COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

11 DISCUSSION

- What has been achieved?
- Universal AI in perspective
- Miscellaneous considerations
- Outlook and open questions
- Philosophical issues

Discussion: Abstract

The course concludes by critically reviewing what has been achieved and discusses some otherwise unmentioned topics of general interest. We summarize the AIXI model and compare various learning algorithms along various dimensions. We continue with an outlook on further research. Furthermore, we collect and state all explicit or implicit assumptions, problems and limitations of AIXI(tl).

The dream of creating artificial devices that reach or outperform human intelligence is an old one, so naturally many philosophical questions have been raised: weak/strong AI, Gödel arguments, the mind-body and the free will problem, consciousness, and various thought experiments. Furthermore, the Turing test, the (non)existence of objective probabilities, non-computable physics, the number of wisdom, and finally ethics, opportunities, and risks of AI are briefly discussed.

11.1 What has been Achieved: Contents

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

The AIXI Model in one Line

complete & essentially unique & limit-computable

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

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

AIXI is an elegant mathematical theory of AI

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

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

Applications: Robots, Agents, Games, Optimization, Supervised Learning, Sequence Prediction, Classification, ...

Issues in AI and how UAI solves them

Kolmogorov complexity:

- generalization
- associative learning
- transfer learning [Mah09]
- knowledge representation
- abstraction
- similarity [CV05]
- regularization, bias-variance [Wal05]

Bayes:

- exploration-exploitation
- learning

History-based:

- partial observability
- non-stationarity
- long-term memory
- large state space

Expectimax:

- planning

UAI deals with these issues in a general and optimal way

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.

11.2 Universal AI in Perspective: Contents

- 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,ΦMDP,...) are (single) agents.
- Utility theory: goal-oriented agent.
- Probability theory: to deal with uncertain environment.
- Decision theory: agent that maximizes utility/reward.
- Planning: in expectimax tree and large DBNs.
- Information Theory: Core in defining and analyzing UAIs.
- Reinforcement Learning: via Bayes-mixture and PAC-MDP to deal with unknown world.
- Knowledge Representation: In compressed history and features Φ.
- Reasoning: To improve compression/planning/search/... algorithms.
- Logic: For proofs in AIXItl and soph. features in ΦDBN.
- Complexity Theory: In AIXItl and PAC-MDP. We need poly-time and ultimately linear-time approx. algorithms for all building blocks.
- Heuristic Search & Optimization: Approximating Solomonoff by compressing history, and minimizing Cost$(\Phi, \text{Structure}|h)$
- Interfaces: Robotics, Vision, Language: In theory learnable from scratch. In practice engineered pre-&post-processing.

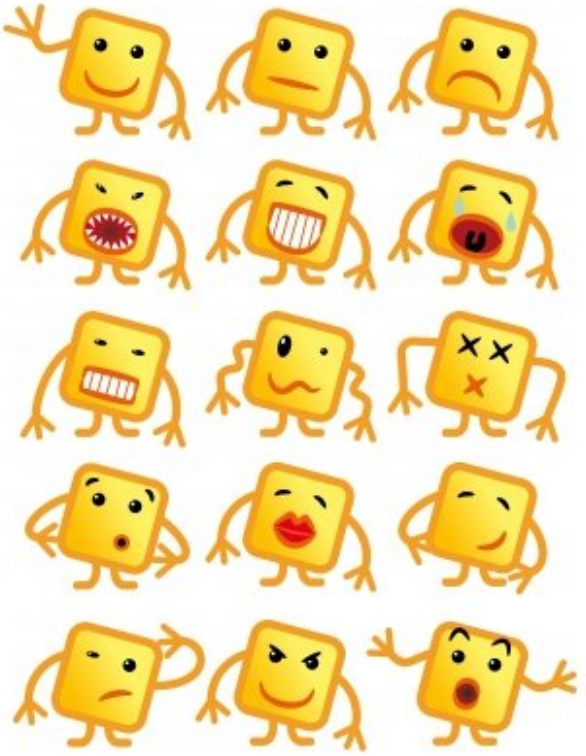
Aspects of Intelligence

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 and PAC-MDP
optimization	compression and expectimax (Cost() in ΦMDP)
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 (partition from Φ in ΦMDP)
induction	Universal Bayesian posterior (Ockham's razor)
deduction	Correctness proofs in AIXItl

