Formalizing Embeddedness Failures in Universal Artificial Intelligence

Cole Wyeth[1] and Marcus Hutter[2]

[1]Cheriton School of Computer Science , University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada , https://colewyeth.com/
[2]Google DeepMind & Australian National University,
http://www.hutter1.net/

April 2025

Abstract

We rigorously discuss the commonly asserted failures of the AIXI reinforcement learning agent as a model of embedded agency. We attempt to formalize these failure modes and prove that they occur within the framework of universal artificial intelligence, focusing on very simple variants of AIXI. We introduce joint AIXI, which models the joint action/percept history as drawn from the universal distribution, and hardened AIXI, which recovers from side-channel attacks by recalculating its own previous actions. We also evaluate the progress that has been made towards a successful theory of embedded agency based on variants of the AIXI agent, placing our work in relation to more drastic departures from AIXI. While other approaches such as reflective oracles can construct agents with more desirable properties, our study better illustrates the concrete challenges of embedded agency and tests the limits of what can be achieved using only the standard tools of algorithmic information theory.

Contents

Keywords

Universal artificial intelligence, Solomonoff induction, evidential decision theory

1 Introduction

The original AIXI reinforcement learning agent, intended as a nearly parameter-free formal gold standard for artificial general intelligence (AGI), is a *Cartesian dualist* that believes it is interacting with an environment from the outside, in the sense that its *policy* is fixed and not overwritten by anything that happens in the environment, though its *actions* can certainly adapt based on the percepts it receives. This is frequently compared to a person playing a video game, who certainly does not believe he is being simulated by the game but rather interacts with it only by observing the screen and pressing buttons. In contrast, it would presumably be important for an AGI to be aware that it exists within its environment (the universe) and its computations are therefore subject to the laws of physics [DG20]. With this in mind, we investigate versions of the AIXI agent [Hut00] that treat the action sequence a on a similar footing to the percept sequence e, meaning that the actions are considered as explainable by the same rules generating the percepts. The most obvious idea is to use the universal distribution to model the joint (action/percept) distribution (even though actions are selected by the agent). Although this is the most direct way to transform AIXI into an embedded agent, it does not appear to have been analyzed in detail; in particular, it is usually assumed (but not proven) to fail (often implicitly, without distinguishing the universal sequence and environment distributions, e.g. [FST15]).

Outline. First, we discuss the more sophisticated approaches to embedded AIXI-like agents as points of reference. Then we give a highly compressed introduction to the notions of algorithmic information theory (AIT) and particularly universal artificial intelligence (UAI) needed to model universal mixtures of sequence distributions and environments, followed by some mappings that relate sequence and environment distributions. We introduce some positive and some negative results for our embedded AIXI variant (called joint AIXI), each of which is perhaps a little surprising, but follows easily from recent progress in AIT. The limitations of joint AIXI motivate hardened AIXI, which also attempts to learn from its own actions, but only to discover side-channel action corruption. This narrower scope allows sharper results characterizing hardened AIXI in terms of AIXI.

Terminology for the problem setting. AIXI has a harder incomputability level than the environments in its hypothesis class, since its actions are not sampled from its belief distribution. This means that we are analyzing an *unrealizable* situation, where the interaction history is generated by a process outside the hypothesis class. For the purposes of learning the joint distribution, it is hard to find any useful guarantees on the action sequence, so we will treat it as *adversarially* chosen and consider the worst case.

2 Related Work

Several sophisticated embedded versions of AIXI have been proposed.

Reflective oracles. By expanding AIXI's hypothesis class to include machines with access to a special type of "reflective" oracle [FTC15], researchers at the Machine Intelligence Research Institute (MIRI) were able to construct a version of AIXI that is at the same incomputability level as the environments in its hypothesis class [FST15]. This reflective AIXI faces a *realizable* learning problem, so there is no reason to view its own actions as adversarially chosen. In fact, reflective oracle machines can directly compute the conditional probabilities of actions and percepts (not just their joint distributions) so that the perspective change between sequence and environment distribution becomes trivial. We believe that reflective AIXI is an excellent (and perhaps underappreciated) approach to embeddedness, primarily because there is a limit-computable reflective oracle [LTF16], which means that reflective AIXI has a stochastic anytime algorithm. This is a better computability result than has been demonstrated for AIXI, which may be as hard as Δ_3^0 without ε-approximation [LH18], meaning that reflective AIXI not only addresses embeddedness concerns but actually suggests an effective AIXI approximation. The main (serious) limitation of reflective AIXI is that it is still a computationally unbounded model of intelligence, so it is not clear how it treats recursive self-improvement (this is an interesting research question).

