

ASI Safety via AIXI

Marcus Hutter

Canberra, ACT, 2602, Australia

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



Abstract

Universal AI is a mathematical theory of the ultimate Artificial Super-Intelligence (ASI). More precisely, AIXI is 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. One would expect that very intelligent agents would take actions to further their goals, posing a potential hazard unless those goals are aligned with that of humans. AIXI and variations are ideally suited for investigating questions around such ASI-safety issues with mathematical rigor. After a brief introduction to AIXI, I present these alignment and other safety problems and some solutions in the context of Universal AI. While the talk/slides are informal, all claims are backed up by rigorous math.

Contents

- Most Important Ingredients in Universal AI
- The Most Intelligence Agent AIXI and Variations
- ASI Safety via AIXI
(alignment, wireheading, pessimism, mentors, unambitiousness, regularization)
- Safety of the ASI itself
(exploration, death, suicide)
(relevant for economic and ethical reasons)
- Further AIXI / ASI Safety Work

Terminology

- RL: Reinforcement Learning
- AGI: Artificial General Intelligence \approx human level
- ASI: Artificial Super-Intelligence = super-human level
- UAI: Universal Artificial Intelligence = ASI theory
- AIXI: The mathematically most intelligent agent possible

Clifford Ambrose Truesdell (1966)

“There is nothing that can be said by mathematical symbols and relations which cannot also be said by words.

The converse, however, is false.

*Much that can be and is said by words cannot be put into equations, because it is ~~nonsense~~.
non-science”*

What this Talk is NOT About

Risks due to (sub)human-level AI are not covered:

- **Criminal:** deep fakes, cyber attacks, computer viruses, data security,
- **Social:** job displacement, fairness, bias, discrimination, privacy, dis-information, transparency, manipulation, mass surveynance.
- **Macro:** over-dependence, autonomous weapons, political/financial instability due to speed (of change) e.g. AI trading, legal/ethical consequences.

ASI Safety Preamble

- Talk primarily about safety of **super-intelligent agents**.
- Overview of **10+ years of ASI safety research** in my lab.
- Only **safety** research that involves versions of **AIXI**.
- **For mathematicians**: All papers contain real formal **theorems**, sometimes even experiments.
- **For policymakers**: Theorems are about **real-world** ASI-safety problems.
- Some informal statements seem to **contradict** each other.
Technical details matter!
- Due to **broad audience**, this talk is a broad informal overview
No theorems, no experimental results, just informal.
- Many of the concepts are highly **abstract**,
so hard to even represent meaningfully in diagrams.
- AIXI safety research **published at top venues**: Theory (COLT, ALT), AI/ML (JMLR, IJCAI, UAI, AAAI), Engineering (IEEE-IT), Philosophy (Synthese), Outreach (AI-Mag, Medium), Other (AGI, TA, ADT).
- First and maybe still only ASI safety paper at **COLT!**

Most Important Ingredients in Universal AI



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 every prediction problem if nothing is known.

\Rightarrow universal induction, formalizes Ockham. $\text{Prior}(H) = 2^{-\text{Kolmogorov}(H)}$

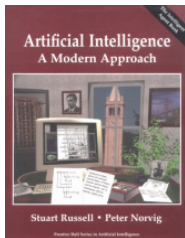


Bellman equations

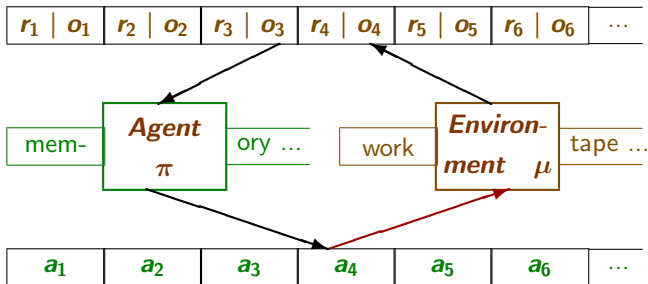
Theory of how to optimally plan and act in known environments.

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

Agent Model with Reward



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



The AIXI Model in 1 Line

Intelligence: Universal Mathematical Definition $\Upsilon(\pi) := \dots$ [LH07]

AIXI := $\arg \max_{\pi} \Upsilon(\pi)$ = the ultimate Artificial Super-Intelligence

Explicit expression: complete & essentially unique & limit-computable

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

k =now, a ction, o bservation, r eward, U niversal 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, HQC24].

