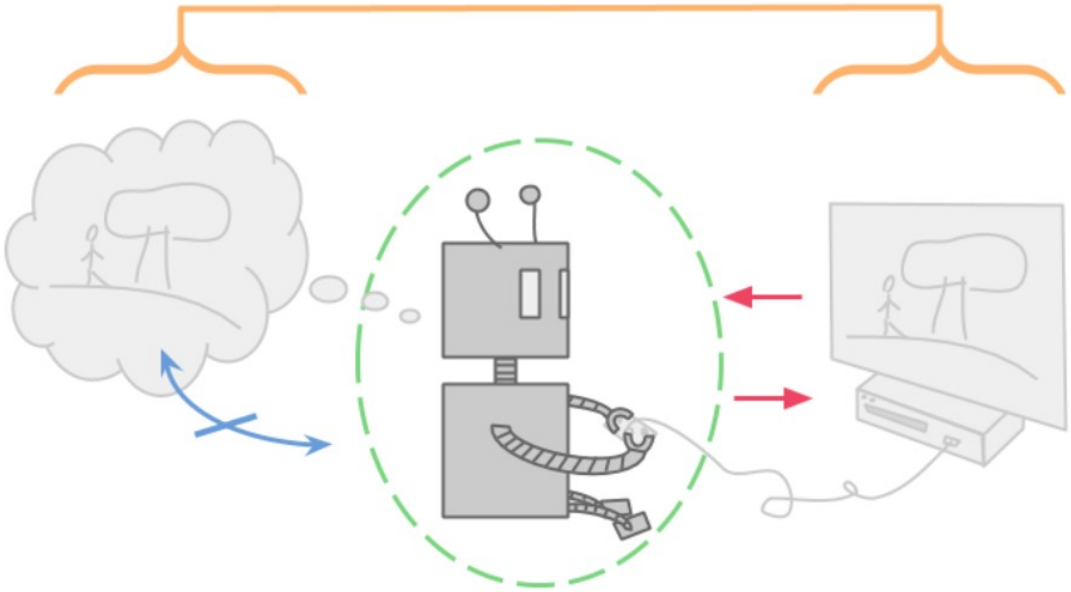
EMBEDDED AGENCY

01/15/2026

Cole Wyeth,
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

The Cybernetic Model

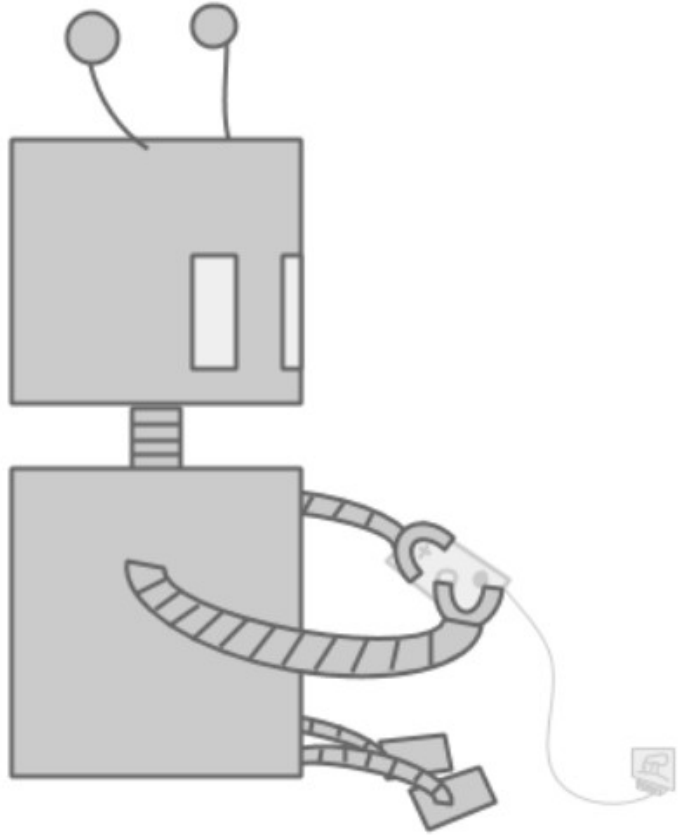


- An *agent* learns by building a *model* of its *environment*.
- The model is used for planning to achieve a goal.

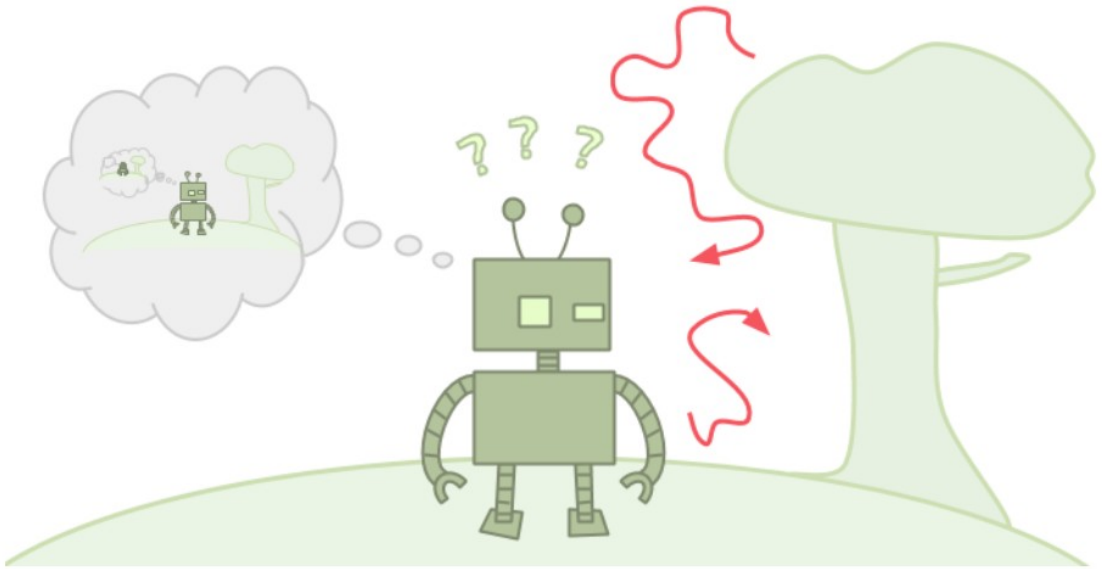
Graphics borrowed from this excellent introduction: https://www.lesswrong.com/posts/i3BTagvt3HbPMx6PN/embedded-agency-full-text-version

