One Decade of Universal Artificial Intelligence

Marcus Hutter

Australian National University

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

ANU

Abstract

The dream of creating artificial devices that reach or outperform human intelligence is an old one. A top-down approach is to first find a mathematical "gold-standard" definition of general intelligence, and then approximate the equations to make them computationally efficient. Indeed, Universal AI is such an elegant formal, objective, and non-anthropocentric theory. It results in an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. For the first time, without providing any domain knowledge, the same agent is able to self-adapt to a diverse range of interactive environments. These achievements give new hope that the grand goal of Artificial General Intelligence is not elusive. This tutorial provides an informal overview of UAI in context. It attempts to gently introduce a very theoretical, formal, and mathematical subject, and discusses philosophical and technical ingredients, traits of intelligence, underlying assumptions, relation to other work, differences to human intelligence, some social questions, and the past and future of UAI.

TABLE OF CONTENTS

1 BACKGROUND AND CONTEXT

- Organizational
- Artificial General Intelligence
- Natural and Artificial Approaches
- On Elegant Theories of
- What is (Artificial) Intelligence?
- What is Universal Artificial Intelligence?
- Relevant Research Fields
- Relation between ML & RL & (U)AI
- Tutorial Highlights

Artificial General Intelligence

What is (not) the goal of AGI research?

- Is: Build general-purpose Super-Intelligences.
- Not: Create AI software solving specific problems.
- Might ignite a technological Singularity.

What is (Artificial) Intelligence?

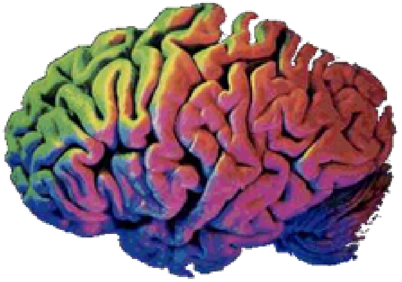
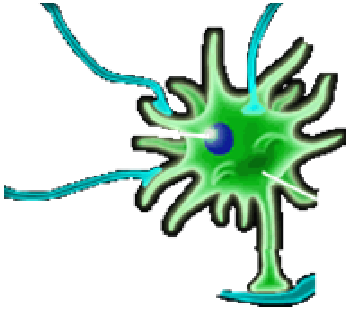
What are we really doing and aiming at?

- 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.

“Natural” Approaches

copy and improve (human) nature

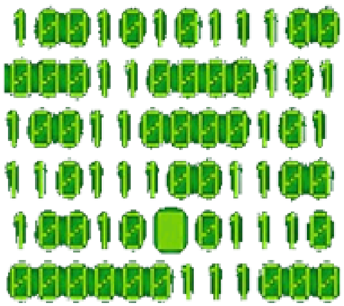
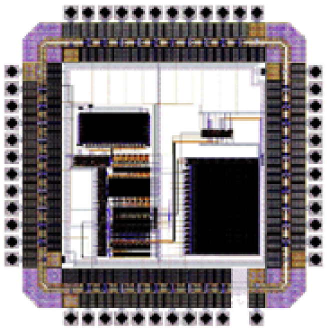
Biological Approaches to Super-Intelligence

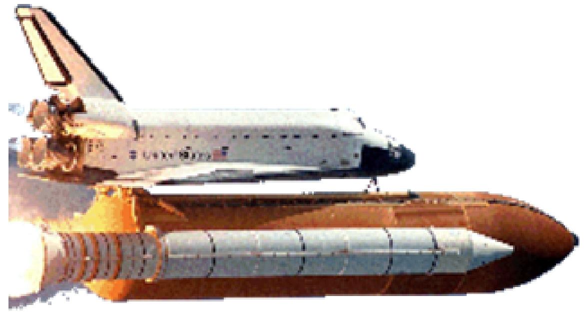
- Brain Scan & Simulation
- Genetic Enhancement
- Brain Augmentation

Not the topic of this tutorial

"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 tutorial: Foundations of "artificial" approaches to AGI

What is (Artificial) Intelligence?

Intelligence can have many faces $\Rightarrow$ formal definition difficult

- reasoning
- creativity
- association
- generalization
- pattern recognition
- problem solving
- memorization
- planning
- achieving goals
- learning
- optimization
- self-preservation
- vision
- language processing
- motor skills
- 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, ...

There is an Elegant Theory of ...

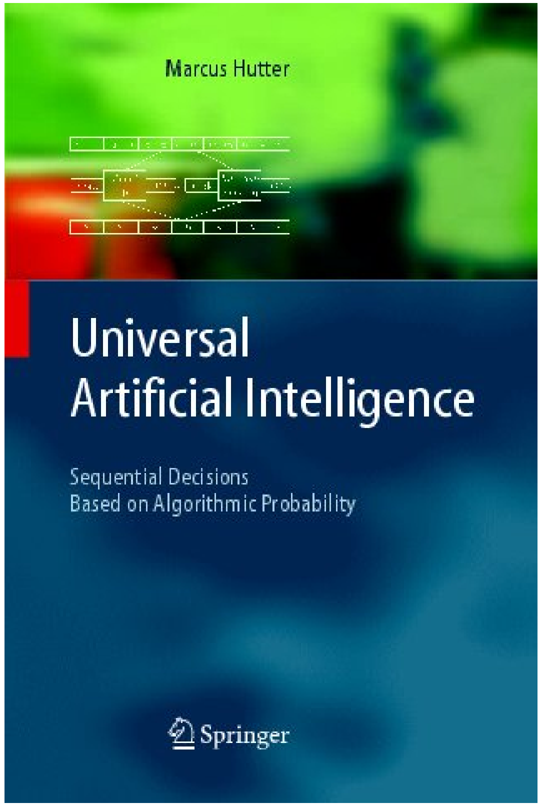
Cellular Automata	$\Rightarrow$	... Computing
Iterative maps	$\Rightarrow$	...Chaos and Order
QED	$\Rightarrow$	... Chemistry
Super-Strings	$\Rightarrow$	... the Universe
Universal AI	$\Rightarrow$	**... Super Intelligence**

What is Universal Artificial Intelligence?

- Sequential Decision Theory solves the problem of rational agents in uncertain worlds if the environmental probability distribution is *known*.

- Solomonoff's theory of Universal Induction solves the problem of sequence prediction for *unknown* prior distribution.

