Universal AI: AIXI

David Quarel

Iliad Intensive

June 2026

The big picture: one agent, two halves

- **Universal AI** asks: what would an *optimally intelligent* agent look like, given unlimited compute and the weakest possible assumptions?
- It splits into two problems:
  1. **Prediction (Solomonoff):** passively observe a sequence $x_1, x_2, \ldots$ and predict the next symbol. *How should you learn?*
  2. **Action (AIXI):** interact with an unknown environment, take actions, collect reward. *How should you act?*
- **The common engine:** a single **Bayesian mixture** ξ over a class $\mathcal{M}$ of candidate environments, weighted by a prior w_ν.
  - Solomonoff: ξ (eventually) predicts as well as the truth μ.
  - AIXI: the Bayes-optimal policy π^*_ξ (eventually) acts as well as the optimal policy π^*_μ for true environment μ.
- **Universal** choice: $\mathcal{M} =$ all computable environments, prior $w_\nu = 2^{-K(\nu)}$ (K is the "complexity" of the environment). Assumption "$\mu \in \mathcal{M}$" becomes "*the universe is computable*".

The Bayesian mixture ξ

Definition

Class $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$ with prior weights $w_\nu > 0$, $\sum_\nu w_\nu = 1$.

$$\xi(x_{1:t}) := \sum_{\nu \in \mathcal{M}} w_\nu \, \nu(x_{1:t}).$$

- **Key fact (mixture is a predictor):** its one-step prediction is a *posterior-weighted* average of the members' predictions (you prove this today!)

$$\xi(x_t \mid x_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu \mid x_{<t}) \, \nu(x_t \mid x_{<t}), \qquad w(\nu \mid x_{<t}) = w_\nu \, \frac{\nu(x_{<t})}{\xi(x_{<t})}.$$

- **Posterior update is multiplicative:** $w(\nu \mid x_{1:t}) = w(\nu \mid x_{<t}) \cdot \dfrac{\nu(x_t \mid x_{<t})}{\xi(x_t \mid x_{<t})}$: we reweight by how well ν predicted vs. ξ (Bayes' rule!)

From prediction to action

- Now the agent doesn't just watch: it **acts**, and its actions shape what it sees.
- At each step t: the agent picks action a_t, the environment returns a **percept** $e_t = (o_t, r_t) =$ observation + reward.

$$h = a_1\, e_1\; a_2\, e_2\; a_3\, e_3\; \cdots \qquad (\text{history } æ_{<t})$$

- Spaces: actions $\mathcal{A}$, observations $\mathcal{O}$, rewards $\mathcal{R} \subset [0,1]$, percepts $\mathcal{E} = \mathcal{O} \times \mathcal{R}$, and (finite) histories $\mathcal{H}^* := (\mathcal{A} \times \mathcal{E})^* := \cup_{t=0}^{\infty} (\mathcal{A} \times \mathcal{E})^t$.
- **No Markov / MDP assumption!** Both the policy and environment can depend on all prior percepts.

The three players

Policy π

$\pi(a_t \mid æ_{<t}) : \mathcal{H} \to \Delta\mathcal{A}$ is the agent's distribution over the next action given history.

Environment ν

$\nu(e_t \mid æ_{<t}\, a_t) : \mathcal{H} \times \mathcal{A} \to \Delta\mathcal{E}$ is the environment's distribution over the next percept given history and action. True (unknown) environment is μ.

Interaction measure ν^π

Run π inside ν and multiply along the history:

$$\nu^\pi(æ_{t:m} \mid æ_{<t}) := \prod_{k=t}^{m} \pi(a_k \mid æ_{<k})\, \nu(e_k \mid æ_{<k} a_k).$$

Value, optimality, and AIXI

Value function (assuming geometric discount γ)

$$V_\nu^{\pi,m}(æ_{<t}) = (1-\gamma)\,\mathbb{E}_\nu^\pi[G_{t-1:m} \mid æ_{<t}] \text{ with } G_{t-1:m} = \sum_{k=t}^{m} \gamma^{k-t}\, r_k$$

$$= (1-\gamma) \sum_{æ_{t:m}} \nu^\pi(æ_{t:m} \mid æ_{<t}) \sum_{k=t}^{m} \gamma^{k-t}\, r_k$$

The **infinite-horizon value** is defined as the pointwise limit

$$V_\nu^\pi \equiv V_\nu^{\pi,\infty}(æ_{<t}) := \lim_{m\to\infty} V_\nu^{\pi,m}(æ_{<t}).$$

The **optimal value** is $V_\nu^{*,m}(æ_{<t}) := \sup_\pi V_\nu^{\pi,m}(æ_{<t})$, with $V_\nu^* \equiv V_\nu^{*,\infty}$.

- **Optimal value** $V_\nu^* := \sup_\pi V_\nu^\pi$; an optimal policy π_ν^* satisfies $V_\nu^{\pi_\nu^*} = V_\nu^*$.
- **Bayes-optimal policy** $\pi_\xi^* \in \arg\max_\pi V_\xi^\pi$: acts optimally w.r.t. the *mixture* ξ.

AIXI

AIXI

π^*_ξ with $\mathcal{M}$ = all lower-semicomputable environments and prior $w_\nu = 2^{-K(\nu)}$. Write ξ_U for ξ with this choice of $\mathcal{M}$ and w_ν, so AIXI $= \pi^*_{\xi_U}$.