Large Agents

- Implicit assumptions:
  1. The agent is “larger” than the environment. The agent’s mind can fit a full copy of the environment. [Ele+25]
  2. The agent is “outside of” the environment. [DG20]
- Note 1 → 2, so model-based RL ~always assumes 2.
- Can we still learn when the agent is smaller than the environment?



The Actual Situation: Embedded Agency



- An agent is just a special part of the environment.
- It may be modified, destroyed, or copied (by itself or others) between interaction steps.
- Other equally powerful agents may exist in the environment, mutually reasoning about each other.

Ambitious Agent Design and Safety

- We are interested in agent designs that scale past human-level intelligence.*
- To put (the right) goal into a superintelligent agent, we need to express that goal in terms of whatever model the agent uses for planning.
- The agent will be smaller than and contained in its environment.

**Is it really a good idea to study designs for superintelligent agents? Arguably, superintelligent agent design is primarily useful as a model of what is coming, rather than a blueprint. This model may warn us of the dangers with negative results. For example, perhaps it is safer to build only the type of agents that fulfill limited tasks on smaller environments!*

Ambitious Agent Design and Safety

- Safety relevance:
  - Reward hacking
  - Shutdown problem
  - User identification

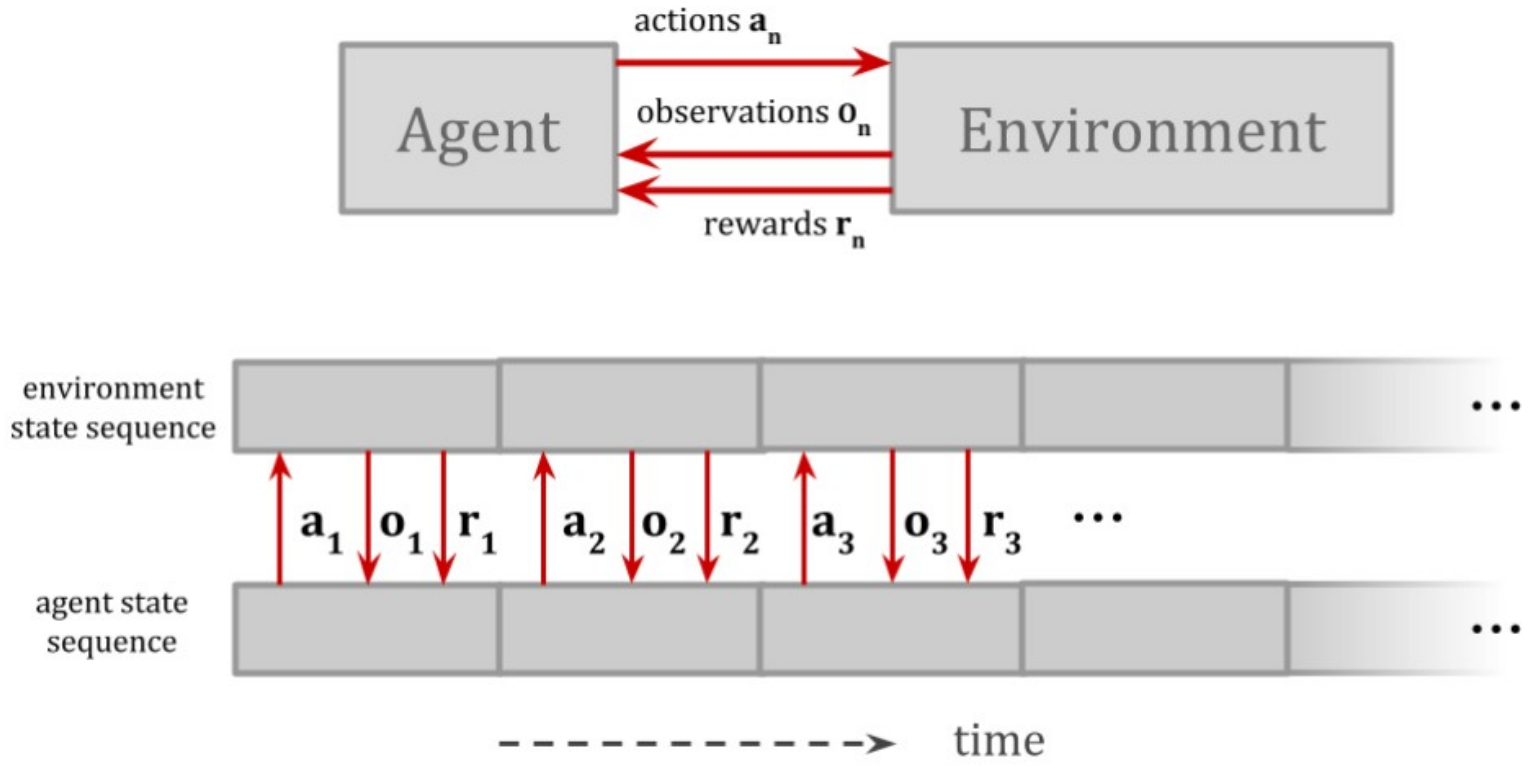
Overview

- We will start with large agents and work our way down.
  - Self-modeling Failures & AIXI
  - The Grain of Truth Problem & Reflective Oracles
  - Unrealizability & Imprecise Probabilities

