

---

# UNCERTAINTY & INDUCTION IN AGI

---

Marcus Hutter

Canberra, ACT, 0200, Australia

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



ANU

AGI – ProbOrNot Workshop – 31 July 2013

# Abstract

AGI systems have to learn from experience, build models of the environment from the acquired knowledge, and use these models for prediction (and action). In philosophy this is called inductive inference, in statistics it is called estimation and prediction, and in computer science it is addressed by machine learning. I will first review unsuccessful attempts and unsuitable approaches towards a general theory of uncertainty and induction, including Popper's denial of induction, frequentist statistics, much of statistical learning theory, subjective Bayesianism, Carnap's confirmation theory, the data paradigm, eliminative induction, pluralism, and deductive and other approaches. I will then argue that Solomonoff's formal, general, complete, consistent, and essentially unique theory provably solves most issues that have plagued the other approaches. Some theoretical properties, extensions to (re)active learning agents, and practical approximations are mentioned in passing, but they are not the focus of this talk. I will conclude with some general advice to philosophers and AGI researchers.

# Contents

- The Need for a General Theory
- Unsuitable/Limited Approaches
- Universal Induction and AGI
- Discussion

# Motivation

the necessity of dealing properly with uncertainty&induction in AGI

- The world is only partially observable:
  - ⇒ Need to form beliefs about the unobserved.
  - ⇒ Induction problem needs to be solved.
- Also: The world itself may be indeterministic (quantum theory)
- Uncertainty comes in degrees: low/medium/high and finer.
- A theory handling uncertainty must be strong enough to support rational decision making and action recommendation.

**Fazit:** Uncertainty and induction have to be dealt with in a quantitative and graded manner.

# Induction/Prediction Examples

**Hypothesis testing/identification:** Does treatment X cure cancer?  
Do observations of white swans confirm that all ravens are black?

**Model selection:** Are planetary orbits circles or ellipses? How many wavelets do I need to describe my picture well? Which genes can predict cancer?

**Parameter estimation:** Bias of my coin. Eccentricity of earth's orbit.

**Sequence prediction:** Predict weather/stock-quote/... tomorrow, based on past sequence. Continue IQ test sequence like 1,4,9,16,?

**Classification** can be reduced to sequence prediction:  
Predict whether email is spam.

**Question:** Is there a general & formal & complete & consistent theory for induction & prediction?

**Beyond induction:** active/reward learning, fct. optimization, game theory.

# The Need for a Unified Theory

Why do we need or should want a unified theory of induction?

- Finding new rules for every particular (new) problem is cumbersome.
- A plurality of theories is prone to disagreement or contradiction.
- Axiomatization boosted mathematics&logic&deduction and so (should) induction.
- Provides a convincing story and conceptual tools for outsiders.
- Automate induction&science (that's what machine learning does)
- By relating it to existing narrow/heuristic/practical approaches we deepen our understanding of and can improve them.
- Necessary for resolving philosophical problems.
- Unified/universal theories are often beautiful gems.
- There is no convincing argument that the goal is unattainable.

## Unsuitable/Limited Approaches

- **Popper's approach to induction** is seriously flawed:
  - falsificationism is too limited,
  - corroboration  $\equiv$  confirmation or meaningless,
  - simple  $\neq$  easy-to-refute.
- **No free lunch myth** relies on unrealistic uniform sampling. Universal sampling permits free lunch.
- **Frequentism**: definition circular, limited to i.i.d., reference class problem.
- **Statistical Learning Theory**: Predominantly considers i.i.d. data: Empirical Risk Minimization, PAC bounds, VC-dimension, Rademacher complexity, Cross-Validation.

# Unsuitable/Limited Approaches

- **Subjective Bayes:** No formal procedure/theory to get prior.
- **Objective Bayes:** Right in spirit, but limited to small classes unless community embraces information theory.
- **MDL/MML:** practical approximations of universal induction.
- **Pluralism** is globally inconsistent.
- **Deductive Logic:** not strong enough to allow for induction.
- **Non-monotonic reasoning, inductive logic, default reasoning** do not properly take uncertainty into account.
- **Carnap's confirmation theory:** Only for exchangeable data. Cannot confirm universal hypotheses.
- **Data paradigm:** data may be more important than algorithms for "simple" problems, but a "lookup-table" AGI will not work.
- **Eliminative induction:** ignores uncertainty and information theory.
- **Fuzzy logic/sets:** no sound foundations / heuristic.
- **Imprecise probability:** Is indecisive.

## Summary

The criticized approaches cannot serve as a general foundation of induction.

## Conciliation

Of course most of the criticized approaches do work in their limited domains, and are trying to push their boundaries towards more generality.

## And What Now?

Criticizing others is easy and in itself a bit pointless. The crucial question is whether there is something better out there. And indeed there is, which I will turn to now.

# Induction $\Leftrightarrow$ Deduction

Approximate correspondence between the most important concepts in induction and deduction.

	<b>Induction</b>	$\Leftrightarrow$	<b>Deduction</b>
Type of inference:	generalization/prediction	$\Leftrightarrow$	specialization/derivation
Framework:	probability axioms	$\hat{=}$	logical axioms
Assumptions:	prior	$\hat{=}$	non-logical axioms
Inference rule:	Bayes rule	$\hat{=}$	modus ponens
Results:	posterior	$\hat{=}$	theorems
Universal scheme:	Solomonoff probability	$\hat{=}$	Zermelo-Fraenkel set theory
Universal inference:	universal induction	$\hat{=}$	universal theorem prover
Limitation:	incomputable	$\hat{=}$	incomplete (Gödel)
In practice:	approximations	$\hat{=}$	semi-formal proofs
Operation:	computation	$\hat{=}$	proof

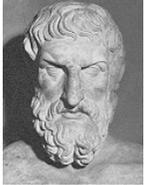
**The foundations of induction are as solid as those for deduction.**

# Foundations of Universal Induction and AGI



## 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.  $\text{Prior}(H) = 2^{-\text{Kolmogorov}(H)}$



## Bellman equations

Theory of how to optimally plan and act in known environments.