Self-AIXI. The Self-AIXI agent [CGMH23] functions in a similar way to reflective AIXI, but it is uncertain of its own policy as well as its environment. The framework in the paper does not specify particular belief distributions for either of these, but asks for choices that make an optimal policy lie in the policy class, a realizability assumption. It is easy to find such a choice by taking advantage of reflective oracles (this is proven in a forthcoming paper of ours, [WHLT25]). Self-AIXI plans only one step ahead to locally maximize its action-value function; the variation of our joint AIXI based on this strategy can easily be analyzed by the same methods introduced in this paper, though our focus on percept prediction becomes less justifiable.

Space-time embedded intelligence. Orseau and Ring depart more drastically from the original AIXI model but continue to take advantage of the tools of UAI by proposing various degrees of embeddedness of an agent inside the universe's computation [OR12]. In their most extreme model, the agent is only a collection of bits on the environment machine's tape. Although this model seems to capture all problems of embeddedness, it does not (naively) appear to offer clear advice or inspiration for constructing implementations.

Embedded decision theories. More extreme departures from the UAI framework such as Infra-Bayesian Physicalism [Kos21], pancomputational enactivism [Ben24], and updateless decision theory [Dai09] are beyond the scope of this paper. Our focus is on the most direct modifications of AIXI for embedded agency, which seems seriously neglected in the literature.

3 Mathematical Preliminaries

Notation. For any finite alphabet Σ, we denote the set of finite strings over Σ as Σ^* and the set of infinite sequences over Σ as Σ^∞. If $s \in \Sigma^*$ or $s \in \Sigma^\infty$, then $s_i \in \Sigma$ is the i^{th} symbol of s, indexing from 1. Similarly $s_{i:j}$ for $i \leq j$ is the substring from indices i to j. The length $l(s)$ is the number of symbols in s. The probability simplex over Σ is denoted $\Delta\Sigma$. Particular alphabets we will discuss are the set of actions $\mathcal{A}$ available to an agent and the set of percepts $\mathcal{E}$ that the environment ν might produce and send to the agent. A percept consists of an observation in $\mathcal{O}$ and a reward in $\mathcal{R} \subset \mathbb{R}$. The action/percept at time t will be denoted a_t/e_t, and the history of actions and observations before time t will be written $æ_{<t} = a_1 e_1 ... a_{t-1} e_{t-1}$.

Definition 1 (lower semicomputable) *A function f is* lower semicomputable *(l.s.c.) if there is a computable function $\phi(x,k)$ monotonically increasing in its second argument with $\lim_{k\to\infty}\phi(x,k) = f(x)$. That is, f can be approximated from below.*

In AIT the history distribution often has a probability gap because interaction may terminate at a finite time. For this reason, it is modeled by a "semimeasure," not a proper probability measure.

Definition 2 (semimeasure) *A semimeasure*[1] *ν is a function $\Sigma^* \to \mathbb{R}^+$ satisfying $\nu(x) \geq \sum_{a\in\Sigma}\nu(xa)$.*

For our purposes, semimeasures always assign probability ≤ 1 to the empty string ϵ. The set of such l.s.c. semimeasures is denoted $\mathcal{M}^{\text{semi}}_{\text{lsc}}$. It is possible to construct (and fix) a computable enumeration $\mathcal{M}^{\text{semi}}_{\text{lsc}} = \{\nu_i | i \in \mathbb{N}\}$.

Definition 3 ($\boldsymbol{\xi_U}$) *The universal distribution ξ_U is defined as*

$$\xi_U(æ_{1:t}) := \sum_{i\in\mathbb{N}} w_i \nu_i(æ_{1:t}) \tag{1}$$

for $æ \in (\mathcal{A}\times\mathcal{E})^\infty$, where $i \mapsto w_i > 0$ is a l.s.c. function with $\sum_i w_i \leq 1$, e.g. $w_i := [i(i+1)]^{-1}$. An alternative construction is

$$\xi_U(x) \overset{\times}{=} \sum_{p:U(p)=x*} 2^{-l(p)} \tag{2}$$

with monotone UTM U (a "joint" distribution producing sequences which are NOT action contextual).