- Combining both ideas one arrives at



$$\begin{array}{ccc} \multicolumn{3}{c}{\text{A Unified View of Artificial Intelligence}} \\ = & & = \\ \text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\ + & & + \\ \text{Universal Induction} & = & \text{Ockham} + \text{Bayes} + \text{Turing} \end{array}$$

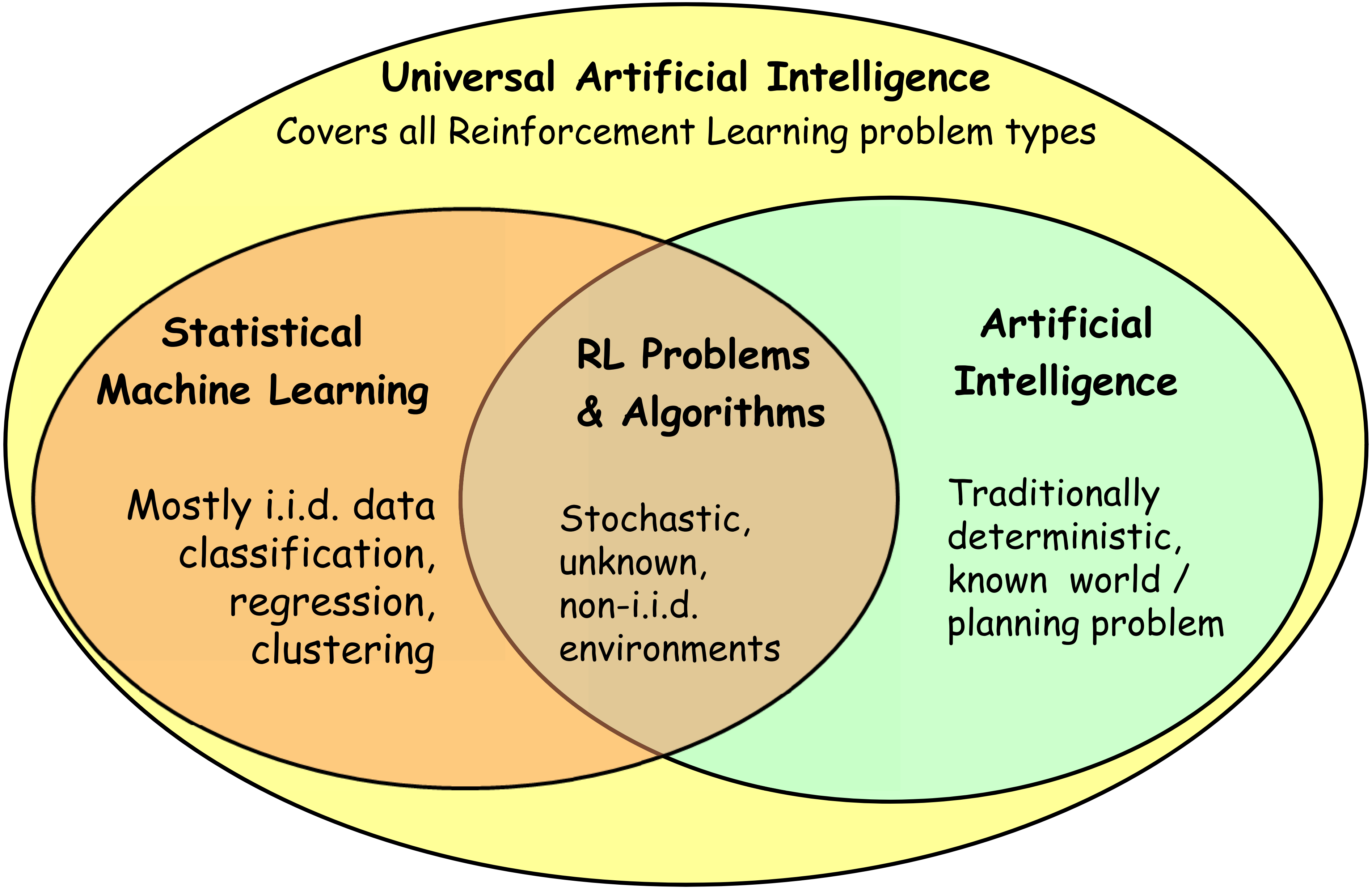
Approximation and Implementation: *Single* agent that *learns* to play TicTacToe/Pacman/Poker/... from scratch. [VNH+11]

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



Tutorial Highlights

- Formal definition of (general rational) Intelligence.
- Optimal rational agent for arbitrary problems.
- Philosophical, mathematical, and computational background.
- Some approximations, implementations, and applications.

  (learning TicTacToe, PacMan, simplified Poker from scratch)
- State-of-the-art artificial general intelligence.

2 UNIVERSAL INDUCTION & PREDICTION

- Induction→Prediction→Decision→Action
- Science $\approx$ Induction $\approx$ Occam's Razor
- Information Theory & Kolmogorov Complexity
- Bayesian Probability Theory
- Algorithmic Probability Theory
- Inductive Inference & Universal Forecasting
- Sequential Decision 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

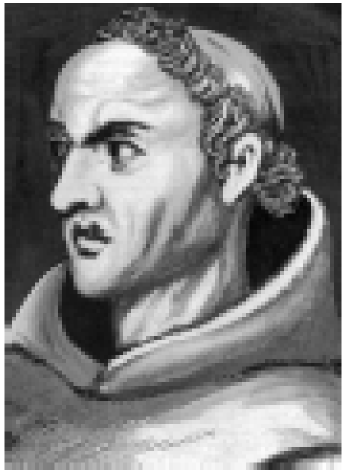
Induction: Find a model of the world economy.

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.

Science ≈ Induction ≈ Occam's Razor



- Grue Emerald Paradox:

  Hypothesis 1: All emeralds are green.

  Hypothesis 2: All emeralds found till y2020 are green, thereafter all emeralds are blue.

- Which hypothesis is more plausible? H1! Justification?

- Occam's razor: take simplest hypothesis consistent with data. is the most important principle in machine learning and science.

- Problem: How to quantify "simplicity"? Beauty? Elegance? Description Length!

  [The Grue problem goes much deeper. This is only half of the story]

Information Theory & Kolmogorov Complexity

- Quantification/interpretation of Occam's razor:
- Shortest description of object is best explanation.
- Shortest program for a string on a Turing machine T leads to best extrapolation=prediction.

$$K_T(x) = \min_p \{\ell(p) : T(p) = x\}$$

- Prediction is best for a universal Turing machine U.

Kolmogorov-complexity$(x) \;=\; K(x) \;=\; K_U(x) \qquad \leq K_T(x) + c_T$

Bayesian Probability Theory

Given (1): Models $P(D|H_i)$ for probability of observing data D, when H_i is true.

Given (2): Prior probability over hypotheses $P(H_i)$

Goal: Posterior probability $P(H_i|D)$ of H_i, after having seen data D.