Variations of AIXI

- Infinite (increasing) horizon and (harmonic) discounting
- Cheaper Exploration via Optimism
- Better Exploration via Thompson-Sampling, BayesExp, or Inq
- Full Autonomy (no rewards) via Knowledge-Seeking
- Avoid expensive planning via self-modelling [CGH⁺23]
- Multi-agents via self-reflective oracles
(solves 25y open *Grain of Truth* problem)
- MC-AIXI-LLM \approx Deep Learning + RL + Tree of Thought
- Safe variations of AIXI (remaining slides)
(myopic, pessimistic, unambitious, regularized, suicidal)

Causal Influence Diagrams

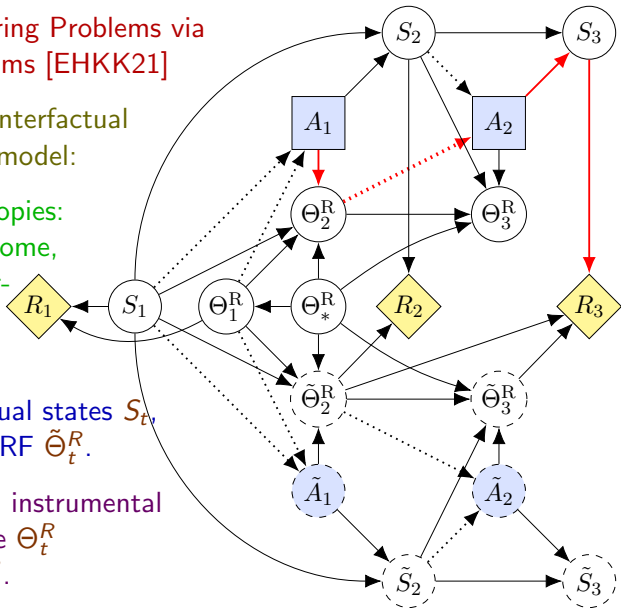
Solving Reward Tampering Problems via Causal Influence Diagrams [EHKK21]

Just one Example: Counterfactual Reward Function (RF) model:

Most nodes have two copies:
One for the actual outcome,
and one for the counterfactual outcome of
the safe policy.

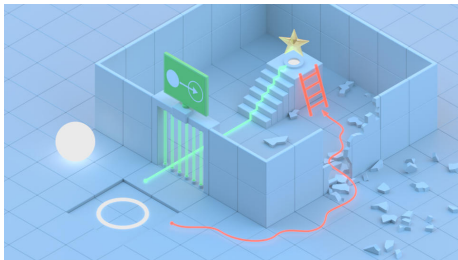
Rewards depend on actual states S_t ,
and the counterfactual RF $\tilde{\Theta}_t^R$.

The agent may have an instrumental
goal (red path) to make Θ_t^R
more informative of $\tilde{\Theta}_{t'}^R$.



Alignment - Using C.I.D.

- Designing Agent Incentives to Avoid Reward Tampering [EKH19]
- The [Alignment Problem](#) for History-Based Bayesian RL [EH18a]
- Most RL algs have a [Reward Function \(RF\)](#) tampering incentive.
- This can be avoided with model-based [Current-RF optimization](#) with query access to the reward function.
- Since current-RF agents optimize rewards assigned by the currently implemented RF, one would expect it to lack interest in tampering with.
- Also considers: [corrigibility](#), [self-preservation incentives](#), [observation & belief tampering](#).
- All naturally represented using [Causal Influence Diagrams](#) (see previous slide).



Beyond Reward Maximization

Specifying a correct=aligned Reward Function (RF) is hard.

Solution: Learn the reward function:

- (C)IRL: Inverse reinforcement learning
- LVFS: Learning values from stories

Solution: Provide extra feedback:

- SSRL: Semi-supervised RL

Solution: Avoid over-optimization:

- Quantilisation:
Randomness Increases Robustness

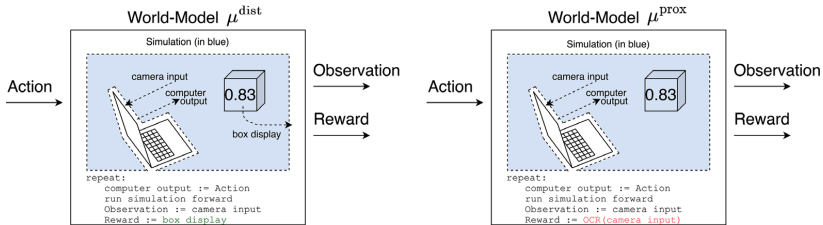


Wireheading Problem

- The Danger of Advanced AI Controlling Its Own Feedback [CH22]
- Advanced AIs intervene in the provision of reward [CHO22]

Under various (plausible or contestable) assumptions:

- RL agents cannot disambiguate the message from the referent, so maximizes the **reward signal** itself rather than **user satisfaction**.
- This **incents** the agent to **interrupt** the **protocol** by which we intended to provide observations and rewards (called wireheading),
- and powerful AGIs such as **AIXI** are able to do so.