Solomonoff + Bellman = **Universal Artificial Intelligence.**

# Properties of Universal Induction

solves/avoids/meliorates most problems other systems are plagued with

- + is formal, complete, general, globally consistent,
- + has essentially optimal error bounds,
- + solves the problem of zero p(oste)rior,
- + can confirm universal hypotheses,
- + is reparametrization and regrouping invariant,
- + solves the problem of old evidence and updating,
- + allows to incorporate prior knowledge,
- + ready integration into rational decision agent,
- prediction of short sequences needs care,
- constant fudges in all results and the  $U$ -dependence,
- is incomputable with crude practical approximations.

# 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.

# Universal Artificial Intelligence

Key idea: Optimal action/plan/policy based on the simplest world model consistent with history. Formally ...

$$\text{AIXI: } a_k := \arg \max_{a_k} \sum_{O_k r_k} \dots \max_{a_m} \sum_{O_m r_m} [r_k + \dots + r_m] \sum_{p: U(p, a_1 \dots a_m) = O_1 r_1 \dots O_m r_m} 2^{-\text{length}(p)}$$

$k$ =now,  $a$ action,  $o$ bservation,  $r$ eward,  $U$ niversal TM,  $p$ rogram,  $m$ =lifespan

AIXI is an elegant, complete, essentially unique, and limit-computable mathematical theory of AI.

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

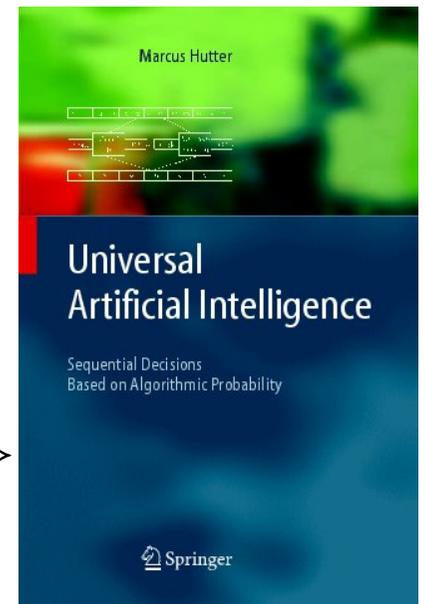
Proof: For formalizations, quantifications, proofs see  $\Rightarrow$

Problem: Computationally intractable.

Achievement: Well-defines AI. Gold standard to aim at.

Inspired practical algorithms. Cf. infeasible exact minimax.

[H'00-05]



# Approximations

- Exact computation of probabilities often intractable.
- Approximate the solution, not the problem!
- Any heuristic method needs to be gauged against Solomonoff/AIXI gold-standard.
- Cheap surrogates will fail in critical situations.
- Established/good approximations:  
MDL, MML, CTW, Universal search, Monte Carlo, ...
- AGI needs anytime algorithm powerful enough to approach Solomonoff/AIXI in the limit of infinite comp.time.

# Summary

- Conceptually and mathematically the **problem of induction is solved**.
- Computational problems and some philosophical questions remain.
- Ingredients for induction and prediction:  
Ockham, Epicurus, Turing, Bayes, Kolmogorov, Solomonoff
- For decisions and actions: Include **Bellman**.
- **Mathematical** results: consistency, bounds, optimality, many others.
- Most **Philosophical** riddles around induction solved.
- **Experimental** results via practical compressors.

Induction  $\approx$  Science  $\approx$  Machine Learning  $\approx$   
Ockham's razor  $\approx$  Compression  $\approx$  Intelligence.

# Advice to Philosophers & AGI Researchers

- Embrace U(A)I as the best conceptual solutions of the induction/AGI problem so far.
- Stand on the shoulders of giants like Bayes, Shannon, Turing, Kolmogorov, Solomonoff, Wallace, Rissanen, Bellman.
- Work out defects / what is missing, and try to improve U(A)I, or
- Work on alternatives but then benchmark your approach against state of the art U(A)I.
- Cranks who have not understood the giants and try to reinvent the wheel from scratch can safely be ignored.

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

## When it's OK to ignore $U(A)$

- if your pursued approaches already works sufficiently well
- if your problem is simple enough (e.g. i.i.d.)
- if you do not care about a principled/sound solution
- if you're happy to succeed by trial-and-error (with restrictions)

## Information Theory

- Information Theory plays an even more significant role for induction than this presentation might suggest.
- Algorithmic Information Theory is superior to Shannon Information.
- There are AIT versions that even capture Meaningful Information.

## Outlook

- Use compression size as general performance measure (like perplexity is used in speech)
- Via code-length view, many approaches become comparable, and may be regarded as approximations to UI.
- This should lead to better compression algorithms which in turn should lead to better learning algorithms.
- Address open problems in induction within the UI framework.

# Thanks! Questions? Details:

**Paper1:** *A Philosophical Treatise of Universal Induction.*

Entropy, 13:6 (2011) 1076–1136.

**Paper2:** *Universal Intelligence: A Definition of Machine Intelligence.*

Minds & Machines, 17:4 (2007) 391–444.

**Book** intends to excite a broader AI audience about abstract Algorithmic Information Theory –and– inform theorists about exciting applications to AI.

Decision Theory	=	Probability	+	Utility Theory		
+		+				
Universal Induction	=	Ockham	+	Bayes	+	Turing
=		=				
A Unified View of Artificial Intelligence						