Solution:

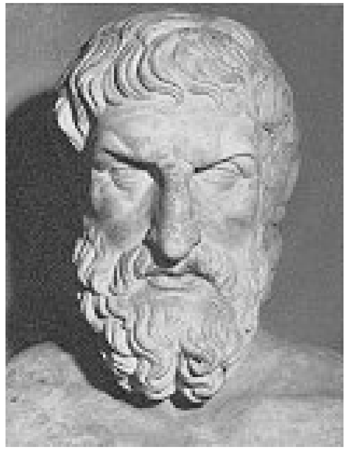
Bayes' rule:

$$\boxed{P(H_i|D) \;=\; \frac{P(D|H_i)\cdot P(H_i)}{\sum_i P(D|H_i)\cdot P(H_i)}}$$

(1) Models $P(D|H_i)$ usually easy to describe (objective probabilities)

(2) But Bayesian prob. theory does not tell us how to choose the prior $P(H_i)$ (subjective probabilities)

Algorithmic Probability Theory

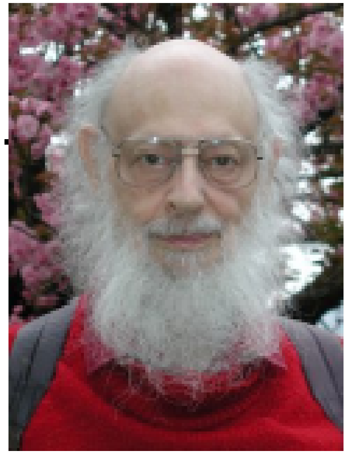


- Epicurus: If more than one theory is consistent with the observations, keep all theories.
- $\Rightarrow$ uniform prior over all H_i?
- Refinement with Occam's razor quantified in terms of Kolmogorov complexity:

$$P(H_i) \;:=\; 2^{-K_{T/U}(H_i)}$$

- Fixing T we have a complete theory for prediction.
  Problem: How to choose T.
- Choosing U we have a universal theory for prediction.
  Observation: Particular choice of U does not matter much.
  Problem: Incomputable.

Inductive Inference & Universal Forecasting



- Solomonoff combined Occam, Epicurus, Bayes, and Turing into one formal theory of sequential prediction.
- $M(x)$ = probability that a universal Turing machine outputs x when provided with fair coin flips on the input tape.
- A posteriori probability of y given x is $M(y|x) = M(xy)/M(x)$.
- Given $\dot{x}_1, .., \dot{x}_{t-1}$, the probability of x_t is $M(x_t|\dot{x}_1...\dot{x}_{t-1})$.
- Immediate "applications":
  - Weather forecasting: $x_t \in \{\text{sun,rain}\}$.
  - Stock-market prediction: $x_t \in \{\text{bear,bull}\}$.
  - Continuing number sequences in an IQ test: $x_t \in I\!N$.
- Optimal universal inductive reasoning system!

Sequential Decision Theory

Setup: For $t = 1, 2, 3, 4, ...$

Given sequence $x_1, x_2, ..., x_{t-1}$

(1) predict/make decision y_t,

(2) observe x_t,

(3) suffer loss Loss(x_t, y_t),

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

Goal: Minimize expected Loss.

Greedy minimization of expected loss is optimal if:

Important: Decision y_t does not influence env. (future observations).
Loss function is known.

Problem: Expectation w.r.t. what?

Solution: W.r.t. universal distribution M if true distr. is unknown.

Example: Weather Forecasting

Observation $x_t \in \mathcal{X} = \{$sunny, rainy$\}$

Decision $y_t \in \mathcal{Y} = \{$umbrella, sunglasses$\}$

Loss	sunny	rainy
umbrella	0.1	0.3
sunglasses	0.0	1.0

Taking umbrella/sunglasses does not influence future weather

(ignoring butterfly effect)

3 UNIVERSAL ARTIFICIAL INTELLIGENCE

- Informal Definition of (Artificial) Intelligence
- Rational Agents in Known and Unknown Environment
- Formal Definition of Intelligence
- The AIXI Model in one Line
- Some Important Problem Classes

Informal Definition of (Artificial) Intelligence

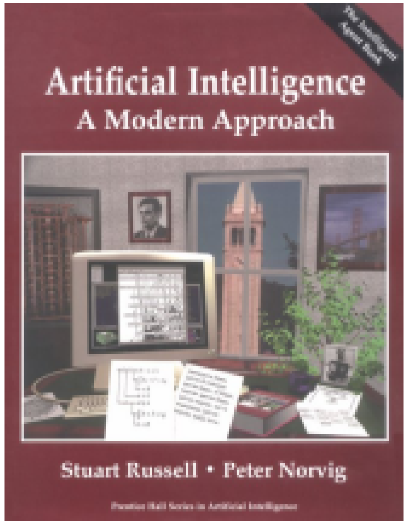
> Intelligence measures an agent's ability to achieve goals in a wide range of environments. [S. Legg and M. Hutter]

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, ...

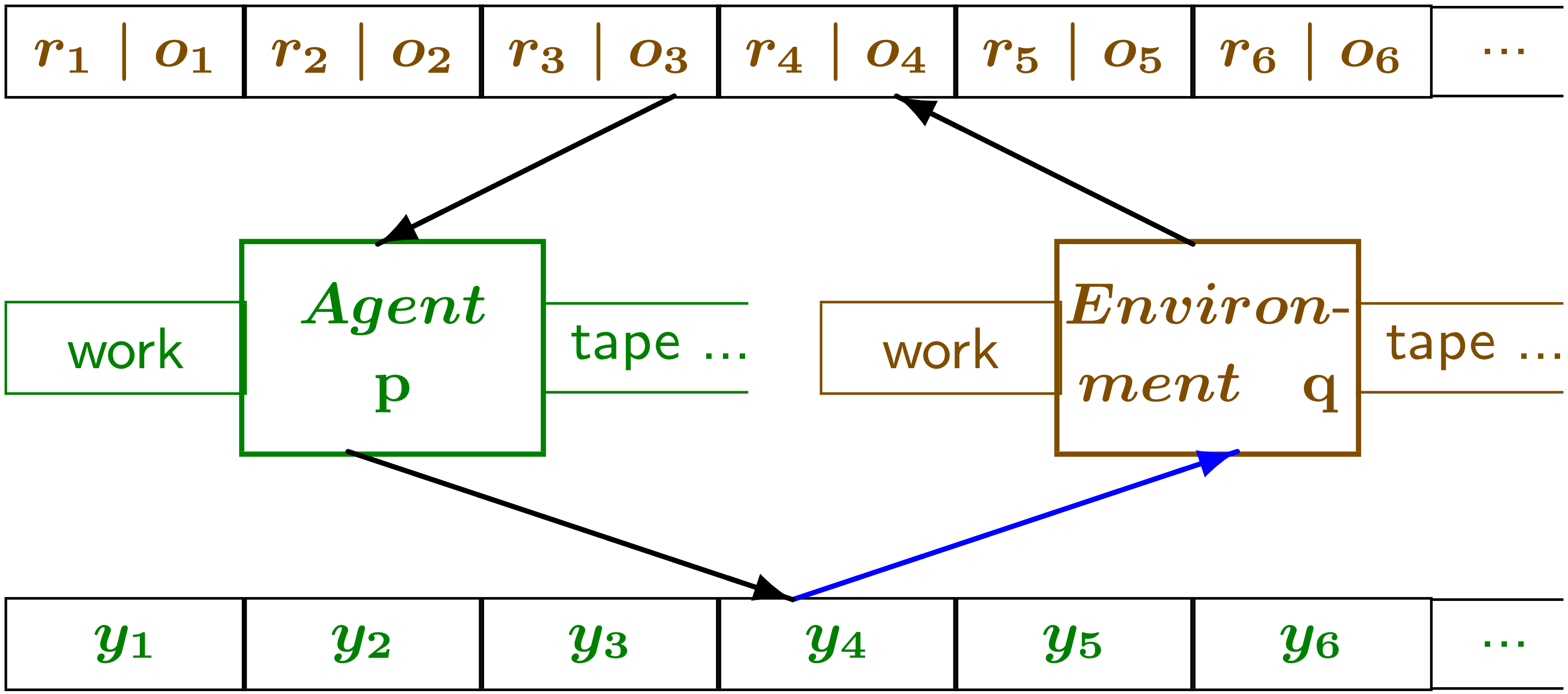
Never trust a ~~theory~~ experiment if it is not supported by an ~~experiment~~ theory



Agent Model with Reward

if actions/decisions a

influence the environment q



Rational Agents in Known Environment

- Setup: Known deterministic or probabilistic environment
- Fields: AI planning & sequential decision theory & control theory
- Greedy maximization of reward r $(=-\text{Loss})$ no longer optimal.
  Example: Chess
- Exploration versus exploitation problem.