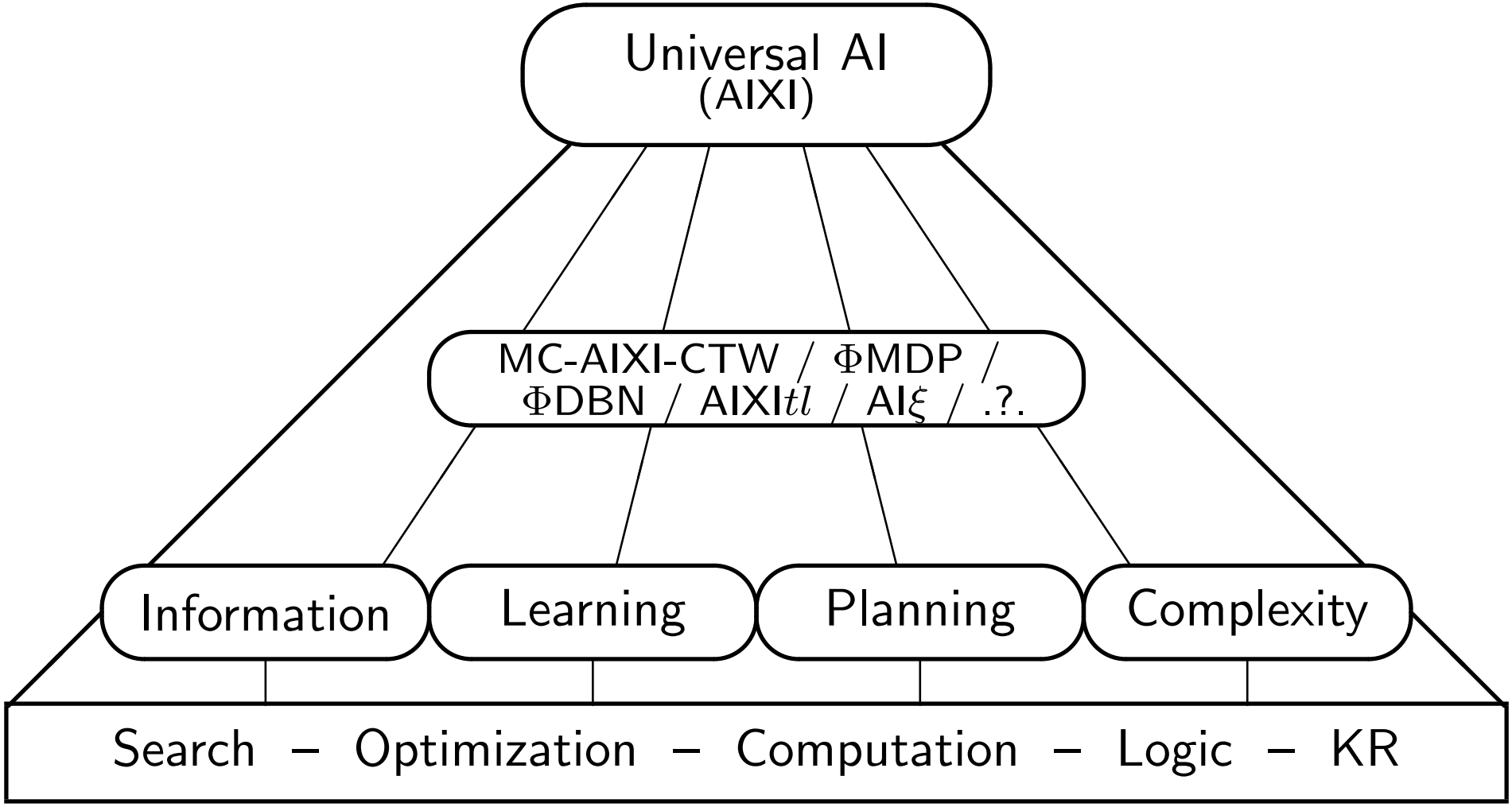
Other Aspects of the Human Mind

- 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

Issues in RL and how AIXI solves them

Kolmogorov complexity:

- generalization
- associative learning
- transfer learning [Mah09]
- knowledge representation
- abstraction
- similarity [CV05]
- regularization, bias-variance [Wal05]

Bayes:

- exploration-exploitation
- learning

History-based:

- partial observability
- non-stationarity
- long-term memory
- large state space

Expectimax:

- planning

AIXI deals with these issues in a general and optimal way

Properties of Learning Algorithms

Comparison of AIXI to Other Approaches

Algorithm \ Properties	time efficient	data efficient	explo-ration	conver-gence	global optimum	genera-lization	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 [Leg08] 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?

