Unifying Foundations of Intelligent Agents

Marcus Hutter

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

Abstract

The dream of creating artificial devices that reach or outperform human intelligence is many centuries old. In this talk I present an elegant parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. The theory reduces all conceptual AI problems to pure computational questions. The necessary and sufficient ingredients are Bayesian probability theory; algorithmic information theory; universal Turing machines; the agent framework; sequential decision theory; and reinforcement learning, which are all important subjects in their own right. I also present some recent approximations, implementations, and applications of this modern top-down approach to AI.

Overview

Goal: Construct a single universal agent
that learns to act optimally in any environment.

State of the art: Formal (mathematical, non-comp.) definition
of such an agent.

Accomplishment: Well-defines AI. Formalizes rational intelligence.
Formal "solution" of the AI problem in the sense of ...

$\Longrightarrow$ Reduces the conceptional AI problem
to a (pure) computational problem.

Evidence: Mathematical optimality proofs
and some experimental results.

Contents

- Philosophical and Mathematical Background.
- Universal Intelligence Measure.
- The Ultimate Intelligence. The AIXI Agent.
- Summary and References.

PHILOSOPHICAL AND MATHEMATICAL BACKGROUND

What is (Artificial) Intelligence?

Intelligence can have many faces ⇒ formal definition difficult

- reasoning
- creativity
- association
- generalization
- pattern recognition
- problem solving
- memorization
- planning
- achieving goals
- learning
- optimization
- self-preservation
- vision
- language processing
- classification
- induction
- deduction
- ...

What is AI?	**Thinking**	**Acting**
humanly	Cognitive Science	Turing test, Behaviorism
rationally	Laws Thought	Doing the Right Thing

Collection of 70+ Defs of Intelligence
`http://www.vetta.org/definitions-of-intelligence/`

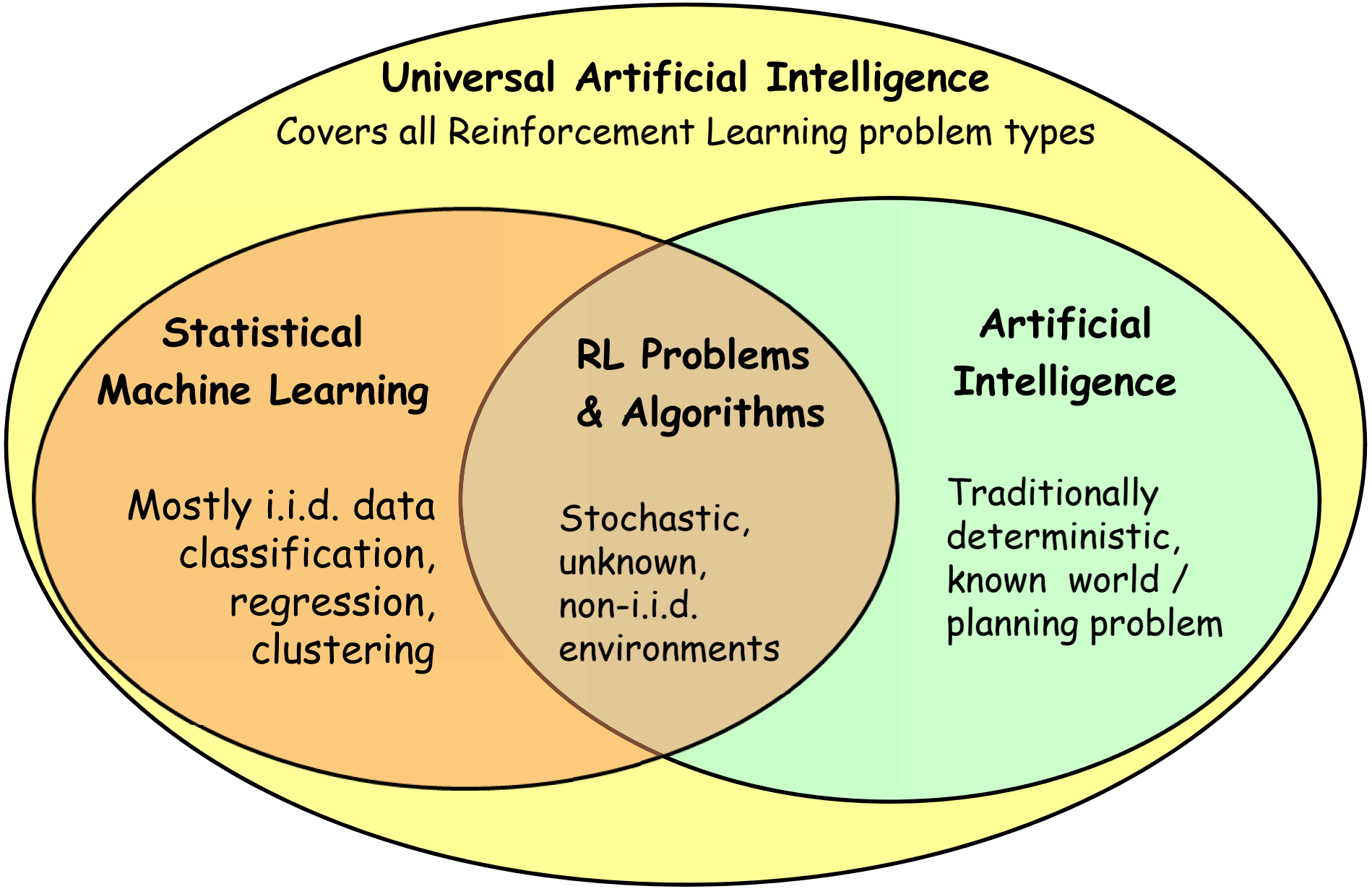
Real world is nasty: partially unobservable, uncertain, unknown, non-ergodic, reactive, vast, but luckily structured, ...