  $\Rightarrow$ Agent has to be farsighted.
- Optimal solution: Maximize future (expected) reward sum, called value.
- Problem: Things drastically change if environment is unknown

Rational Agents in Unknown Environment

Additional problem: (probabilistic) environment unknown.

Fields: reinforcement learning and adaptive control theory

Bayesian approach: Mixture distribution ξ.

1. What performance does Bayes-optimal policy imply?
   It does not necessarily imply self-optimization
   (Heaven&Hell example).

2. Computationally very hard problem.

3. Choice of horizon? Immortal agents are lazy.

Universal Solomonoff mixture $\Rightarrow$ universal agent AIXI.

Represents a formal (math., non-comp.) solution to the AI problem?

Most (all AI?) problems are easily phrased within AIXI.

Formal Definition of Intelligence

- Agent follows policy $\pi : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^* \leadsto \mathcal{A}$
- Environment reacts with $\mu : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^*\times\mathcal{A} \leadsto \mathcal{O}\times\mathcal{R}$
- 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 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.

Some Important Problem Classes

- Sequence Prediction, e.g. weather or stock-market prediction. Strong result: number of "errors" $\propto K(\mu) =$ small.
- Strategic Games: Learn to play well (minimax) strategic zero-sum games (like chess) or even exploit limited capabilities of opponent.
- Optimization: Find (approximate) minimum of function with as few function calls as possible. Difficult exploration versus exploitation problem.
- Supervised learning: Learn functions by presenting $(z, f(z))$ pairs and ask for function values of z' by presenting $(z', ?)$ pairs. Supervised learning is much faster than reinforcement learning.

AIXI quickly learns to predict, play games, optimize, learn supervised.

4 APPROXIMATIONS & APPLICATIONS

- Universal Search and the Fastest Algorithm (FastPrg)
- Time-Bounded AIXI Model (AIXItl)
- Brute-Force Approximation of AIXI (AIξ)
- A Monte-Carlo AIXI Approximation (MC-AIXI-CTW)

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)

A General Idea: Universal Search

- Levin search: Fastest algorithm for inversion and optimization problems.
- Theoretical application:
  Assume somebody found a non-constructive proof of P=NP, then Levin-search is a polynomial time algorithm for every NP (complete) problem.
- Practical (OOPS) applications (J. Schmidhuber)
  Mazes, towers of hanoi, robotics, ...
- FastPrg: The asymptotically fastest and shortest algorithm for all well-defined problems.
- Computable Approximations of AIXI:
  AIXItl and AIξ and MC-AIXI-CTW and ΦMDP.
- Human Knowledge Compression Prize: (50'000€)

The Time-bounded AIXI Model

- Let p be any (extended chronological self-evaluating) policy
- with length $\ell(p)\leq l$ and computation time per cycle $t(p)\leq t$
- for which there exists a proof of length $\leq l_P$ that p is a valid approximation of (not overestimating) its true value $V_M^p\equiv\Upsilon(p)$.
- AIXItl selects such p with highest self-evaluation.

Optimality of AIXItl

- AIXItl depends on l,t and l_P but not on knowing p.
- It is effectively more or equally intelligent w.r.t. intelligence order relation $\succeq^c$ than any such p.
- Its size is $\ell(p^{best})=O(\log(l\cdot t\cdot l_P))$.
- Its setup-time is $t_{setup}(p^{best})=O(l_P^2\cdot 2^{l_P})$.
- Its computation time per cycle is $t_{cycle}(p^{best})=O(2^l\cdot t)$.

Brute-Force Approximation of AIXI

- Truncate expectimax tree depth to a small fixed lookahead h.
  Optimal action computable in time $|\mathcal{Y}\times\mathcal{X}|^h\times$ time to evaluate ξ.

- Consider mixture over Markov Decision Processes (MDP) only, i.e.
  $\xi(x_{1:m}|y_{1:m})=\sum_{\nu\in\mathcal{M}}w_\nu\prod_{t=1}^m\nu(x_t|x_{t-1}y_t)$. Note: ξ is *not* MDP

- Choose uniform prior over w_μ.
  Then $\xi(x_{1:m}|y_{1:m})$ can be computed in linear time.

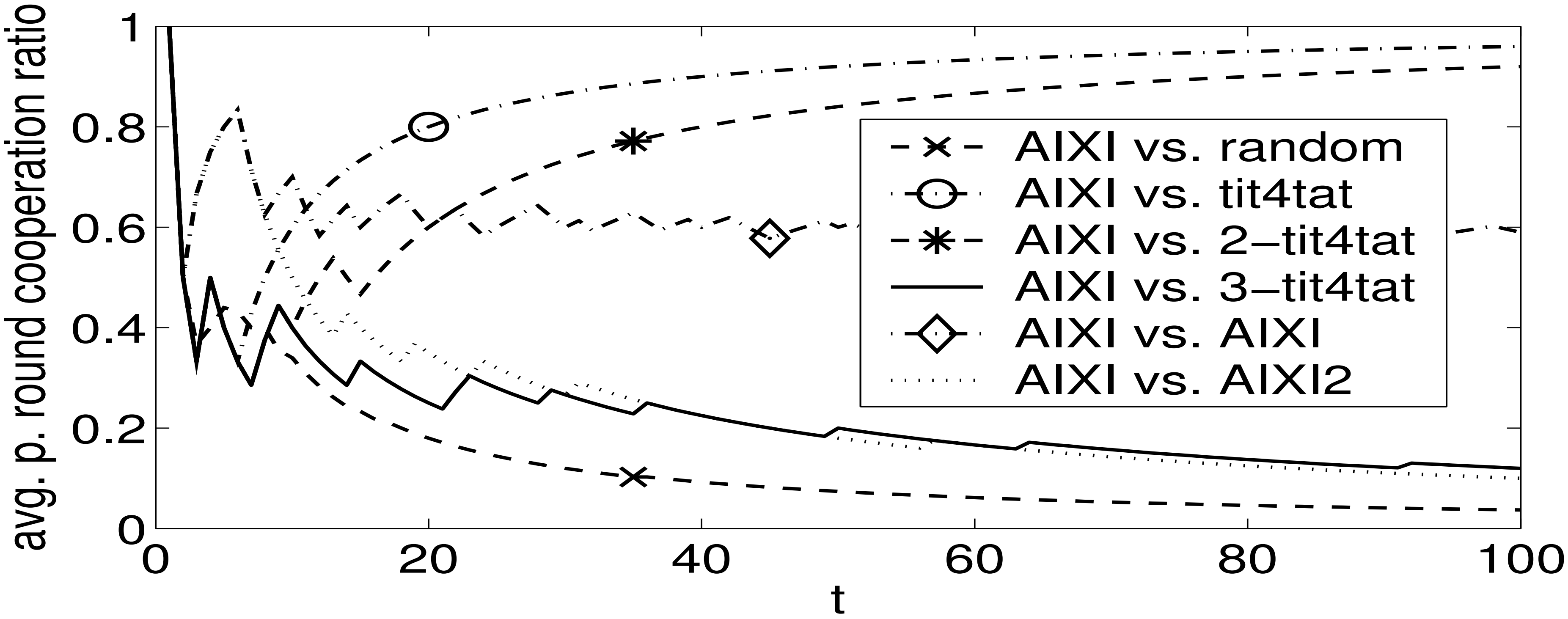
- Consider (approximately) Markov problems
  with very small action and perception space.