11.3 Miscellaneous Considerations: Contents

- 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
- Creativity – An Algorithmic View
- 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 $\mathcal{X}=\mathcal{Y}=\mathbb{N}$.

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]
- Universal (instrumental) values: survival, spreading, information, rationality, space, time, matter, energy, power, security, truth ?

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 expects negative (positive) life-time reward. [MEH16]
- 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, LH14]
- 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?

[Hut09d, Sec.5j],[PH06, LTF16]

Creativity – An Algorithmic View

- Definition: the process of producing something original&worthwhile.
- The process: combining and modifying existing thoughts or artifacts in novel ways, driven by random choice and filtering out bad results.
- Analogy: Ecosystems appear to be creatively designed, but blind evolutionary process was sufficient.
- Solving complex problems requires (apparent) creativity.
- Since AIXI is able to solve complex problems, it will appear creative.
- Analogy: Brute-force MiniMax chess programs appear to make (occasionally) creative moves.
- Creativity emerges from long-term reward maximization.
- Science $\approx$ finding patterns $\approx$ Compression
  is creative process is formal procedure
- Exploratory actions can appear creative.
- Fazit: Creativity is just exploration, filtering, and problem solving.

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).

11.4 Outlook and Open Questions: Contents

- 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.

11.5 Philosophical AI Questions: Contents

- Can machines act or be intelligent or conscious?
  (weak/strong AI, Gödel, mind-body, free will,
  brain dissection, Chinese room, lookup table)
- Turing Test & Its Limitations
- (Non)Existence of Objective Probabilities
- Non-Computable Physics & Brains
- Evolution & the Number of Wisdom
- Ethics and Risks of AI
- What If We Do Succeed?
- Countdown To Singularity
- Three Laws of Robotics

Can Weak AI Succeed?

The argument from disability:

– A machine can never do X.

\+ These claims have been disproven for an increasing # of things X.

The mathematical objection (Lucas 1961, Penrose 1989,1994):

– No formal system incl. AIs, but only humans can "see" that Gödel's unprovable sentence is true.

\+ Lucas cannot consistently assert that this sentence is true.

The argument from informality of behavior:

– Human behavior is far too complex to be captured by any simple set of rules. Dreyfus (1972,1992) "What computers (still) can't do".

\+ Computers already can generalize, can learn from experience, etc.

The Mathematical Objection to Weak AI

Applying Gödel's incompleteness theorem:

- G(F) := "This sentence cannot be proved in the formal axiomatic system F"
- We humans can easily see that G(F) must be true.
- Lucas (1961), Penrose (1989,1994):
  Since any AI is an F, no AI can prove G(F).
- Therefore there are things humans, but no AI system can do.

Counter-argument:

- L := "J.R.Lucas cannot consistently assert that this sentence is true"
- Lucas cannot assert L, but now we can conclude that it is true.
- Lucas is in the same situation as an AI.

Strong AI versus Weak AI

Argument from consciousness:

– A machine passing the Turing test would not prove that it actually really thinks or is conscious about itself.

\+ We do not know whether other humans are conscious about themselves, but it is a polite convention, which should be applied to AIs too.

Biological naturalism:

– Mental states can emerge from neural substrate only.

Functionalism:

\+ Only the functionality/behavior matters.

Strong AI: Mind-Body and Free Will

Mind-body problem:

+ Materialist: There exists only the a mortal body.
– Dualist: There also exists an immortal soul.