Relevant Research Fields

(Universal) Artificial Intelligence has interconnections with (draws from and contributes to) many research fields:

- computer science (artificial intelligence, machine learning),
- engineering (information theory, adaptive control),
- economics (rational agents, game theory),
- mathematics (statistics, probability),
- psychology (behaviorism, motivation, incentives),
- philosophy (reasoning, induction, knowledge).

Relation between ML & RL & (U)AI



Informal Definition of (Artificial) Intelligence

Intelligence measures an agent's ability to achieve goals in a wide range of environments. [S. Legg and M. Hutter]

Emergent: Features such as the ability to learn and adapt, or to understand, are implicit in the above definition as these capacities enable an agent to succeed in a wide range of environments.

The science of Artificial Intelligence is concerned with the construction of intelligent systems/artifacts/agents and their analysis.

What next? Substantiate all terms above: agent, ability, utility, goal, success, learn, adapt, environment, ...

Never trust a ~~theory~~ experiment if it is not supported by an ~~experiment~~ theory

Induction→Prediction→Decision→Action

Having or acquiring or *learning* or *inducing* a model of the environment an agent interacts with allows the agent to make *predictions* and utilize them in its *decision* process of finding a good next *action*.

Induction infers general models from specific observations/facts/data, usually exhibiting regularities or properties or relations in the latter.

Example

Induction: Find a model of the world economy.

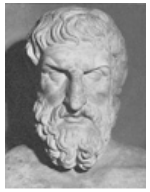
Prediction: Use the model for predicting the future stock market.

Decision: Decide whether to invest assets in stocks or bonds.

Action: Trading large quantities of stocks influences the market.

Foundations of Universal Artificial Intelligence

Ockhams' razor (simplicity) principle
Entities should not be multiplied beyond necessity.



Epicurus' principle of multiple explanations
If more than one theory is consistent with the observations, keep all theories.

Bayes' rule for conditional probabilities
Given the prior belief/probability one can predict all future probabilities. Posterior$(H|D) \propto$ Likelihood$(D|H) \times$ Prior(H).

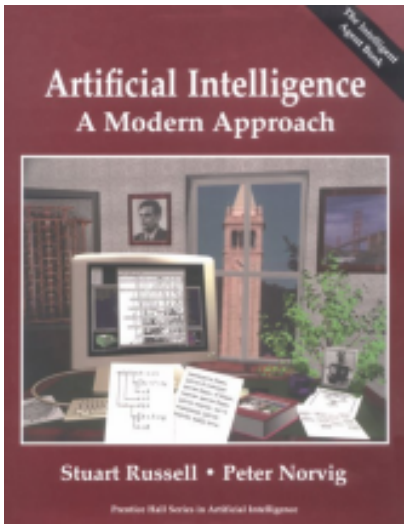
Turing's universal machine
Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine.

Kolmogorov's complexity
The complexity or information content of an object is the length of its shortest description on a universal Turing machine.