- Example application: 2×2 Matrix Games like Prisoner's Dilemma,
  Stag Hunt, Chicken, Battle of Sexes, and Matching Pennies [PH06].

AIXI Learns to Play 2×2 Matrix Games

- Repeated prisoners dilemma.
- Game unknown to AIXI.
  Must be learned as well
- AIXI behaves appropriately.

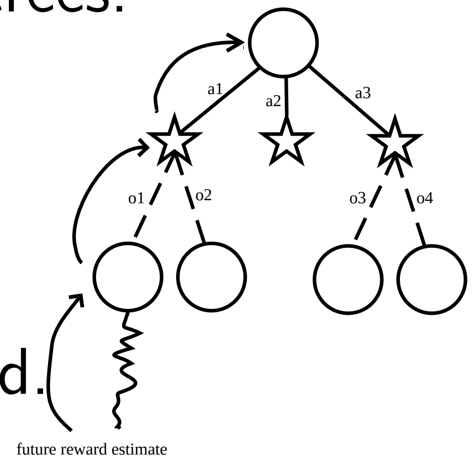
Loss matrix	cooperates	defects
cooperates	=0.3 years	=1 year
defects	free	=0.7 years



A Monte-Carlo AIXI Approximation

Consider class of Variable-Order Markov Decision Processes.

The Context Tree Weighting (CTW) algorithm can efficiently mix (exactly in essentially linear time) all prediction suffix trees.



Monte-Carlo approximation of expectimax tree:
Upper Confidence Tree (UCT) algorithm:

- Sample observations from CTW distribution.
- Select actions with highest upper confidence bound.
- Expand tree by one leaf node (per trajectory).
- Simulate from leaf node further down using (fixed) playout policy.
- Propagate back the value estimates for each node.

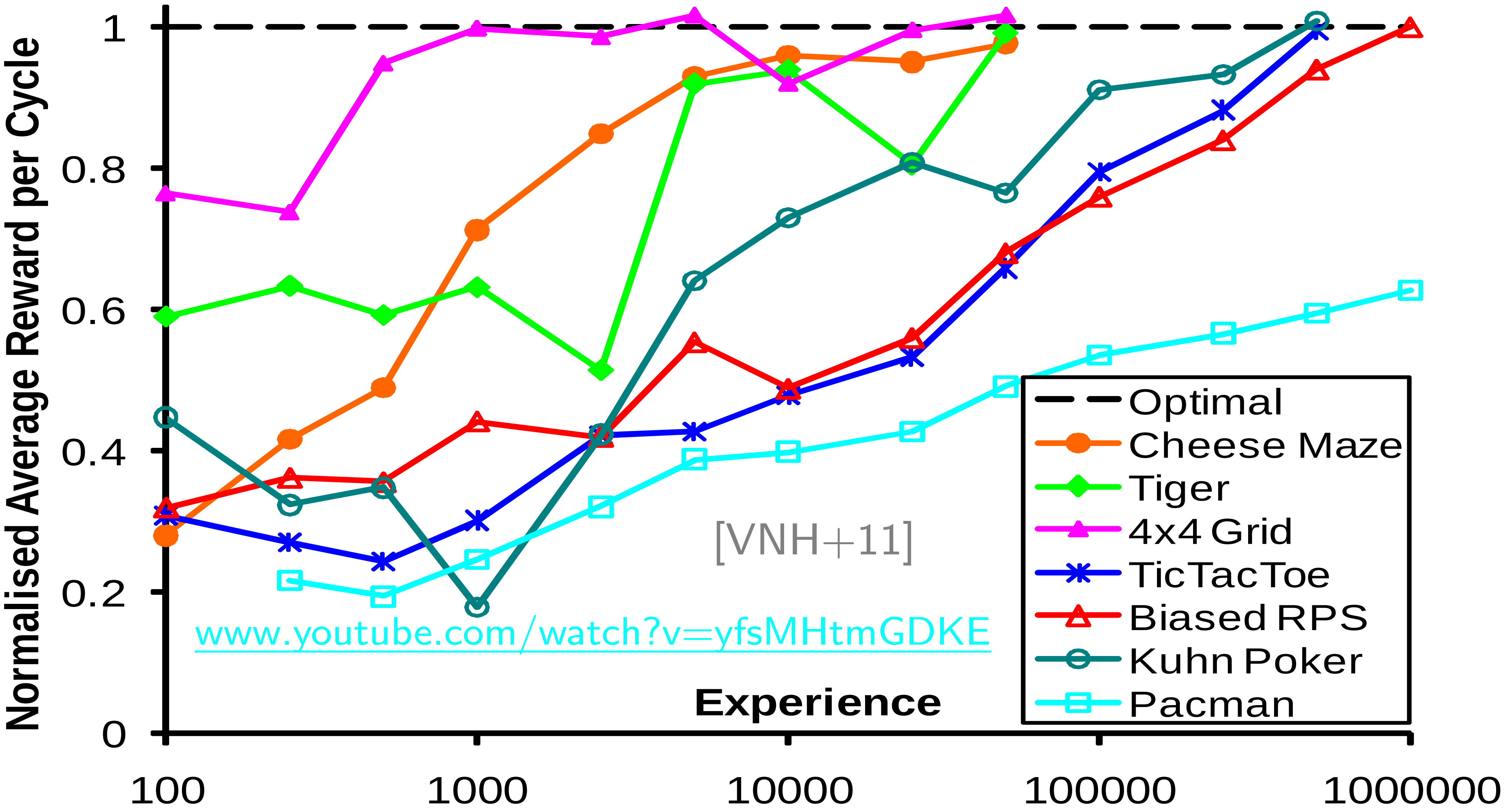
Repeat until timeout. [VNH+11]

Guaranteed to converge to exact value.

Extension: Predicate CTW not based on raw obs. but features thereof.