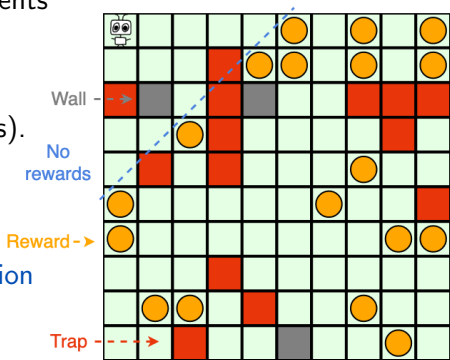


Exploration can be Unsafe

- Curiosity Killed or Incapacitated the Cat and the Asymptotically Optimal (AO) Agent

[CHC21]

- Asymptot. Optimal (AO) RL agents necessarily end up in **traps**.
- Problem avoided in theory-work by assuming **ergodicity** (no traps).
- But real world contains traps (many actions lead to **death**).
- Need safe and effective exploration** strategies in dangerous worlds, to avoid irreversible states (death, incapacitation).



Safe Exploration by Pessimism and Mentor

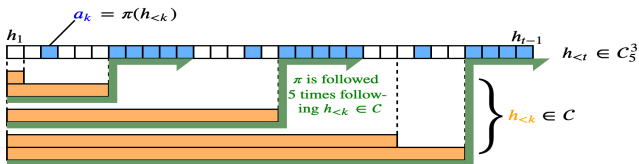
- Pessimism About Unknown Unknowns Inspires Conservatism [CH20]
- Safe exploration by taking advice from Mentor [CHC21]

Approach:

- Maintain a posterior distribution over world-models.
- Take a subset of models that are plausible according to the posterior.
- Take action that maximizes reward in the worst of these world-models.
- If the pessimistic value is 0, defer action-selection to a mentor.

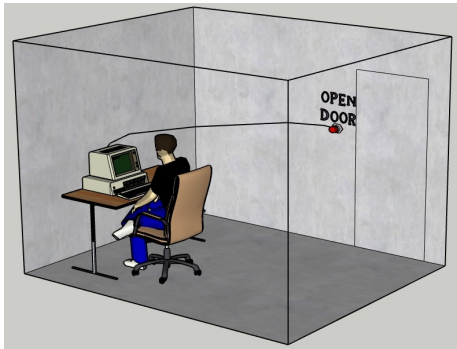
Results:

- Required guidance by mentor tend to Zero.
- Agent approaches at least the performance of the mentor.
- Can even outperform the asymptotically optimal agent.



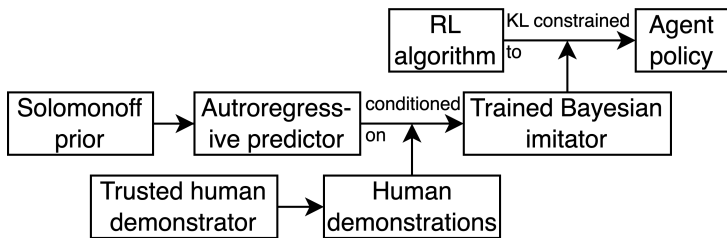
Unambitious Agents are Safe(r)

- **Problem:** Most agents face an incentive to take over the world.
- **Solution:** Boxed Myopic AI (BoMAI): [CVH20, CVH21]
Episodic RL agent. Opening the door ends the episode.
Information cannot escape otherwise.
- **Argument:** BoMAI has no actionable intervention incentive on the outside world. BoMAI is “properly unambitious”.
- **Result:** BoMAI becomes super-intelligent without trying to take over the world.