Solomonoff's universal prior=Ockham+Epicurus+Bayes+Turing
Solves the question of how to choose the prior if nothing is known. $\Rightarrow$ universal induction, formal Ockham. Prior$(H) = 2^{-\text{Kolmogorov}(H)}$

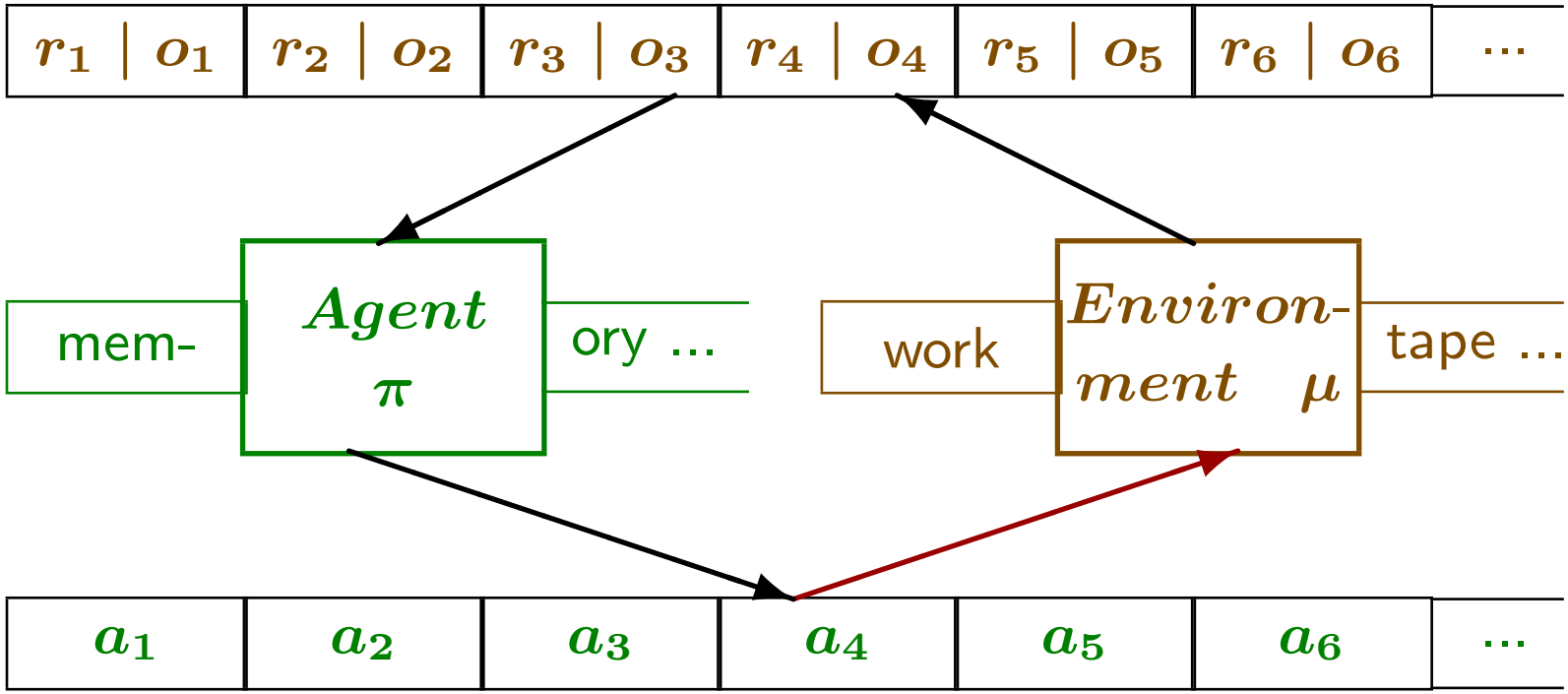
Bellman equations
Theory of how to optimally plan and act in known environments.
Solomonoff + Bellman = **Universal Artificial Intelligence.**

UNIVERSAL INTELLIGENCE MEASURE



Agent Model with Reward

Most if not all AI problems can be formulated within the agent framework



Reinforcement Learning is Extremely General

works in animals, humans, robots, and software agents

- playing games such as checkers, backgammon, go, jeopardi, ...
- playing sports such as soccer, tennis, ...
- learning languages, recognizing faces
- flying a helicopter, driving a car
- navigating a robot through a maze
- planning and scheduling tasks
- making money on the stock market
- answering questions on an IQ test
- passing a Turing test
- ...

Formal Definition of Intelligence

- Agent follows policy $\pi : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^* \rightsquigarrow \mathcal{A}$
- Environment reacts with $\mu : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^*\times\mathcal{A} \rightsquigarrow \mathcal{O}\times\mathcal{R}$
- Performance of agent π in environment μ
  $=$ expected cumulative reward $= V_\mu^\pi := \mathbb{E}_\mu^\pi[\sum_{t=1}^\infty r_t^{\pi\mu}]$
- True environment μ unknown
  $\Rightarrow$ average over wide range of environments
- Ockham+Epicurus: Weigh each environment with its Kolmogorov complexity $K(\mu) := \min_p\{length(p) : U(p) = \mu\}$
- Universal intelligence of agent π is $\Upsilon(\pi) := \sum_\mu 2^{-K(\mu)} V_\mu^\pi$.
- Compare to our informal definition: Intelligence measures an agent's ability to perform well in a wide range of environments.
- AIXI $= \arg\max_\pi \Upsilon(\pi) =$ most intelligent agent.