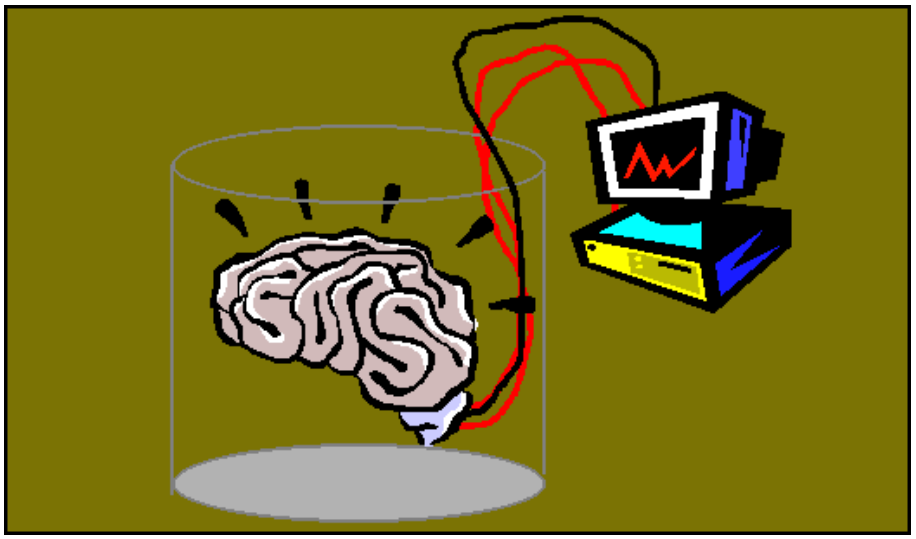
Free will paradox:

– How can a purely physical mind, governed strictly by physical laws, have free will?
+ By carefully reconstructing our naive notion of free will:
If it is impossible to predict and tell my next decision,
then I have effective free will.

Strong AI: Brain Dissection

The "brain in a vat" experiment:
(no) real experience:

+ [see movie *Matrix* for details]



The brain prosthesis experiment:

+ Replacing some neurons in the brain by functionally identical electronic prostheses would neither effect external behavior nor internal experience of the subject.
+ Successively replace one neuron after the other until the whole brain is electronic.

Strong AI: Chinese Room & Lookup Table

Strong AI: Chinese Room & Lookup Table

Assume you have a huge table or rule book containing all answers to all potential questions in the Turing test (say in Chinese which you don't understand).

- – You would pass the Turing test without understanding anything.
- + There is no big enough table.
- + The used rule book is conscious.
- + Analogy: Look, the brain just works according to physical rules without understanding anything.

Strong AI versus Weak AI: Does it Matter?

The phenomenon of consciousness is mysterious, but likely it is not too important whether a machine simulates intelligence or really *is* self aware. Maybe the whole distinction between strong and weak AI makes no sense.

Analogy:

– Natural $\leftrightarrow$ artificial: urea, wine, paintings, thinking.

– Real $\leftrightarrow$ virtual: flying an airplane versus simulator.

Is there a fundamental difference? Should we care?

Turing Test & Its Limitations

Turing Test (1950): If a human judge cannot reliably tell whether a teletype chat is with a machine or a human, the machine should be regarded as intelligent.

Standard objections:

- Tests for humanness, not for intelligence:
  - Some human behavior is unintelligent.
  - Some intelligent behavior is inhuman.
- The test is binary rather than graded.

Real problem: Unlike the Universal Intelligence Measure [LH07] and AIXI, the Turing test involves a human interrogator and, hence, cannot be formalized mathematically, therefore it does also not allow the development of a computational theory of intelligence.

(Non)Existence of Objective Probabilities

- The assumption that an event occurs with some objective probability expresses the opinion that the occurrence of an individual stochastic event has no explanation.

$\Rightarrow$ i.e. the event is inherently impossible to predict for sure.

- One central goal of science is to explain things.
- Often we do not have an explanation (yet) that is acceptable,
- but to say that "something can principally not be explained" means to stop even trying to find an explanation.

$\Rightarrow$ It seems safer, more honest, and more scientific to say that with our current technology and understanding we can only determine (subjective) outcome probabilities.

Objective=InterSubjective Probability

- If a sufficiently large community of people arrive at the same subjective probabilities from their prior knowledge, one may want to call these probabilities objective.
- Example 1: The outcome of tossing a coin is usually agreed upon to be random, but may after all be predicted by taking a close enough look.
- Eaxmple 2: Even quantum events may be only pseudo-random (Schmidhuber 2002).
- Conclusion: All probabilities are more or less subjective. Objective probabilities may actually only be inter-subjective.

Non-Computable Physics & Brains

Non-computable physics (which is not too odd)
could make Turing-computable AI impossible.

At least the world that is relevant for humans seems to be computable,
so non-computable physics can likely be ignored in practice.
(Gödel argument by Penrose&Lucas has loopholes).