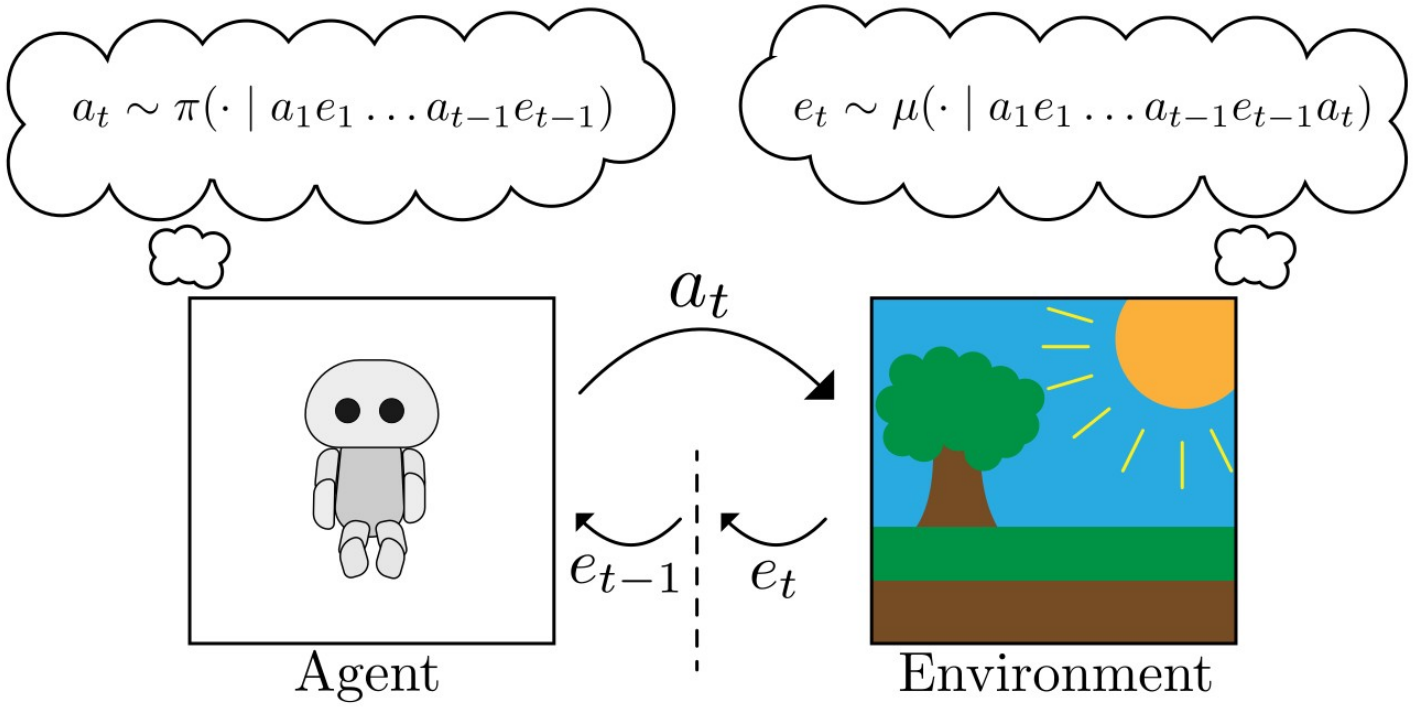
Self-modeling Failures & AIXI

History-based Reinforcement Learning

Under the cybernetic assumption, we can model a very general class of tasks using the history-based RL framework.



