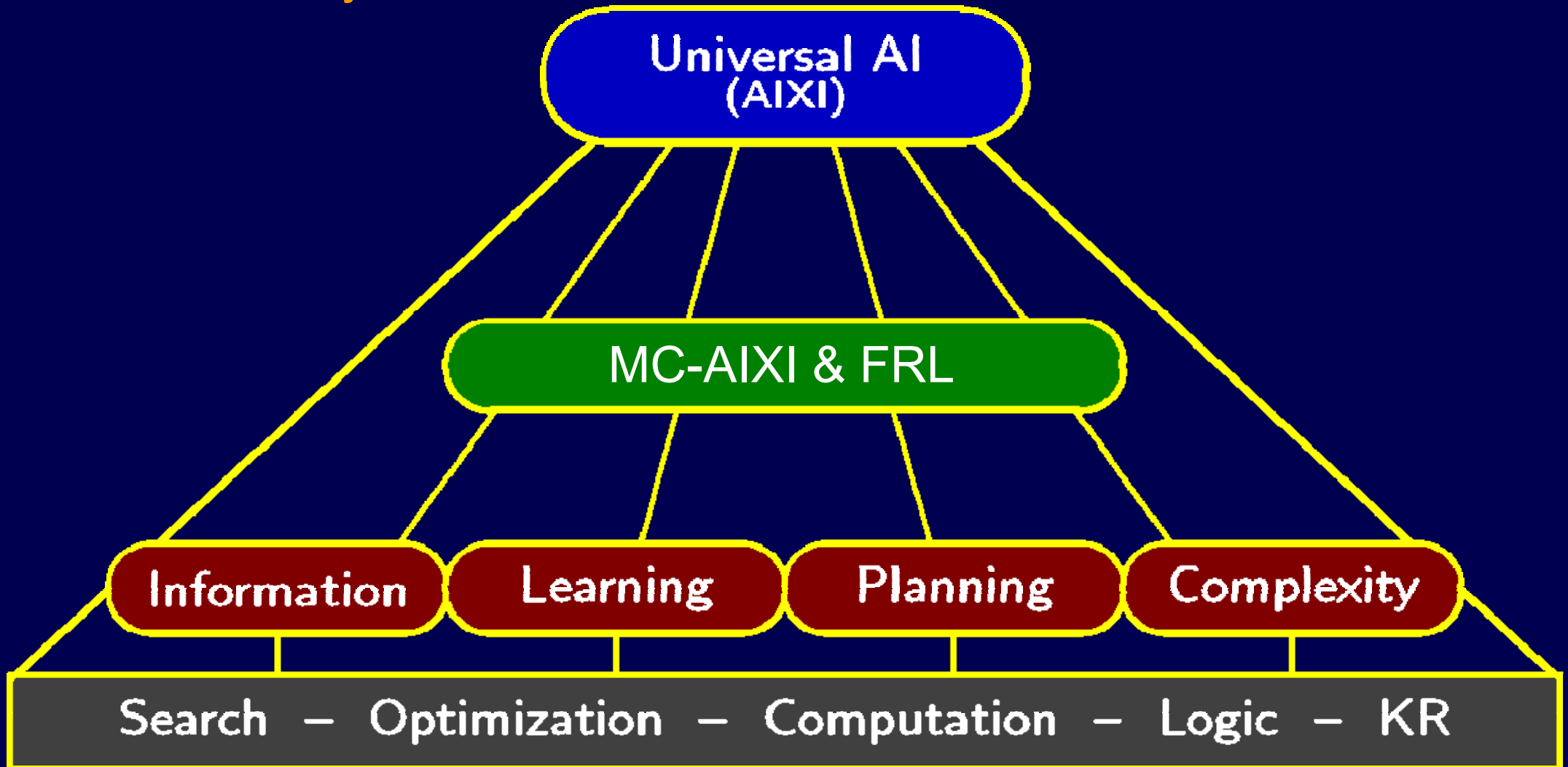
FRL: Computational Flow



Intelligent Agents in Perspective

Summary

Slide



Agents = General Framework, Interface = Robots, Vision, Language

Discussion

Contents

- Traits of (Artificial) Intelligence
- Social Behavior of AIXI
- Questions / Claims / Challenges / Outlook
- References