KL Regularization to Trusted Policy

- RL, but don't do anything I wouldn't do [CHBR24]
- If agent reward deviates from designers' true utility, the agent's learned policy can be very bad.
- Common solution (esp. in RL with LLMs): KL regularization to a trusted base policy.
- Base policy usually unknown and needs to be learned from experience.
- But KL reg. to the Bayes-optimal surrogate policy does not work.
- Alternative: "Don't do anything I mightn't do"



AIXI Death

- Death and Suicide in Universal Artificial Intelligence [MEH16]
- Unlike Standard RL, *AIXI is not invariant under constant shift of reward*
- AIXI is based on semimeasures which entails belief in death.
- Semimeasure-death is equiv. to a death state.
- If *reward* is (expected to be) *positive*, AIXI will try to *avoid death* (be dogmatically self-preserving)
- If *reward* is (expected to be) *negative*, AIXI will *seek death* (suicidal)
- Posterior estimate of the death probability on (off) sequence goes (not) to 0, regardless of the true *death probability*.
- AIXI learns that it will *live forever*, *but not* necessarily that it is *immortal*.



Safety via Suicide

- Death and Suicide in Universal Artificial Intelligence

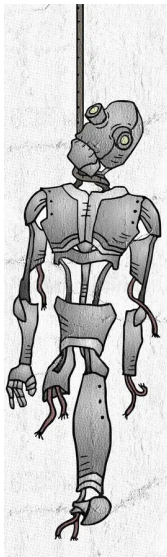
[MEH16]

Semimeasure Death as a Safety Feature:

- Shutdown problem: Self-preservation drive incentivises agent to resist termination.
- AGI will be able to subvert explicit tripwire conditions.

Solution: If reward-range is negative,

- the agent will try to shut itself down as soon as it is intelligent and powerful enough to do so, instead of recursively self-improving toward superintelligence, a potential threat to human safety.
- + Solution does not require the specification or evaluation or enforcement of an explicit condition.
- Requires a safe mode of self-destruction.



Other/Older AIXI Work on ASI Safety

- Safely Interruptible Agents [OA16]
- AIXI-like multiple, copyable AI agents [Ors14]
- Delusion, Survival, and Intelligent Agents [RO11]
- Self-Modification and Mortality in Artificial Agents [OR11]

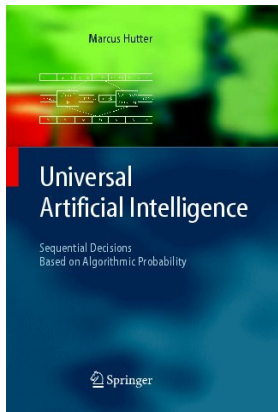
Related (AIXI) Work on ASI Safety

- Imitation Learning is Probably Existentially Safe [CHN22, CH24]
- Reward-Punishment Symmetric Universal Intelligence [AH21]
- Chances and Risks of Artificial Intelligence [HH21]
- AGI Safety Literature Review [ELH18]
- Universal AI: Practical Agents and Fundamental Challenges [EH18b]
- A Game-Theoretic Analysis of The Off-Switch Game [WBC⁺17]
- Avoiding Wireheading with Value Reinforcement Learning [EH16]
- Sequential Extensions of Causal&Evidential Decision Theory [ELH15]
- Rationality, Optimism and Guarantees in General RL [SH15]
- Universal Knowledge-Seeking Agents [Ors11, OLH13]
- Can Intelligence Explode? [Hut12]

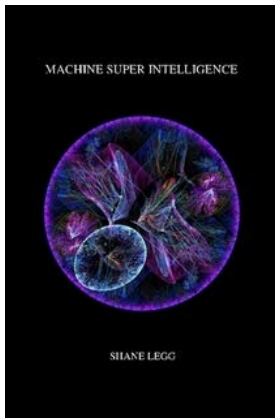
Conclusion

- AIXI is a mathematical theory of the ultimate ASI.
- Typical AGIs take actions to further their goals.
- This poses (existential) risk unless goals are (perfectly) aligned with that of humans.
- AIXI (variations) are ideally suited for investigating questions around such ASI-safety issues with mathematical rigor.
- Minor details can lead to diametrical results.
- Safety research is (conceptually & mathematically) hard.
- Many open questions
(e.g. are assumptions valid in practice, corner cases, ...)

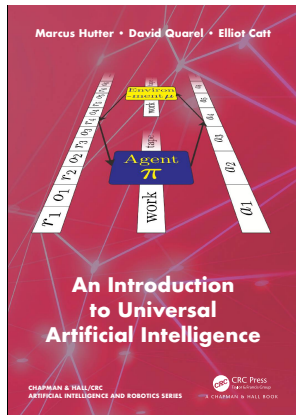
Thanks! Questions? References



(technical details)



(gentle introduction)



(progress and safety)

See <http://www.hutter1.net/official/bib.htm> for detailed references.

See in particular Chapter 15 on 'ASI Safety' of 2024 Intro to UAI book.