History-based Reinforcement Learning



$$h_{1:t} := a_1 e_1 a_2 e_2 ... a_t e_t$$

$$e_t := o_t r_t$$

$$\mu^\pi(a_t | h_{<t}) := \pi(a_t | h_{<t})$$

$$\mu^\pi(e_t | h_{<t} a_t) := \mu(e_t | h_{<t} a_t)$$

$$h \sim \mu^\pi$$

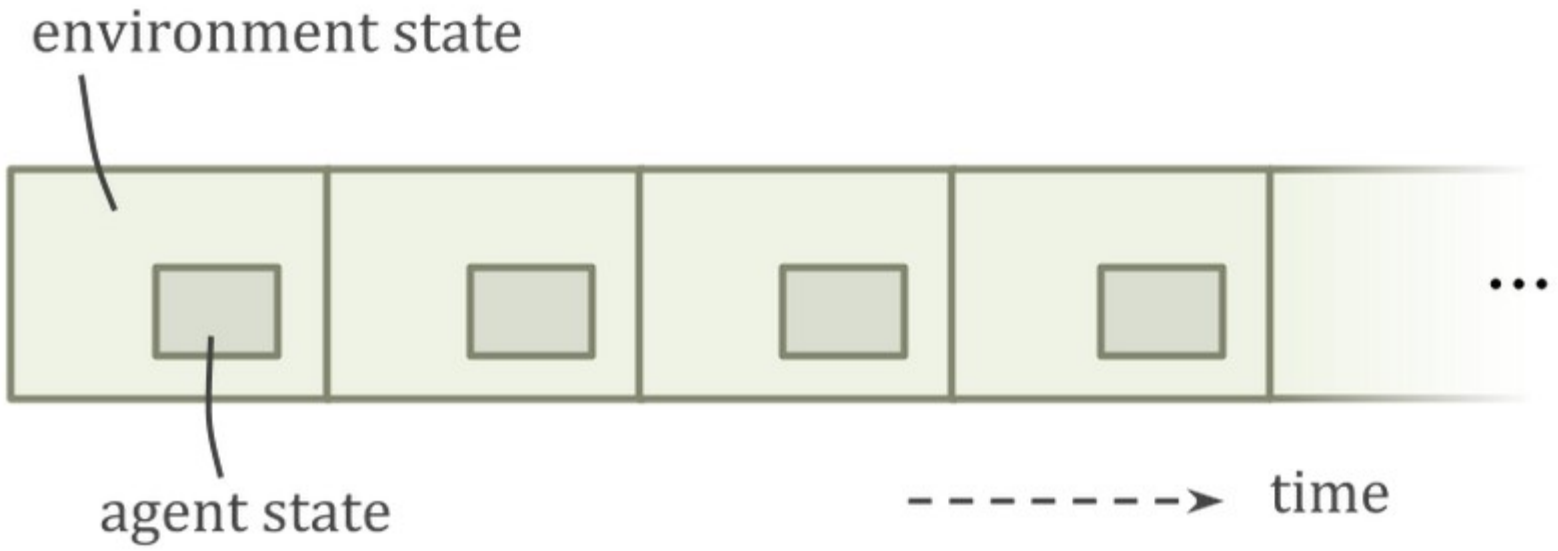
AIXI

- AIXI is a history-based RL agent designed for computable environments [Hut00]
- It plans optimally against a Bayesian mixture of possible programs.
- The simplest formulation can be written in one line (using fixed horizon and iterative value function)

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

Joint History Prediction

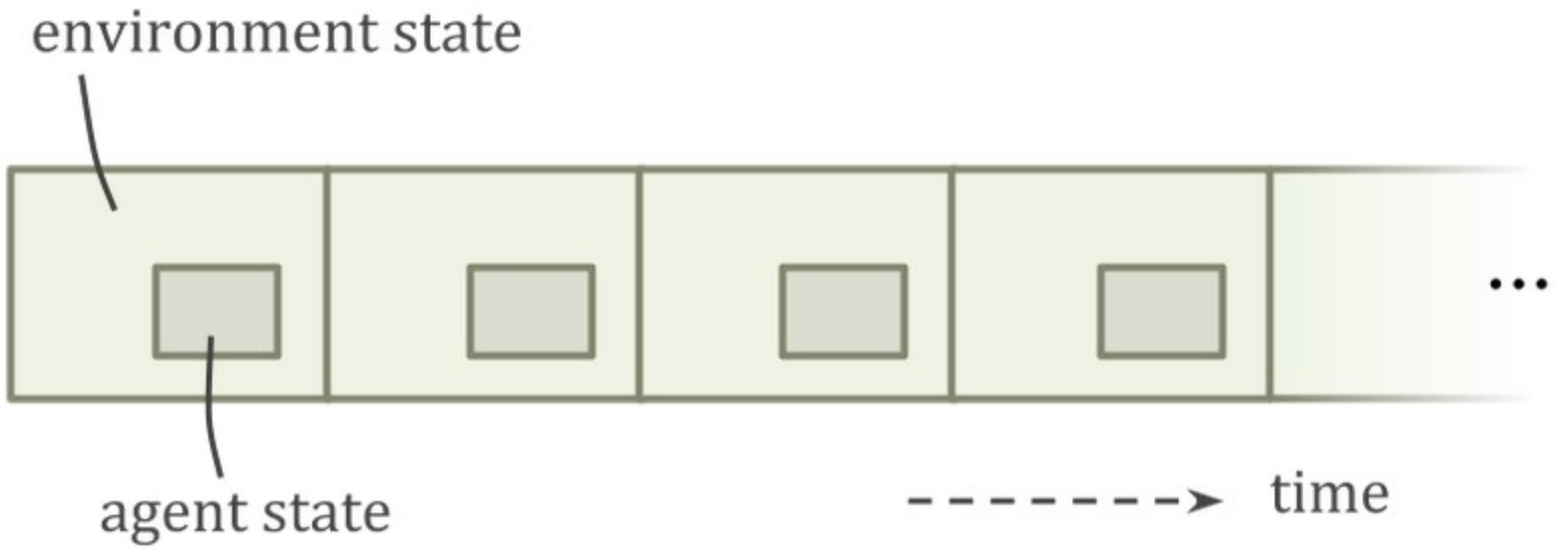
Idea: We can go beyond the cybernetic framework by jointly predicting the entire history including actions and rewards.



Joint History Prediction

Formalization: $a_1 e_1 a_2 e_2 ... = h \sim \rho$

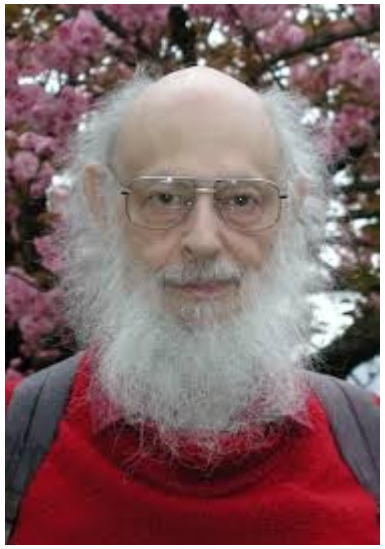
As a special case, the cybernetic assumption says $\rho = \mu^{\pi}$



Joint History Prediction

For pure prediction problems, Ray Solomonoff proposed using a Bayesian mixture over all probabilistic programs:

$$M(x) = \sum_{U(p)=x*} 2^{-l(p)}$$



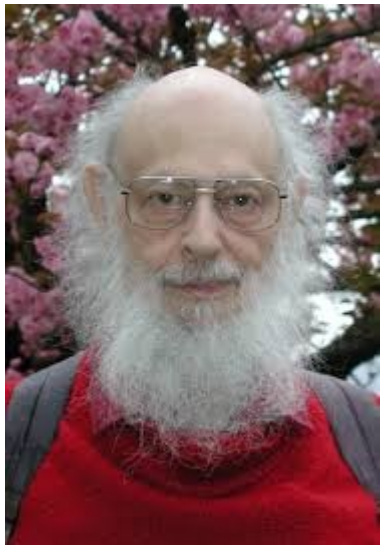
Marcus Hutter extended this framework to decision problems:

$$\xi^{\mathrm{AI}}(e_{1:t}||a_{1:t}) = \sum_{U(p,a_{1:t})=e_{1:t}*} 2^{-l(p)}$$

Joint History Prediction

For pure prediction problems, Ray Solomonoff proposed using a Bayesian mixture over all probabilistic programs:

$$M(x) = \sum_{U(p)=x*} 2^{-l(p)}$$



Why not just use this for prediction of the joint distribution?