Evolution & the Number of Wisdom

The enormous computational power of evolution could have developed
and coded information into our genes,
(a) which significantly guides human reasoning,
(b) cannot efficiently be obtained from scratch (Chaitin 1991).

Cheating solution: Add the information from our genes or brain
structure to any/our AI system.

Ethics and Risks of AI

– People might lose their jobs to automation.
+ So far automation (via AI technology) has created more jobs and wealth than it has eliminated.

– People might have too much (or too little) leisure time
+ AI frees us from boring routine jobs and leaves more time for pretentious and creative things.

– People might lose their sense of being unique.
+ We mastered similar degradations in the past (Galileo, Darwin, physical strength)
+ We will not feel so lonely anymore (cf. SETI)

– People might lose some of their privacy rights.

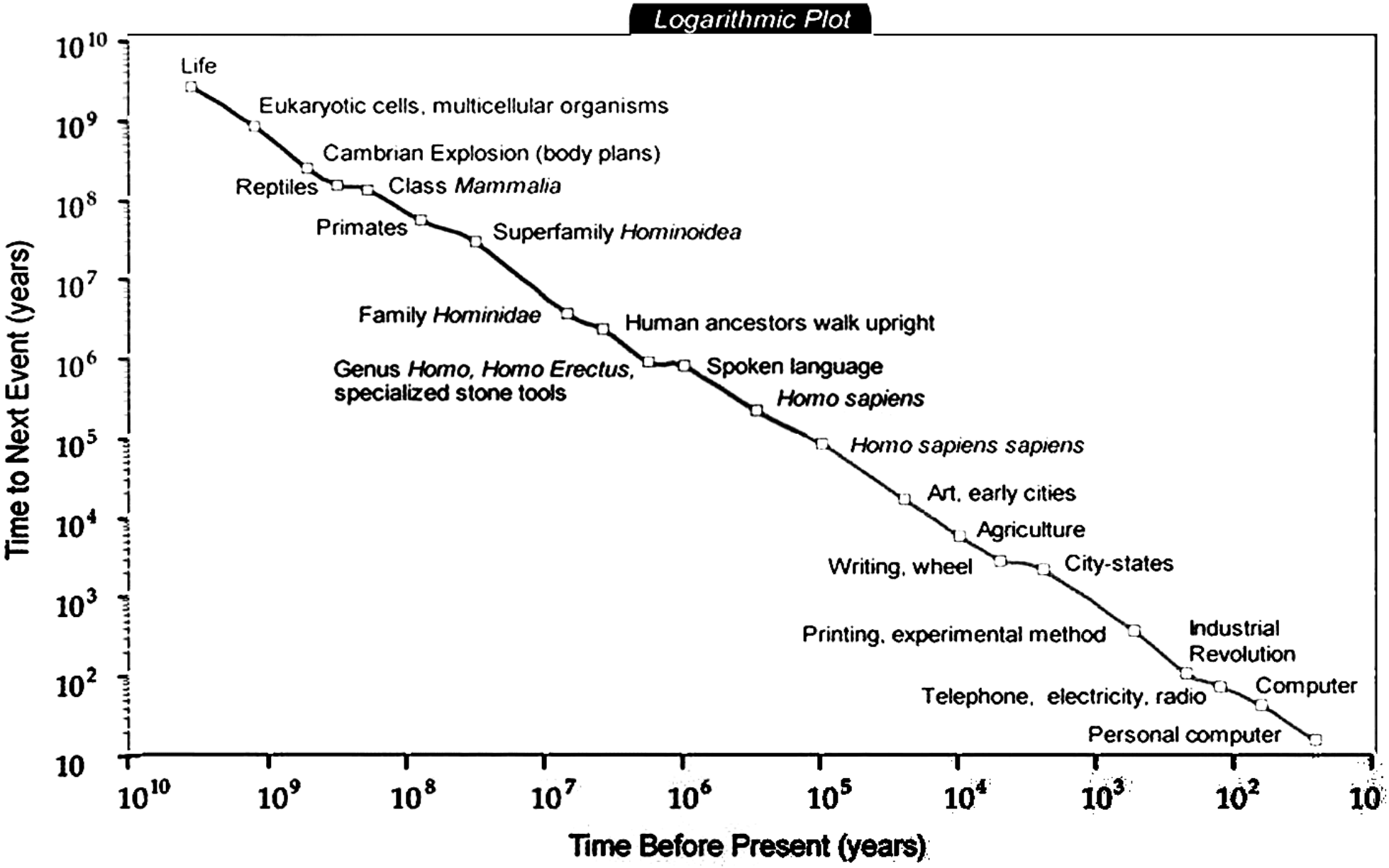
– The use of AI systems might result in a loss of accountability.
? Who is responsible if a physician follows the advice of a medical expert system, whose diagnosis turns out to be wrong?