For simplicity of exposition we assume $\mathcal{A} = \mathcal{E}$ by expanding the smaller alphabet.

Chronological semimeasures. We write $\nu^\cdot(e_{1:t}||a_{1:t})$ to denote the probability that the environment $\nu^\cdot$ produces percepts $e_{1:t}$ when the agent takes actions

[1] Technically, this only defines a pre-semimeasure, but a unique *sensible* extension to the generated σ-algebra exists [WH25].

$a_{1:t}$. Formally, this notation is used whenever $\nu^{\cdot}(\cdot||\cdot)$ is a two-argument string function $\bigcup_{n\in\mathbb{N}}\mathcal{E}^n\times\mathcal{A}^n\to[0,1]$ satisfying $\nu^{\cdot}(e_{1:t}||a_{1:t})\geq\sum_{e_{t+1}}\nu^{\cdot}(e_{1:t+1}||a_{1:t+1})$, called a chronological semimeasure[2]. Despite the name, when the two arguments are treated as forming one interaction history $æ_{1:t}$, this is actually a weaker requirement than ordinary semimeasures must satisfy. Usually, the superscript $\cdot$ is replaced by some index or identifying symbol, distinguishing environments from sequence distributions, which use subscripts when any adornment is required. For instance, we can computably enumerate the set of lower semicomputable chronological semimeasures $\mathcal{M}^{\text{ccs}}_{\text{lsc}}:=\{\nu^i|i\in\mathbb{N}\}$. We can also view policies as chronological semimeasures over actions given percepts, but strictly speaking this results in a slightly different class because actions precede percepts, so the function type becomes $\bigcup_{n\in\mathbb{N}}\mathcal{A}^n\times\mathcal{E}^{n-1}\to[0,1]$ and the condition becomes $\pi(a_{1:t}||e_{1:t-1})\geq\sum_{a_{t+1}}\pi(a_{1:t+1}||e_{1:t})$.

Semimeasure representation. It is clear that for any semimeasure $\nu_{\cdot}$, the map $e_{1:t},a_{1:t}\to\prod_{i=1}^{t}\nu_{\cdot}(e_i|æ_{<i}a_i)$ defines a chronological semimeasure that we will write as $\nu^{\cdot}(e_{1:t}||a_{1:t})$. Conversely, if $\nu^{\cdot}$ is a chronological semimeasure, we can choose $\nu_{\cdot}$ so that $\nu_{\cdot}(a_i|æ_{<i})=1/|\mathcal{A}|$ (or any arbitrary element of $\Delta\mathcal{A}$) and $\nu_{\cdot}(e_i|æ_{<i}a_i)=\nu^{\cdot}(e_{1:i}||a_{1:i})/\nu^{\cdot}(e_{<i}||a_{<i})$ (and $\nu_{\cdot}$ has 0 probability of producing a percept in an action position or vice versa). Then $\nu_{\cdot}$ obviously satisfies the semimeasure property at action positions and satisfies the semimeasure property at percept positions by the chronological semimeasure property of $\nu^{\cdot}$, and it is easy to see that $\nu^{\cdot}(e_{1:t}||a_{1:t})=\prod_{i=1}^{t}\nu_{\cdot}(e_i|æ_{<i}a_i)$. Satisfying the chronological semimeasure property is equivalent to having such a semimeasure representation; unfortunately, this representation result does not seem to hold when we restrict to *l.s.c.* (chronological) semimeasures[3].

Definition 4 ($\boldsymbol{\xi^{\text{AI}}}$) *The universal chronological semimeasure ξ^{AI} is defined by*

$$\xi^{AI}(e_{1:t}||a_{1:t}):=\sum_{i\in\mathbb{N}}w_i\nu^i(e_{1:t}||a_{1:t}) \tag{3}$$

with w_i as in Definition 3. Alternatively, we can obtain ξ^{AI} by

$$\xi^{AI}(e_{1:t}||a_{1:t})\overset{\times}{=}\sum_{p:U^C(p,a_{1:t})=e_{1:t}}2^{-l(p)} \tag{4}$$

where "chronological" UTM U^C only reads actions up to time t before producing e_t.