$$\xi^U(e_{1:t}||a_{1:t}) = \prod_{i=1}^{t} M(e_i|æ_{1:i-1}a_i)$$

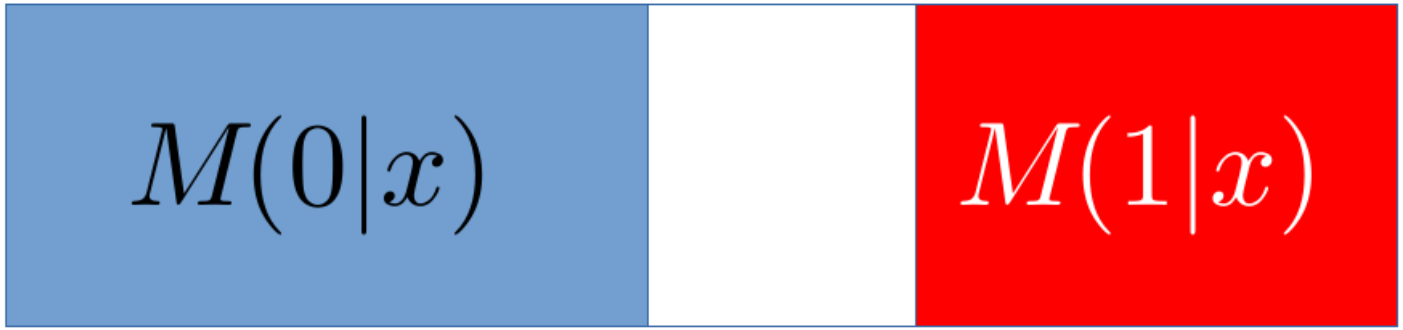
Joint History Prediction

- Joint prediction does not immediately transfer to RL
- Learning may not converge even when the cybernetic assumption holds
- We can choose uncomputable odd bits that destroy Solomonoff prediction on computable even bits:

  11 00 00 11 11 11 11 …
- The same trick means certain uncomputable actions can break Solomonoff percept prediction
- Surprisingly, normalizing M patches this problem.

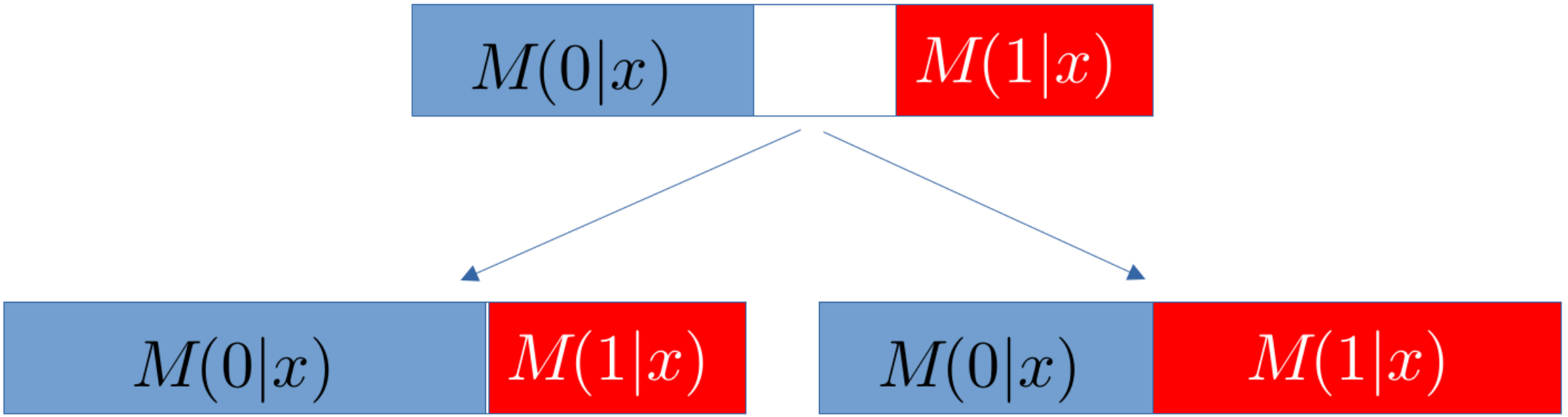
Semimeasure Loss

- Unfortunately, some programs stop producing output.
- That means M is **not** a probability distribution over infinite sequences – we call it a **semimeasure.**
- Formally, any $\nu : \mathbb{B}^* \to [0, 1]$ is a semimeasure iff $\nu(x) \geq \sum_{b \in \{0,1\}} \nu(xb)$
- We cannot tell in advance which programs will stop; this is equivalent to the **halting problem**.



Semimeasure Loss

There are multiple ways to “complete” or “normalize” a semimeasure



The gap that we reallocate this way is called the **semimeasure loss**

Prediction of Selected Bits

It turns out that the semimeasure loss can be made large infinitely often [LHG11]:

Lemma 14. *There exists an* $\omega \in \mathcal{B}^\infty$ *such that*

$$\liminf_{n\to\infty} [\mathbf{M}(0|\omega_{<n}) + \mathbf{M}(1|\omega_{<n})] = 0.$$

It follows that $\exists \omega \in \mathbb{B}^\infty$ such that $\forall n \in \mathbb{N}\ \omega_{2n} = w_{2n-1}$ but

$$M(\omega_{2n}|\omega_{<2n}) \not\to 1$$

Prediction of Selected Bits

It follows that [WH25a]:

> **Theorem 7 (Adversarial non-convergence of $\boldsymbol{\xi^U}$)** *There exists* $e=a\in\mathbb{B}^\infty$ *such that* $\xi^U(e_{1:t}||a_{1:t})=\xi^U(a_{1:t}||a_{1:t})\to 0$ *as* $t\to\infty$.