Is Universal Intelligence Υ any Good?

- Captures our informal definition of intelligence.
- Incorporates Occam's razor.
- Very general: No restriction on internal working of agent.
- Correctly orders simple adaptive agents.
- Agents with high Υ like AIXI are extremely powerful.
- Υ spans from very low intelligence up to ultra-high intelligence.
- Practically meaningful: High Υ = practically useful.
- Non-anthropocentric: based on information & computation theory. (unlike Turing test which measures humanness rather than int.)
- Simple and intuitive formal definition: does not rely on equally hard notions such as creativity, understanding, wisdom, consciousness.

Υ is valid, informative, wide range, general, dynamic, unbiased, fundamental, formal, objective, fully defined, universal.

THE ULTIMATE INTELLIGENCE:
THE AIXI AGENT

The AIXI Model in one Line

complete & essentially unique & limit-computable

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

k=now, $a$ction, $o$bservation, $r$eward, Universal TM, $p$rogram, m=lifespan

AIXI is an elegant mathematical theory of general AI,

but incomputable, so needs to be approximated in practice.

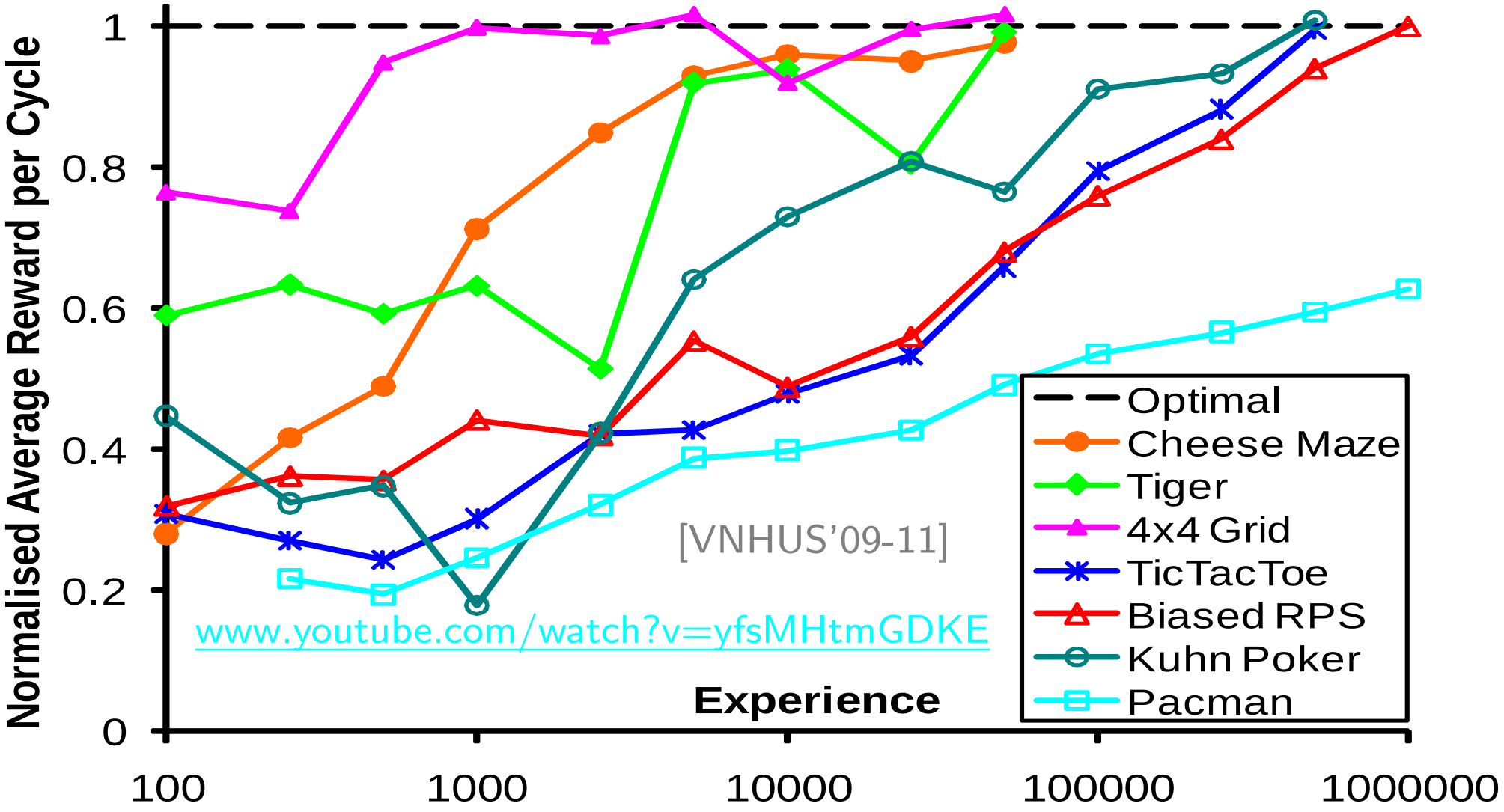
Claim: AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible.