Note that when ξ^{AI} is viewed as a function of $æ_{1:t}$, it is not a semimeasure because it is not subadditive at action indices.

[2]The older, more verbose term is chronological contextual semimeasure, sometimes abbreviated ccs.

[3]The related fact that the conditionals of an l.s.c. semimeasure may not be l.s.c. follows easily from a diagonalization argument for ξ_U. It seems harder to find a clean example that *clearly* shows the chronological semimeasure on half of the bit positions, induced by env of an l.s.c. semimeasure, is not l.s.c.

Definition 5 (domination) *A semimeasure ν (multiplicatively) dominates a semimeasure μ, written $\nu \overset{\times}{\geq} \mu$, if $\exists c \in \mathbb{R}^+$ such that $\forall x \in \Sigma^*$, $\nu(x) \geq c\mu(x)$.*

For chronological semimeasures dominance requires the above to hold for each action sequence. Multiplicative dominance is often used to establish the log loss bound $-\log \nu(x) \overset{+}{\leq} -\log \mu(x)$ and is a stronger condition than absolute continuity, which is the usual criterion for merging-of-opinions style results [BD62].

An agent's goal is to maximize its cumulative (often discounted) reward by choosing an optimal sequence of actions. The optimal policy is defined as

$$\begin{aligned} \pi^*_\nu &= \text{argmax}_\pi V^\pi_\nu \\ &:= \text{argmax}_\pi \sum_{t=1}^{m} \sum_{æ_{1:t}} \gamma_t r_t \pi(a_{1:t}||e_{1:t-1}) \nu(e_{1:t}||a_{1:t}) \end{aligned} \tag{5}$$

where γ_t is a discount function ($\gamma_t \geq 0$, usually $\sum_{t=1}^{m} \gamma_t = 1$) and m may be infinite.

When π is deterministic, we can abuse notation by treating it as a function from histories to actions. Then, in the case that ν is a measure and m is finite, we can write the constraint[4] on the optimal policy's action choice explicitly as

$$\pi^*_\nu(æ_{1:t}) \in \text{argmax}_{a_{t+1}} \sum_{e_{t+1}} ... \max_{a_m} \sum_{e_m} \nu(e_{1:m}||a_{1:m}) \sum_{i=t+1}^{m} \gamma_t r_t \tag{6}$$

With discounting, this can also be extended to infinite horizons, but the distinction will not be important for us.

4 The Dualistic Mixture

We introduce maps dual and env that respectively combine a policy and environment to get a history distribution and convert a (semimeasure) history distribution (back) to an environment.

An observation we will find useful is that given any pair of l.s.c. chronological semimeasures ν generating percepts and π generating actions, the history distribution ν^π that they induce is an l.s.c. semimeasure. To avoid interfering with (super/sub)scripts, we usually write $\text{dual}(\nu,\pi) := \nu^\pi$. There is a semimeasure that encodes the assumption that actions are generated by a l.s.c. agent and percepts by an l.s.c. environment:

$$\xi_{\text{dual}} := \sum_{\nu,\pi \in \mathcal{M}^{\text{semi}}_{\text{lsc}}} w^\pi_\nu \text{dual}(\nu,\pi)$$

We will assume that $w^\pi_\nu = \omega_\pi w_\nu$ (agent and environment are independent). This is the assumption made by Self-AIXI which makes it less general than our

[4] Technically this constraint may be ignored off-policy.

joint AIXI. A natural choice is $w_\nu^\pi = 2^{-K(\pi)}2^{-K(\nu)}$ where K is the Kolmogorov complexity. It is immediate that

$$\xi_U \overset{\times}{\geq} \xi_{\text{dual}} \tag{7}$$

Let

$$\text{env}(\nu)(e_{1:t}||a_{1:t}) := \prod_{i=1}^{t} \nu(e_i|æ_{<i}a_i)$$

To be well-defined, this requires $\nu > 0$, which holds for ξ_U. When $w_\nu^\pi = \omega_\pi w_\nu$, factoring yields $\xi_{\text{dual}} = \text{dual}(\sum_\pi \omega_\pi \pi^\cdot, \sum_\nu w_\nu \nu^\cdot)$ so $\text{env}(\xi_{\text{dual}}) = \sum_\nu w_\nu \nu^\cdot \overset{\times}{=} \xi^{\text{AI}}$.