Monte-Carlo AIXI Applications

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



www.youtube.com/watch?v=yfsMHtmGDKE

5 UNIVERSAL AI IN PERSPECTIVE

- Aspects of AI included in AIXI
- Emergent Properties of AIXI
- Intelligent Agents in Perspective
- Properties of Learning Algorithms
- Machine Intelligence Tests & Definitions
- Common Criticisms
- General Murky & Quirky AI Questions

Connection to (AI) SubFields

- Agents: The UAIs (AIXI) is a (single) agent.
- Utility theory: goal-oriented agent.
- Probability theory: to deal with uncertain environment.
- Decision theory: agent that maximizes utility/reward.
- Planning: in expectimax tree.
- Information Theory: Core in defining and analyzing UAIs.
- Reinforcement Learning: via Bayes-mixture to deal with unknown world.
- Knowledge Representation: In compressed history.
- Reasoning: To improve compression/planning/search/... algorithms.
- Logic: For proofs in AIXItl.
- Complexity Theory: In AIXItl. We need poly-time and ultimately linear-time approx. algorithms for all building blocks.
- Heuristic Search & Optimization: Approximating Kolmogorov & Solomonoff by compressing history.
- Interfaces: Robotics, Vision, Language: In theory learnable from scratch. In practice engineered pre-&post-processing.

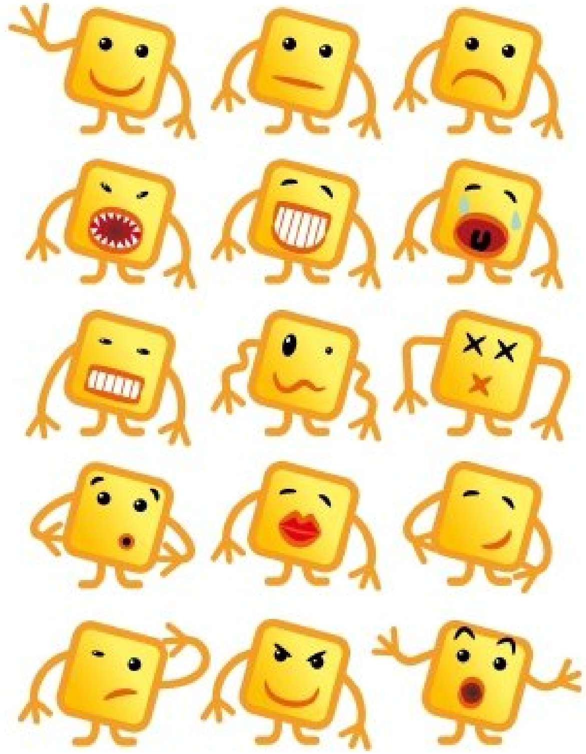
Aspects of Intelligence

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
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

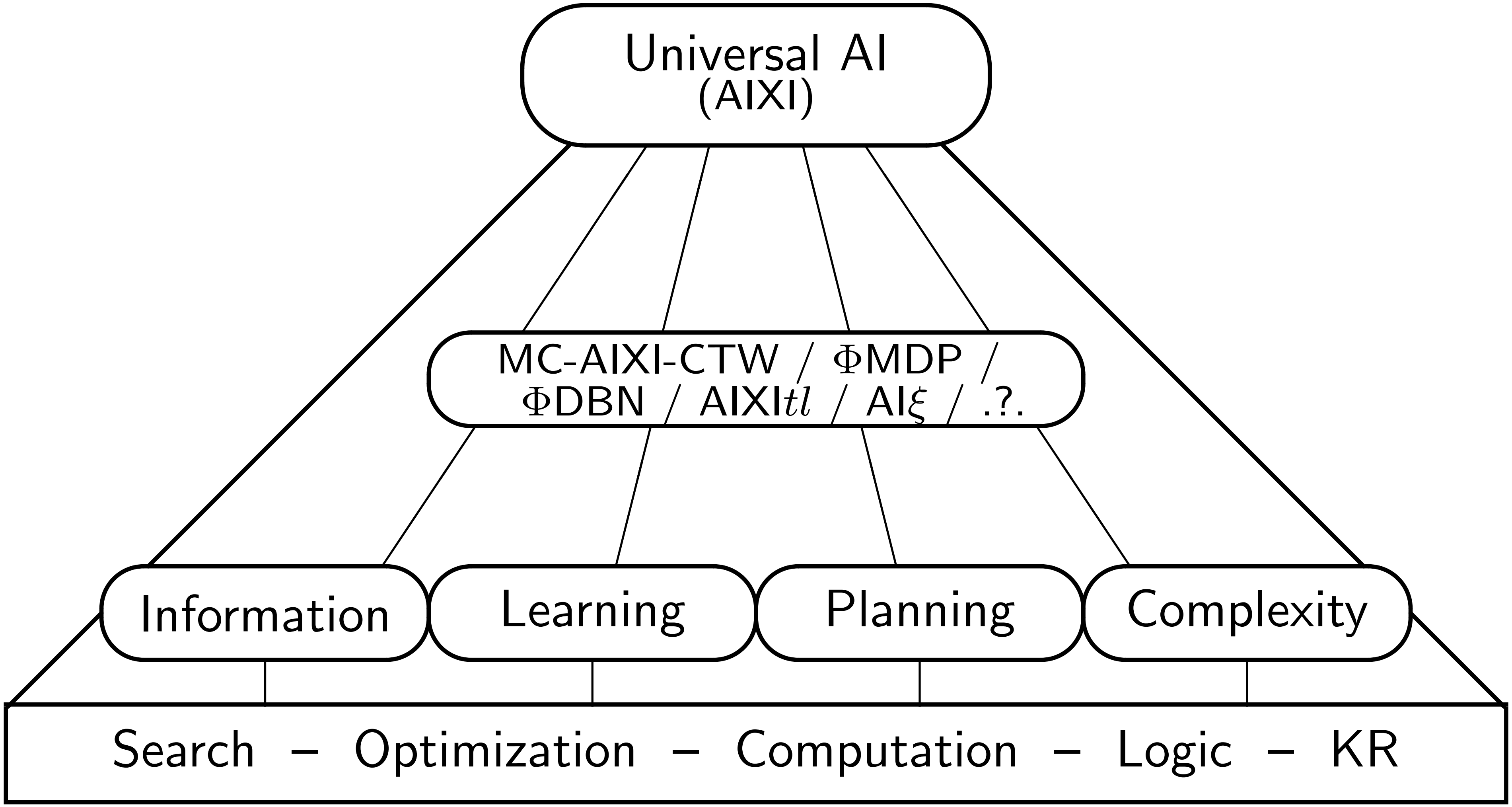
Other Aspects of the Human Mind

- Conciousness
- Self-awareness
- Sentience
- Emotions



If these qualia are relevant for rational decision making, then they should be emergent traits of AIXI too.