What If We Do Succeed?

The success of AI might mean the end of the human race.

- Natural selection is replaced by artificial evolution.
  AI systems will be our mind children (Moravec 1988,2000)
- Once a machine surpasses the intelligence of a human it can design even smarter machines (I.J.Good 1965).
- This will lead to an intelligence explosion and a technological singularity at which the human era ends.
- Prediction beyond this event horizon will be impossible
  (Vernor Vinge 1993)
- Alternative 1: We keep the machines under control.
- Alternative 2: Humans merge with or extend their brain by AI.
  Transhumanism (Ray Kurzweil 2005)

Countdown To Singularity



Three Laws of Robotics

Robots (should) have rights and moral duties

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

(Isaac Asimov 1942)

Conclusions

- We have developed a parameterless model of AI based on Decision Theory and Algorithm Information Theory.
- We have reduced the AI problem to pure computational questions.
- A formal theory of something, even if not computable, is often a great step toward solving a problem and also has merits in its own.
- All other systems seem to make more assumptions about the environment, or it is far from clear that they are optimal.
- Computational questions are very important and are probably difficult. This is the point where AI could get complicated as many AI researchers believe.
- Elegant theory rich in consequences and implications.

Literature

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

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*, Chapter 8. Springer, Berlin, 2005.

[RN10] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*, Part VII. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition, 2010.

[Mor00] H. Moravec. *Robot: Mere Machine to Transcendent Mind*. Oxford University Press, USA, 2000.

[Kur05] R. Kurzweil. *The Singularity Is Near*. Viking, 2005.

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

[Bos14] N. Bostrom, *SuperIntelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Main Course Sources

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

[CV05] R. Cilibrasi and P. M. B. Vitányi. *Clustering by compression.* IEEE Trans. Information Theory, 51(4):1523–1545, 2005. http://arXiv.org/abs/cs/0312044

[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

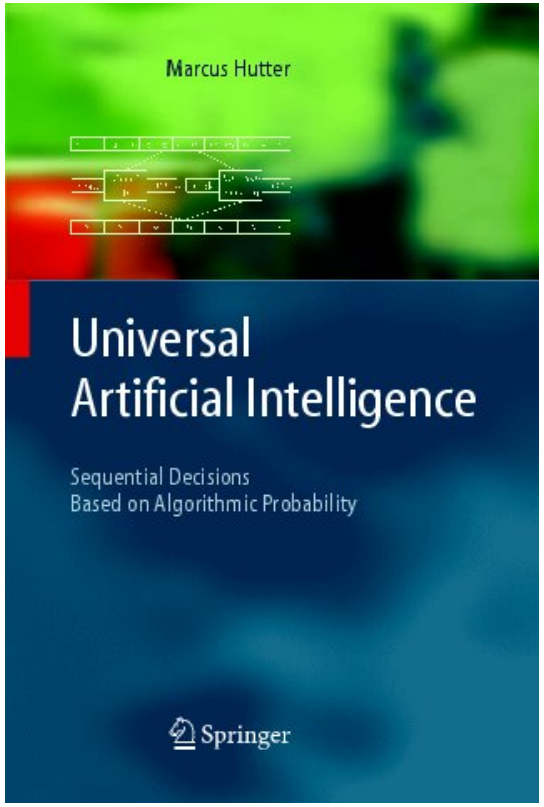
[VNH+11] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011. http://dx.doi.org/10.1613/jair.3125

[Hut12] M. Hutter. One Decade of Universal Artificial Intelligence. In *Theoretical Foundations of Artificial General Intelligence*, 4:67–88, 2012. http://arxiv.org/abs/1202.6153

Thanks! Questions? Details:

A Unified View of Artificial Intelligence

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



Open research problems:

at www.hutter1.net/ai/uaibook.htm

Compression contest:

with 50'000€ prize at prize.hutter1.net

Projects: www.hutter1.net/official/projects.htm