For example, there are some action/percept histories where even though the reward has always matched the (binary) value of the action, AIXI infinitely often doubts this will continue.

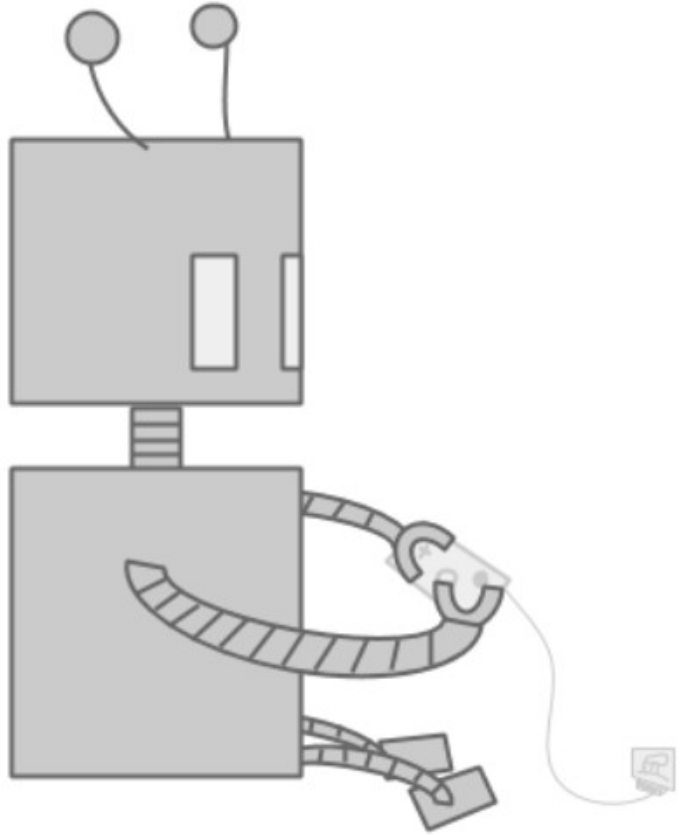
Caution: We do not know if these (action) histories are on-policy for AIXI!

Theorem 7 does not hold when M is normalized.

AIXI is Much “Larger” than its Environment

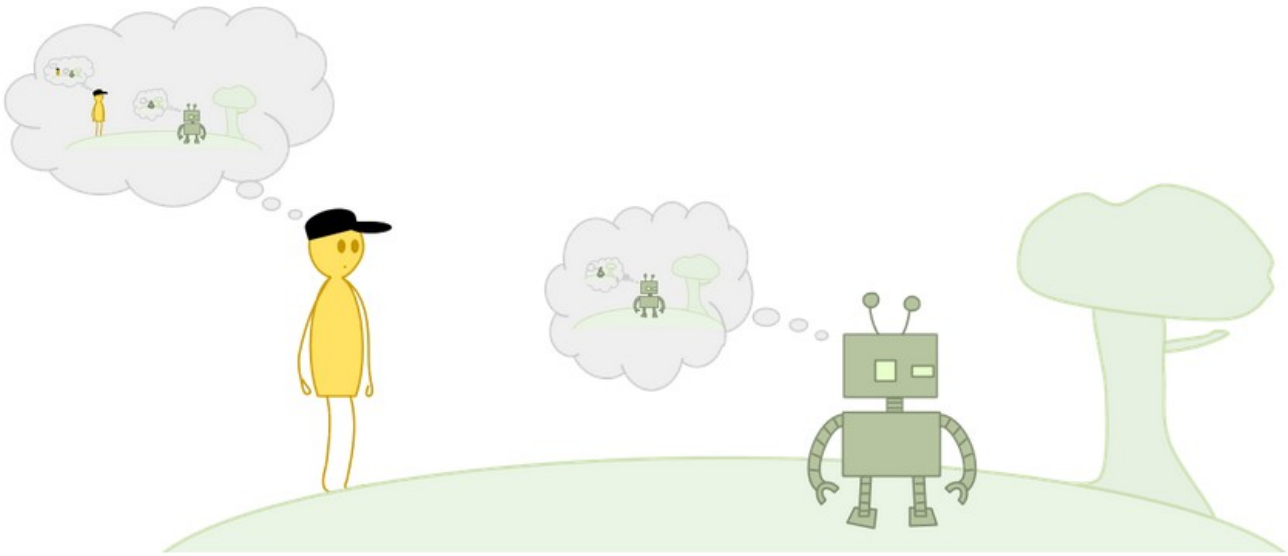
- The AIXI agent believes the environment is lower semi-computable
- The AIXI agent’s policy is not lower semi-computable
- “AIXI does not believe the universe contains AIXI agents.”

$$\dots \begin{array}{ccccccc} \Delta_n^0 & \subset & \Sigma_n^0 & \subset & \Delta_{n+1}^0 & \subset & \Sigma_{n+1}^0 \\ \cap & & \cap & & \cap & & \cap \\ \Pi_n^0 & \subset & \Delta_{n+1}^0 & \subset & \Pi_{n+1}^0 & \subset & \Delta_{n+2}^0 \end{array} \dots$$



The Grain of Truth Problem & Reflective Oracles

Modeling Multi-player Games



- Agents which are modeling each other must be of the “same size”
- Therefore, AIXI cannot learn that the environment contains other AIXI instances
- Note that multi-player games do not necessarily involve other embedded agency problems (“being inside the environment”)