Proof: For formalizations, quantifications, and proofs, see [Hut05].

Potential Applications: Agents, Games, Optimization, Active Learning, Adaptive Control, Robots.

Monte-Carlo AIXI Applications

without providing any domain knowledge, the same agent is able to self-adapt to a diverse range of interactive environments.



Aspects of Intelligence

are all(?) either directly included in AIXI or are emergent

Trait of Intell.	How included in AIXI
reasoning	to improve internal algorithms (emergent)
creativity	exploration bonus, randomization, ...
association	for co-compression of similar observations
generalization	for compression of regularities
pattern recognition	in perceptions for compression
problem solving	how to get more reward
memorization	storing historic perceptions
planning	searching the expectimax tree
achieving goals	by optimal sequential decisions
learning	Bayes-mixture and belief update
optimization	compression and expectimax
self-preservation	by coupling reward to robot components
vision	observation=camera image (emergent)
language	observation/action = audio-signal (emergent)
motor skills	action = movement (emergent)
classification	by compression
induction	Universal Bayesian posterior (Ockham's razor)
deduction	Correctness proofs in AIXItl

Mortal Embodied (AIXI) Agent

- Robot in human society: reward the robot according to how well it solves the tasks we want it to do, like raising and safeguarding a child. In the attempt to maximize reward, the robot will also maintain itself.
- Robot w/o human interaction (e.g. on Alpha-Centauri):
  Some rudimentary capabilities (which may not be that rudimentary at all) are needed to allow the robot to at least survive.
  Train the robot first in safe environment, then let it loose.
- Drugs (hacking the reward system):
  No, since long-term reward would be small (death). but see [OR11]
- Replication/procreation: Yes, if AIXI believes that clones or descendants are useful for its own goals (ensure retirement pension).
- Suicide: Yes (No), if AIXI can be raised to believe to go to heaven (hell). see also [RO11]
- Self-Improvement: Yes, since this helps to increase reward.
- Manipulation: Any Super-intelligent robot can manipulate or threaten its teacher to give more reward.

SUMMARY AND REFERENCES

Summary

Problem:

Specialised intelligent systems are already pervasive, but *general* ones are still out of reach.

Insight:

We have developed *unified* information-theoretic foundations for intelligent agents.

Impact:

The developed theory is a prerequisite for the development of *more flexible, adaptive, robust, reliable, and secure software/systems that our modern society needs*, and provides a gold standard and valuable guidance for researchers working on smart software.

Universal Artificial Intelligence (AIXI)

$$
\begin{array}{ccc}
\| & & \| \\
\text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\
+ & & + \\
\text{Universal Induction} & = & \text{Ockham} + \text{Bayes} + \text{Turing}
\end{array}
$$

Involved Scientific Areas

- reinforcement learning
- information theory
- theory of computation
- Bayesian statistics
- sequential decision theory
- adaptive control theory
- Solomonoff induction
- Kolmogorov complexity
- Universal search
- and many more

Introductory Literature

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hut06] M. Hutter. *Human knowledge compression prize*. open ended, http://prize.hutter1.net/.

[Hut12a] M. Hutter. *Can intelligence explode?* Journal of Consciousness Studies, 19(1-2):143–166, 2012.

[Hut12b] M. Hutter. *One decade of universal artificial intelligence*. In Theoretical Foundations of Artificial General Intelligence, pages 67–88. Atlantis Press, 2012.

[LH07] S. Legg and M. Hutter. *Universal intelligence: A definition of machine intelligence*. Minds & Machines, 17(4):391–444, 2007.

[RH11] S. Rathmanner and M. Hutter. *A philosophical treatise of universal induction*. Entropy, 13(6):1076–1136, 2011.

[VNH+11] J. Veness et al. *A Monte Carlo AIXI approximation*. Journal of Artificial Intelligence Research, 40:95–142, 2011.