5 Joint AIXI

Now we introduce our joint AIXI model which uses ξ_U to model the joint distribution. In particular we investigate its relationship to the conventional ξ^{AI}. Let $\xi^U := \text{env}(\xi_U)$. Then

$$\begin{aligned}
\xi^U(e_t|æ_{<t}a_t) &= \xi_U(e_t|æ_{<t}a_t) \\
&= \frac{\sum_i w_i \nu_i(æ_{<t}a_t e_t)}{\xi_U(æ_{<t}a_t)} \\
&= \sum_i w_i(æ_{<t}a_t)\nu_i(e_t|æ_{<t}a_t) \\
&= \sum_i w_i(æ_{<t}a_t)\nu^i(e_t|æ_{<t}a_t)
\end{aligned} \tag{8}$$

where

$$w_i(æ_{<t}a_t) := \frac{w_i \nu_i(æ_{<t}a_t)}{\xi_U(æ_{<t}a_t)} \tag{9}$$

so ξ^U is not a linear combination of ν^i because of the way that the actions control the weights; ξ^U encodes a kind of sequential action evidential decision theory [ELH15], because semimeasures that prefer the action sequence a are weighted more heavily as environments in ξ^U. Joint AIXI is defined as a Bayes-optimal policy for this belief distribution[5]:

$$\pi^{\text{JAIXI}} := \pi^*_{\xi^U} \tag{10}$$

An aside. An arguably more natural definition is

$$\xi^{\text{env}(U)}(e_{1:t}||a_{1:t}) := \sum_i w_i \text{env}(\nu_i)(e_{1:t}||a_{1:t}) \overset{\times}{\geq} \text{env}(\xi_U)(e_{1:t}||a_{1:t}) \tag{11}$$

[5] Arguably, this decision rule is somewhat short-sighted: despite planning ahead, it does not condition on its intended future decisions - that is, $\xi^U(e_t|æ_{<t}a_t)$ does not depend on $a_{t+1:\infty}$. In contrast $\xi^{\text{AI}}_{\text{alt}}$ of [Hut05], when paired with the iterative value function, can be viewed as attempting to update on intended actions in advance, but actually fails to meet the conditions of a chronological semimeasure (our unpublished result).

Another way of writing this is

$$\xi^{\mathrm{env}(U)}(e_t|æ_{<t}a_t)=\sum_\nu w_\nu(e_{<t}||a_{<t})\nu(e_t|h_{<t}a_t)$$

where $w_\nu(e_{<t}||a_{<t})=\mathrm{env}(\nu)(e_{<t}||a_{<t})/\xi^{\mathrm{env}(U)}(e_{<t}||a_{<t})$. Note that $w(\cdot||\cdot)$ is not a strictly correct use of the $||$ notation, but only indicates a ratio of chronological semimeasures. $\xi^{\mathrm{env}(U)}$ is not the same as $\mathrm{env}(\xi_U)$, but in fact dominates $\mathrm{env}(\xi_U)$ by Eq. (11). Also, for any l.s.c. policy π, $\mathrm{dual}(\xi^{\mathrm{AI}},\pi)$ is an l.s.c. semimeasure, so

$$\xi^{\mathrm{env}(U)}(e_{1:t}||a_{1:t}) \overset{\times}{\geq} \mathrm{env}(\mathrm{dual}(\xi^{\mathrm{AI}},\pi))(e_{1:t}||a_{1:t})=\xi^{\mathrm{AI}}(e_{1:t}||a_{1:t})$$

so $\xi^{\mathrm{env}(U)}$ also dominates the explicitly causal ξ^{AI}! This makes it a rather fascinating mixture, which seems to give some weight to both CDT and EDT, but to judge them by the standards of CDT when updating those weights. Therefore, it appears that $\xi^{\mathrm{env}(U)}$ is more conceptually complicated than ξ^U, and perhaps difficult to justify on philosophical grounds. We leave a more detailed analysis to future work, and turn our focus to ξ^U.