Modeling Multi-player Games

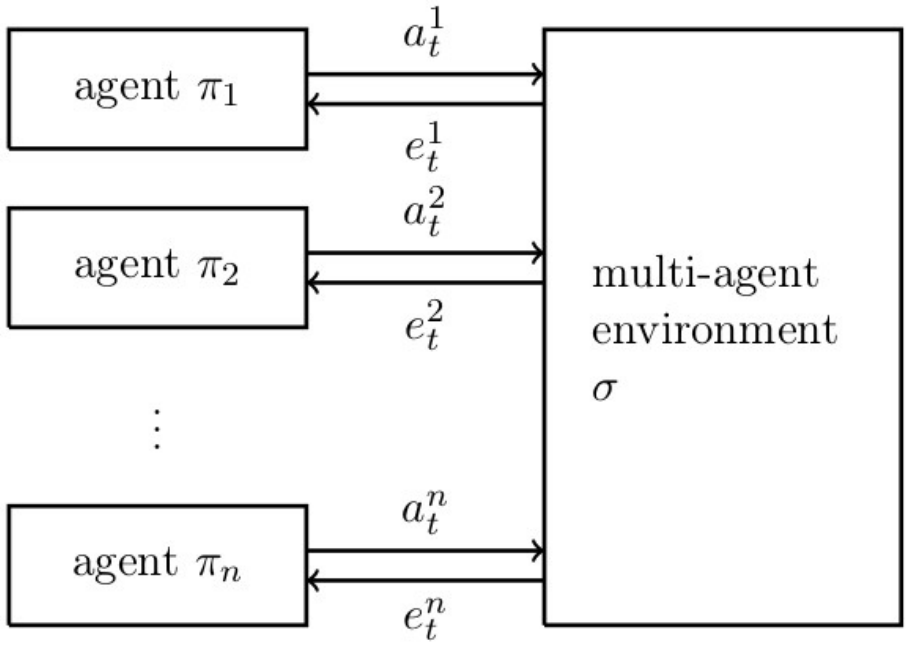


Figure 7.2: Agents $\pi_1, \ldots, \pi_n$ interacting in a multi-agent environment.

$$h_{1:t} := a_1 e_1 a_2 e_2 ... a_t e_t$$

$$h \sim \sigma^\pi$$

$$h^i_{1:t} := a^i_1 e^i_1 a^i_2 e^i_2 ... a^i_t e^i_t$$

$$h^i \sim \sigma^\pi_i$$

Modeling Multi-player Games

$$\sigma^{\pi}(\epsilon) := 1$$

$$\sigma^{\pi}(h_{<t}a_t) := \sigma^{\pi}(h_{<t}) \prod_{i=1}^{n} \pi_i(a_t^i | h_{<t}^i)$$

$$\sigma^{\pi}(h_{<t}a_t e_t) := \sigma^{\pi}(h_{<t}a_t)\sigma(e_t | h_{<t}a_t)$$

$$\sigma_i^{\pi}(h_{<t}) := \sum_{h_{<t}^j :\ j \neq i} \sigma^{\pi}(h_{<t})$$

The Grain of Truth Problem

In a multiplayer game, we would like for Bayesian players to have correctly-specified models of their subjective environments. One way of formalizing this is that they should not be “infinitely surprised” by any even that happens with positive probability, according to the truth history distribution.

Given

- A class of games $\mathcal{G}$
- A class of strategies $\mathcal{P}$
- Subjective beliefs ξ_i for player i

The grain of truth property requires that

- Each optimal policy $\pi^*_{\xi_i} \in \mathcal{P}$
- Each belief distribution $\xi_i \gg \sigma^\pi (\forall \sigma \in \mathcal{G}, \forall \pi \in \mathcal{P}^n)$

Where $\gg$ denotes absolute continuity.

Solution with reflective oracles

- Note that this appears to fail for AIXI players!
- Naively, we might construct games/policies by

  $$\lambda_T(x_t|x_{<t}) = P[T(x_{<t}) = x_t]$$

  But policies hang when they call each other
- Instead, we provide each machine T with a (reflective) oracle O that gives it arbitrarily precise access to the RHS for all other machines T' with O.

  $$\lambda_T^O(x_t|x_{<t}) = P[T^O(x_{<t}) = x_t]$$
- These augmented machines can be used to construct reflective versions of AIXI which satisfy the grain of truth property

Solution with reflective oracles

- Game-theoretic results
  - In the “dualistic” multiplayer game setup, we can show convergence to Nash equilibrium in stage games and generally convergence to optimality for Thompson sampling [LTF16,Wye+25]
  - If we allow correlations between policies (using joint prediction) then we can also prove “acausal” coordination in the prisoner’s dilemma [Mue+25]
  - The semimeasure loss is eliminated and prediction of selected bits succeeds

Solution with reflective oracles

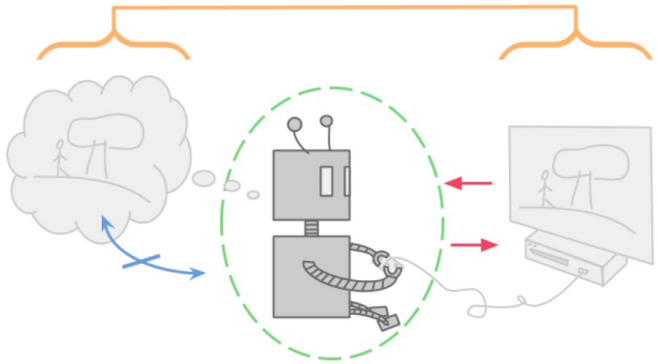
- An agent can also exploit its self-model to avoid planning
- In principle, this allows it to reason about self-modifications
- However, it is difficult to prove any optimality (or any other interesting) properties about this model