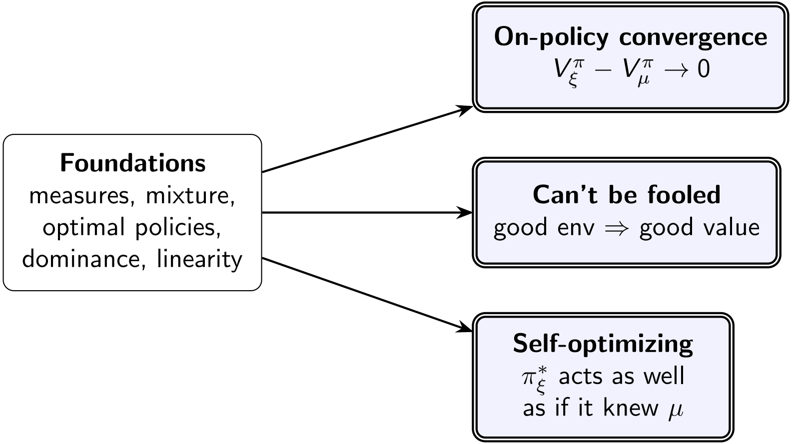
Intuition: *The Bayes-optimal agent with prior according to Occam's razor: Act optimally according to environments that you believe to be in, assuming the simplest environments consistent with observations are more likely to be true.*

AIXI, explicitly

max over actions, $\sum$ over percepts, choosing actions that maximise the ξ-weighted discounted return:

$$a_t = \arg\max_{a_t} \lim_{m\to\infty} \sum_{e_t} \max_{a_{t+1}} \sum_{e_{t+1}} \cdots \max_{a_m} \sum_{e_m} \xi(e_{t:m} \mid æ_{<t}\, a_{t:m}) \sum_{k=t}^{m} \gamma^{k-t} r_k$$

The three main AIXI results



- All three: the Bayesian agent inherits the guarantees it would have if it knew the truth.

Result 1: On-policy value convergence

Theorem

For any $\mu \in \mathcal{M}$ and any policy π,

$$V_\xi^\pi(æ_{<t}) - V_\mu^\pi(æ_{<t}) \to 0 \qquad \mu^\pi\text{-almost surely.}$$

Intuition: *A policy that acts, pretending the environment is ξ, will eventually act as well as if it were in the true environment μ.*

Result 2: AIXI can't be fooled

Theorem (deterministic μ)

If the truth is reachable-good, $V^*_\mu(æ_{<t}) > \varepsilon$ along the played history, then

$$V^*_\xi(æ_{<t}) \;\geq\; w_\mu\,\varepsilon \;>\; 0.$$

Intuition: *Whenever good behaviour in μ is achievable, the Bayes-optimal agent secures at least a fraction proportional to the prior weight of μ.*

Result 3: Self-optimizing property (stretch goal)

Theorem

If *some* policy can learn to act optimally in every $\nu \in \mathcal{M}$, then π^*_ξ does too:

$$V^*_\mu(æ_{<t}) - V^{\pi^*_\xi}_\mu(æ_{<t}) \to 0 \quad \mu^\pi\text{-a.s.}$$

Intuition: *The Bayes-optimal agent that is unaware of which environment it is in learns to act just as well as the best possible policy that already knows which environment it is in (μ).*

Takeaways

- **Solomonoff** solves induction: mix over all computable hypotheses, weighted by simplicity. Total prediction error $\leq K(\mu) \ln 2$.
- **AIXI** lifts this to action: the Bayes-optimal policy over the same universal mixture. It learns to predict *and* to act as well as if it knew the world.
- **The catch:** both are *incomputable* (K is incomputable; $\mathcal{M}$ is intractable to enumerate). They are gold standards, not computable algorithms, and the prior's $2^{-K(\nu)}$ hides a dependency on the choice of universal Turing machine.
- **Why care for safety:** a precise, assumption-light model of idealized intelligence providing a lens on what optimal agents do. For example, we can show under what circumstances AIXI would wirehead [RO11], safely rather than recklessly explore [CCH20], or self-modify [OR11] as a model of what superintelligent agents would do.

References I

[CCH20] Michael K. Cohen, Elliot Catt, and Marcus Hutter.
Curiosity killed or incapacitated the cat and the asymptotically optimal agent.
CoRR, abs/2006.03357, 2020.
URL: https://arxiv.org/abs/2006.03357.

[OR11] Laurent Orseau and Mark Ring.
Self-modification and mortality in artificial agents.
In *Artificial General Intelligence (AGI 2011)*, volume 6830 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2011.
URL: https://link.springer.com/chapter/10.1007/978-3-642-22887-2_1, doi:10.1007/978-3-642-22887-2_1.

[RO11] Mark Ring and Laurent Orseau.
Delusion, survival, and intelligent agents.
In *Artificial General Intelligence (AGI 2011)*, volume 6830 of *Lecture Notes in Computer Science*, pages 11–20. Springer, 2011.
URL: https://people.idsia.ch/~ring/AGI-2011/Paper-B.pdf, doi:10.1007/978-3-642-22887-2_2.