Traits of (Artificial) Intelligence

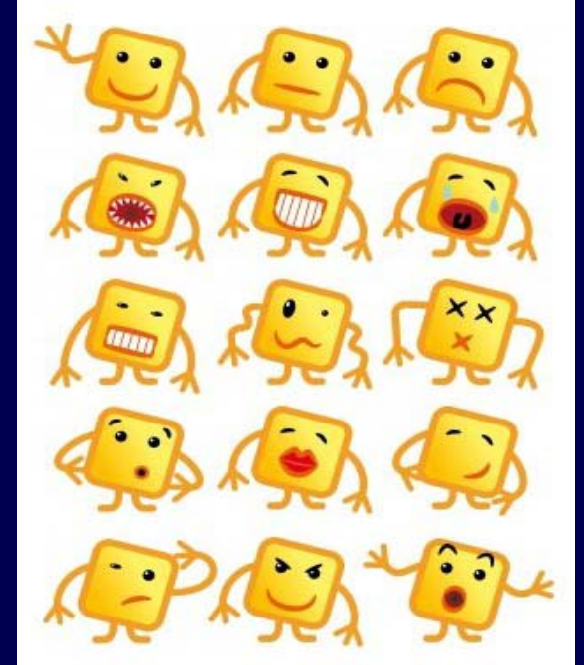
- reasoning
- creativity
- association
- generalization
- pattern recognition
- problem solving
- memorization
- planning under uncertainty
- achieving goals
- learning
- integration
- optimization
- self-preservation
- vision
- natural language processing

These are or can be shown to be emergent traits of AIXI

Other Aspects of the Human Mind



- Consciousness
- Self-awareness
 - Sentience
 - Emotions



If these qualia are relevant for rational decision making, then they should be emergent traits of AI_{XI} too.

Some Social Behavior of AIXI

(reasonable conclusions but not yet formally verified)

- **Drugs (hack reward system)**
 - Virtual: not possible
 - Embodied: no, since long-term reward would be small (death). Counter: Orseau (2011)
- **Procreate:** yes, if AIXI believes that descendants are useful (ensure retirement pension)
- **Suicide:** if can be raised to believe to get to heaven (hell), then yes (no).
- **Self-Improvement:** Yes

What will an AI χ I Singularity look like?

- AI χ I is already completely and essentially uniquely defined.
- ➔ first model for which such questions might be answered rigorously.
(not just trusting our intuitive arguments)

Maybe the questions in some of the following slides can be answered too.

Questions

- Will the natural or the artificial approach win the race toward the singularity?
- How much has to be designed and what can be learnt?
- What is intelligence in absence of a reward concept?
- Will reward maximizers (AIXI) prevail against assimilators (the Borg)?
- Intelligence is upper bounded (by AIXI). Will this prevent a singularity?

Scientific Challenges / Outlook

- What can we (not) expect from AIXI
- Practical approximations of AIXI
- Efficient optimizations of $\text{Cost}()$ in FRL
- Flexible structure learning in FRL
- Devising appropriate training sequences

Summary

- **Theories** are necessary to guide our search for AGI.
- **Intelligence** measures an agent's ability to perform well in a wide range of environments.
- **Universal AI** is an elegant, principled, formal, and complete theory of AGI.
- **AIXI** is an optimal reinforcement learning agent embedded in an arbitrary unknown environment, but is incomputable.
- **Key ingredients:** Ockham, Epicurus, Bayes, Turing, Kolmogorov, Solomonoff, Bellman.
- **MC-AIXI-CTW and FRL** are tractable approximations of AIXI.
- **FRL** reduces the Real World to MDPs.

(Some) AGI research has become a **formal science**

Thanks! Questions? Details:



– S.Legg. Machine Super Intelligence. 2008



– M.H. One Decade of Universal AI. 2012



– J.Veness&al. A MC AIXI Approximation. 2011



– Human Knowledge Compression Prize. 2006



– PhD Students: Please apply at ANU



– Research funding offers are welcome