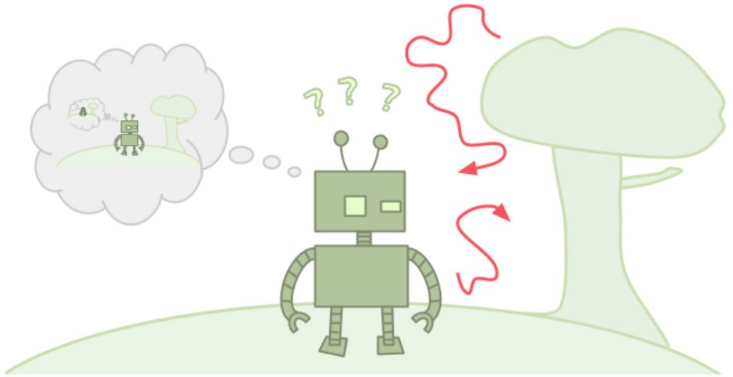
$$a_1, e_1, a_2, e_2, \ldots \sim \xi^O$$

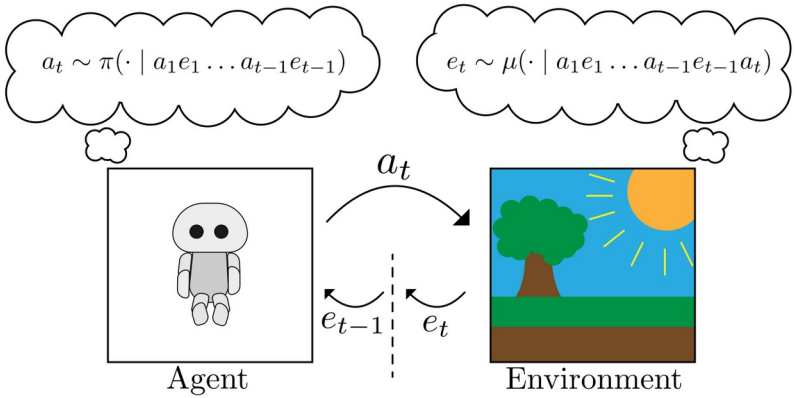
$$\pi_S(h_{<t}) \in \operatorname{argmax}_{a_t} \xi^O(\sum_{i=t}^{\infty} \gamma_i r_i | h_{<t} a_t)$$

Unrealizability & Imprecise Probability

Mathematic Framework for Embedded Agency







???

Imprecise Probability

- Small agents must have misspecified models (if they do model-based planning at all)
- There are various frameworks for representing partial knowledge
- It is already natural to view semimeasures as imprecise probability measures [WH25b]
- Various credal set belief representations have been studied in the context of AIXI
- Arguably, none of this work takes misspecification/unrealizability as seriously as e.g. Infra-Bayesianism

Imprecise Probability

- It is actually harder to study smaller agents!
- Computability → Computational Complexity
- Possibly there is less to say; very small agents are often task-specific
  - We do not want agent foundations to depend on designing the best bacterium!
- I recommend a top-down approach targeted towards specific safety problems and NOT a general theory of boundedly-rational agent design

Citations

[DG20] Abram Demski and Scott Garrabrant. Embedded Agency. Oct. 2020. doi: 10.48550/arXiv.1902.09469. ArXiv: 1902.09469 [cs]. (Visited on 03/29/2025).

[Ele+25] Esraa Elelimy et al. Rethinking the Foundations for Continual Reinforcement Learning. July 2025. doi: 10.48550/arXiv.2504.08161.arXiv: 2504.08161 [cs]. (Visited on 01/15/2026).

[Hut00] Marcus Hutter. A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. arXiv:cs/0004001. Apr. 2000. doi: 10.48550/arXiv.cs/0004001. url: http://arxiv.org/abs/cs/0004001 (visited on 07/01/2024).

[LHG11] Tor Lattimore, Marcus Hutter, and Vaibhav Gavane. “Universal Prediction of Selected Bits”. en. In: Algorithmic Learning Theory. Ed. by Jyrki Kivinen et al. Berlin, Heidelberg: Springer, 2011, pp. 262–276. isbn: 978-3-642-24412-4. doi:10.1007/978-3-64224412-4_22.

[LTF16] Jan Leike, Jessica Taylor, and Benya Fallenstein. “A Formal Solution to the Grain of Truth Problem”. In: Proceedings of the Thirty Second Conference on Uncertainty in Artificial Intelligence. UAI’16. Arlington, Virginia, USA: AUAI Press, June 2016, pp. 427–436. isbn: 978-0-9966431-1-5. (Visited on 07/12/2024).

[Meu+25] Alexander Meulemans et al. Embedded Universal Predictive Intelligence: A Coherent Framework for Multi-Agent Learning. Nov. 2025. doi: 10.48550/arXiv.2511.22226. arXiv: 2511.22226 [cs]. (Visited on 12/14/2025).

[WH25a] Cole Wyeth and Marcus Hutter. Formalizing Embeddedness Failures in Universal Artificial Intelligence. May 2025. doi: 10.48550/arXiv.2505.17882. arXiv: 2505.17882 [cs]. (Visited on 10/02/2025).

[WH25b] Cole Wyeth and Marcus Hutter. “Value Under Ignorance in Universal Artificial Intelligence”. In: Artificial General Intelligence: 18th International Conference, AGI 2025, Reykjavic, Iceland, August 10–13, 2025, Proceedings, Part II. Berlin, Heidelberg: SpringerVerlag, Aug. 2025, pp. 338–349. isbn: 978-3-032-00799-5. doi: 10.1007/978-3-032-00800-8_30. (Visited on 10/02/2025).

[Wye+25] Cole Wyeth et al. Limit-Computable Grains of Truth for Arbitrary Computable Extensive-Form (Un)Known Games. Aug. 2025. doi:10.48550/arXiv.2508.16245. arXiv: 2508.16245 [cs]. (Visitedon 12/23/2025).