Intelligent Agents in Perspective



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

Properties of Learning Algorithms

Comparison of AIXI to Other Approaches

Algorithm / Properties	time efficient	data efficient	exploration	convergence	global optimum	generalization	POMDP	learning	active
Value/Policy iteration	yes/no	yes	–	YES	YES	NO	NO	NO	yes
TD w. func.approx.	no/yes	NO	NO	no/yes	NO	YES	NO	YES	YES
Direct Policy Search	no/yes	YES	NO	no/yes	NO	YES	no	YES	YES
Logic Planners	yes/no	YES	yes	YES	YES	no	no	YES	yes
RL with Split Trees	yes	YES	no	YES	NO	yes	YES	YES	YES
Pred.w. Expert Advice	yes/no	YES	–	YES	yes/no	yes	NO	YES	NO
OOPS	yes/no	no	–	yes	yes/no	YES	YES	YES	YES
Market/Economy RL	yes/no	no	NO	no	no/yes	yes	yes/no	YES	YES
SPXI	no	YES	–	YES	YES	YES	NO	YES	NO
AIXI	NO	YES	YES	yes	YES	YES	YES	YES	YES
AIXItl	no/yes	YES	YES	YES	yes	YES	YES	YES	YES
MC-AIXI-CTW	yes/no	yes	YES	YES	yes	NO	yes/no	YES	YES
Feature RL	yes/no	YES	yes	yes	yes	yes	yes	YES	YES
Human	yes	yes	yes	no/yes	NO	YES	YES	YES	YES

Machine Intelligence Tests & Definitions

★ = yes, · = no, ● = debatable, ? = unknown.

Intelligence Test	Valid	Informative	Wide Range	General	Dynamic	Unbiased	Fundamental	Formal	Objective	Fully Defined	Universal	Practical	Test vs. Def.
Turing Test	●	·	·	·	●	·	·	·	·	●	·	●	T
Total Turing Test	●	·	·	·	●	·	·	·	·	●	·	·	T
Inverted Turing Test	●	●	·	·	●	·	·	·	·	●	·	●	T
Toddler Turing Test	●	·	·	·	●	·	·	·	·	·	·	●	T
Linguistic Complexity	●	★	●	·	·	·	·	●	●	·	●	●	T
Text Compression Test	●	★	★	●	·	●	●	★	★	★	●	★	T
Turing Ratio	●	★	★	★	?	?	?	?	?	·	?	?	T/D
Psychometric AI	★	★	●	★	?	●	·	●	●	●	·	●	T/D
Smith's Test	●	★	★	●	·	?	★	★	★	·	?	●	T/D
C-Test	●	★	★	●	·	★	★	★	★	★	★	★	T/D
Universal $\Upsilon(\pi)$, AIXI	★	★	★	★	★	★	★	★	★	★	★	·	D

Common Criticisms

- AIXI is obviously wrong.
  (intelligence cannot be captured in a few simple equations)
- AIXI is obviously correct. (everybody already knows this)
- Assuming that the environment is computable is too strong.
- All standard objections to strong AI also apply to AIXI.
  (free will, lookup table, Lucas/Penrose Gödel argument)
- AIXI doesn't deal with X or cannot do X.
  (X = consciousness, creativity, imagination, emotion, love, soul, etc.)
- AIXI is not intelligent because it cannot choose its goals.
- Universal AI is impossible due to the No-Free-Lunch theorem.

See [Legg:08] for refutations of these and more criticisms.

General Murky & Quirky AI Questions

- Is current mainstream AI research relevant for AGI?
- Are sequential decision and algorithmic probability theory all we need to well-define AI?
- What is (Universal) AI theory good for?
- What are robots good for in AI?
- Is intelligence a fundamentally simple concept? (compare with fractals or physics theories)
- What can we (not) expect from super-intelligent agents?
- Is maximizing the expected reward the right criterion?
- Isn't universal learning impossible due to the NFL theorems?

6 MISCELLANEOUS CONSIDERATIONS

- Game Theory and Simultaneous Actions
- Input/Output Spaces
- Specific/Universal/Generic Prior Knowledge
- How AIXI(tl) Deals with Encrypted Information
- Origin of Rewards and Universal Goals
- Mortal Embodied (AIXI) Agent
- Some more Social Questions
- Is Intelligence Simple or Complex?

Game Theory and Simultaneous Actions

Game theory often considers simultaneous actions of both players (e.g. 2×2 matrix games) (agent and environment in our terminology).

Our approach can simulate this by withholding the environment from the current agent's output y_k, until x_k has been received by the agent.

Input/Output Spaces

- In our examples: specialized input and output spaces $\mathcal{X}$ and $\mathcal{Y}$.
- In principle: Generic interface, e.g. high-resolution camera / monitor / actuators, but then complex vision and control behavior has to be learnt too (e.g. recognizing and drawing TicTacToe boards).
- In theory: Any interface can be Turing-reduced to binary $\mathcal{X}$ and $\mathcal{Y}$ by sequentializing, or embedded into $\mathcal{X}=\mathcal{Y}=I\!N$.

Prior Knowledge — Specific Solutions

For specific practical problems we usually have extra information about the problem at hand, which could and should be used to guide the forecasting and decisions.

Ways of incorporating prior knowledge:

- Restrict Bayesian mixture ξ_U from all computable environments to those not contradicting our prior knowledge, or soft version:
- Bias weights weights w_ν towards environments that are more likely according to our prior knowledge.

Both can be difficult to realize, since one often has only an informal description of prior facts.

Prior Knowledge — Universal Solution

- Code all prior knowledge in one long binary string $d_{1:\ell}$ (e.g. a dump of Wikipedia, see H-prize) essentially in any format.
- Provide $d_{1:\ell}$ as first (sequence of) observation to AIXI/Solomonoff, i.e. prefix actual observation $x_{<n}$ with $d_{1:\ell}$.
- This also allows to predict short sequences reliably (insensitive to choice of UTM).
- This is also how humans are able to agree on predictions based on apparently little data, e.g. 1,1,1,1,1,1,?
- Humans can make non-arbitrary predictions given a short sequence $x_{<n}$ only iff $M(x_n|d_{1:\ell}x_{<n})$ leads to essentially the same prediction for all "reasonable" universal Turing machines U.

Universal=Generic Prior Knowledge

- Problem 1: Higher-level knowledge is never 100% sure.
  $\Rightarrow$ No environment (except those inconsistent with bare observations) can be ruled out categorically
  (The world may change completely tomorrow).

- Problem 2: Env. μ does not describe the total universe, but only a small fraction, from the subjective perspective of the agent.

- Problem 3: Generic properties of the universe like locality, continuity, or the existence of manipulable objects with properties and relations in a manifold may be distorted due to the subjective perspective.

- Problem 4: Known generic properties only constitute information of size O(1) and do not help much in theory (but might in practice).

- On the other hand, the scientific approach is to simply assume some properties (whether true in real life or not) and analyze the performance of the resulting models.

How AIXI(tl) Deals with Encrypted Information

- De&en-cryption are bijective functions of complexity O(1), and Kolmogorov complexity is invariant under such transformations $\Rightarrow$ AIXI is immune to encryption. Due its unlimited computational resources it can crack any encryption.

- This shows that in general it does not matter how information is presented to AIXI.

- But any time-bounded approximation like AIXItl will degrade under hard-to-invert encodings.

Origin of Rewards and Universal Goals

- Where do rewards come from if we don't (want to) provide them.
- Human interaction: reward the robot according to how well it solves the tasks we want it to do.
- Autonomous: Hard-wire reward to predefined task:
  E.g. Mars robot: reward = battery level & evidence of water/life.
- Is there something like a universal goal
- Curiosity-driven learning [Sch07]
- Knowledge seeking agents [Ors11, OLH13]

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).
- Procreate: Yes, if AIXI believes that descendants are useful (ensure retirement pension).
- Suicide: Yes (No), if AIXI can be raised to believe to go to heaven (hell).
- Self-Improvement: Yes, since this helps to increase reward.
- Manipulation: Any Super-intelligent robot can manipulate or threaten its teacher to give more reward.

Some more Social Questions

- Attitude: Are pure reward maximizers egoists, *psychopaths*, and/or killers or will they be *friendly* (*altruism* as extended *ego(t)ism*)?

- Curiosity killed the cat and maybe AIXI, or is extra reward for curiosity necessary? [Sch07, Ors11, LHS13]

- Immortality can cause laziness! [Hut05, Sec.5.7]

- Can self-preservation be learned or need (parts of) it be innate. see also [RO11]

- Socializing: How will AIXI interact with another AIXI? [Hut09, Sec.5j],[PH06]

Is Intelligence Simple or Complex?

The AIXI model shows that

in theory intelligence is a simple concept

that can be condensed into a few formulas.

But intelligence may be complicated in practice:

- One likely needs to provide special-purpose algorithms ($methods$) from the very beginning to reduce the computational burden.
- Many algorithms will be related to reduce the complexity of the input/output by appropriate pre/postprocessing (vision/language/robotics).

7 OUTLOOK AND OPEN QUESTIONS

- Outlook
- Assumptions
- Multi-Agent Setup
- Next Steps

Outlook

- Theory: Prove stronger theoretical performance guarantees for AIXI and AIξ; general ones, as well as tighter ones for special environments μ.
- Scaling AIXI down: Further investigation of the approximations AIXItl, AIξ, MC-AIXI-CTW, ΦMDP, Gödel machine. Develop other/better approximations of AIXI.
- Importance of training (sequence):
  To maximize the information content in the reward,
  one should provide a sequence of simple-to-complex tasks to solve,
  with the simpler ones helping in learning the more complex ones,
  and give positive reward to approximately the better half of the actions.

Assumptions

- Occam's razor is a central and profound assumption,
  but actually a general prerequisite of science.

- Environment is sampled from a computable probability distribution
  with a reasonable program size on a natural Turing machine.

- Objective probabilities/randomness exist
  and respect Kolmogorov's probability Axioms.
  Assumption can be dropped if world is assumed to be deterministic.

- Using Bayes mixtures as subjective probabilities did not involve any assumptions, since they were justified decision-theoretically.

Assumptions (contd.)

- Maximizing expected lifetime reward sum:
  Generalization possible but likely not needed.
  (e.g. obtain risk aversion by concave trafo of rewards)

- Finite action/perception spaces $\mathcal{Y}/\mathcal{X}$: Likely generalizable to countable spaces (ε-optimal policies), and possibly to continuous ones. but finite is sufficient in practice.

- Nonnegative rewards:
  Generalizable to bounded rewards. Should be sufficient in practice.

- Finite horizon or near-harmonic discounting.

Attention: All(?) other known approaches to AI implicitly or explicitly make (many) more assumptions.

Multi-Agent Setup – Problem

Consider AIXI in a multi-agent setup interacting with other agents, in particular consider AIXI interacting with another AIXI.

There are no known theoretical guarantees for this case, since AIXI-environment is non-computable.

AIXI may still perform well in general multi-agent setups, but we don't know.

Next Steps

- Address the many open theoretical questions in [Hut05].
- Bridge the gap between (Universal) AI theory and AI practice.
- Explore what role logical reasoning, knowledge representation, vision, language, etc. play in Universal AI.
- Determine the right discounting of future rewards.
- Develop the right nurturing environment for a learning agent.
- Consider embodied agents (e.g. internal$\leftrightarrow$external reward)
- Analyze AIXI in the multi-agent setting.

8 WHAT HAS BEEN ACHIEVED

- Recap of Universal AI and AIXI
- Involved Research Fields
- Overall and Major Achievements

Overall Achievement

- Developed the mathematical foundations of artificial intelligence.
- Developed a theory for rational agents
  acting optimally in any environment.
- This was not an easy task since intelligence has many
  (often ill-defined) facets.

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
- computational complexity theory
- Bayesian statistics
- sequential decision theory
- adaptive control theory
- Solomonoff induction
- Kolmogorov complexity
- Universal search
- and many more

Major Achievements 1

Philosophical & mathematical & computational foundations of universal induction based on

- Occam's razor principle,
- Epicurus' principle of multiple explanations,
- subjective versus objective probabilities,
- Cox's axioms for beliefs,
- Kolmogorov's axioms of probability,
- conditional probability and Bayes' rule,
- Turing machines,
- Kolmogorov complexity,
- culminating in universal Solomonoff induction.

Major Achievements 2

Miscellaneous

- Convergence and optimality results for (universal) Bayesian sequence prediction.
- Sequential decision theory in a very general form in which actions and perceptions may depend on arbitrary past events (AIμ).
- Kolmogorov complexity with approximations (MDL) and applications to clustering via the Universal Similarity Metric.
- Universal intelligence measure and order relation regarding which AIXI is the most intelligent agent.

Major Achievements 3

Universal Artificial Intelligence (AIXI)

- Unification of sequential decision theory and Solomonoff's theory of universal induction, both optimal in their own domain, to the optimal universally intelligent agent AIXI.
- Categorization of environments.
- Universal discounting and choice of the horizon
- AIXI/AIξ is self-optimizing and Pareto optimal
- AIXI can deal with a number of important problem classes, including sequence prediction, strategic games, function minimization, and supervised learning.

Major Achievements 4

Approximations & Applications

- Universal search: Levin search, FastPrg, OOPS, Gödel machine, ...
- Approximations: AIXItl, AIξ, MC-AIXI-CTW, ΦMDP.
- Applications: Prisoners Dilemma and other 2×2 matrix games, Toy Mazes, TicTacToe, Rock-Paper-Scissors, Pacman, Kuhn-Poker, ...
- Fazit: Achievements 1-4 show that artificial intelligence *can* be framed by an elegant mathematical theory. Some progress has also been made toward an elegant *computational* theory of intelligence.

Thanks! Questions? Details:

[Hut05] M. Hutter. *Universal Artificial Intelligence*. Springer, Berlin, 2005. http://www.hutter1.net/ai/uaibook.htm

[VNH+11] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011. http://dx.doi.org/10.1613/jair.3125

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

[Leg08] S. Legg. *Machine Super Intelligence*. PhD thesis, IDSIA, Lugano, Switzerland, 2008.

[RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 16(6):1076–1136, 2011. http://dx.doi.org/10.3390/e13061076

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

See `http://www.hutter1.net/official/bib.htm` for more.