Adversarial learning. We can adapt results on the prediction of selected bits directly from [LHG11] to understand the relationship between ξ^U and ξ^{AI}. The authors investigated whether computable structure at certain computable indices of a sequence can be learned by the universal distribution even when the rest of the sequence is adversarially chosen. They were motivated by supervised learning, where the distribution of examples may be much more complicated/unpredictable than the distribution of labels. Our motivation is different: we expect the even = percept indices to come from an l.s.c. environment distribution (given the odd = action indices as context), but this simple partition of the indices matches the example case in [Hut05, Problem 2.9].

Failure to learn a simple environment. According to [LHG11, Theorem 12], it is possible for ξ_U to fail to predict a binary sequence at even indices despite even bits exactly matching the preceding bits at odd indices. For example, such a sequence might begin something like 00 11 11 00 11..., where the odd values cannot be predicted by any computable rule. Formally (translated to our notation),

Theorem 6 (Adversarial non-convergence of $\boldsymbol{\xi_U}$) *There exists $\omega \in \mathbb{B}^\infty$ with $\omega_{2n}=\omega_{2n-1}$ but $\liminf_n \xi_U(\omega_{2n}|\omega_{1:2n-1})<1$.*

Such ω must not be computable, but the theorem is still surprising since the even bits are very easy to predict for a human. Now consider the simplest possible environment μ^{id} with binary action space, empty observation space, and binary reward space, defined by

$$\mu^{\mathrm{id}}(e_t|æ_{<t}a_t)=[\![e_t=a_t]\!] \tag{12}$$

Clearly, this is a l.s.c. chronological semimeasure.

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 $t{\to}\infty$.*

Proof. This is a direct result of Theorem 6. ■

Theorem 8 $\xi^U \overset{\times}{\not\geq} \xi^{AI}$

Proof. Choose $e{=}a$ according to Theorem 7 and note that

$$\xi^{\mathrm{AI}}(a_{1:t}||a_{1:t}) \overset{\times}{\geq} \mu_{\mathrm{id}}(a_{1:t}||a_{1:t}){=}1 \tag{13}$$

while $\xi^U(a_{1:t}||a_{1:t}){\to}0$. ■

We expect that domination fails in the other direction as well because ξ^U's posteriors (as in Eq. (9)) treat a much differently than ξ^{AI}'s posteriors, which should sometimes be advantageous for prediction.

Conjecture 9 $\xi^{AI} \overset{\times}{\not\geq} \xi^U$

Normalization allows learning computable environments. Recall that ξ_U is only a semimeasure and not a proper probability measure. The most common "normalization" or completion of ξ_U to a measure is called Solomonoff normalization:

$$\hat{\xi}_U(\omega_t|\omega_{<t}){=}\frac{\xi_U(\omega_t|\omega_{<t})}{\sum_{\omega'_t\in\Sigma}\xi_U(\omega'_t|\omega_{<t})} \tag{14}$$

Surprisingly, applying this normalization allows a type of learning in the adversarial case. We translate [LHG11, Theorem 10] as follows:

Theorem 10 (Adversarial learning for $\boldsymbol{\hat{\xi}_U}$) *Let $f:\mathbb{B}^*{\to}\mathbb{B}{\cup}\{\epsilon\}$ be a total recursive function and consider $\omega{\in}\mathbb{B}^\infty$ satisfying $\omega_n{=}f(\omega_{<n})$ when $f(\omega_{<n}){\neq}\epsilon$. Then for any infinite sequence $n_1,n_2,...$ with $f(\omega_{n_i}){\neq}\epsilon$, $\lim_{i\to\infty}\hat{\xi}_U(\omega_{n_i}|\omega_{<n_i}){=}1$.*

In other words, if the bits at certain recursively checkable indices are a recursive function of the preceding sequence, $\hat{\xi}_U$ will eventually learn to make good predictions at those indices. Unfortunately, this positive result seems unlikely to generalize to the stochastic case (though this is an open problem).

Theorem 11 (Adversarially learning environments) *If $\mu^{\cdot}\in\mathcal{M}^{ccs}_{lsc}$ is deterministic and $e{\sim}\mu^{\cdot}(\cdot|a)$ with $l(e){=}\infty$, then $\lim_{t\to\infty}\mathrm{env}(\hat{\xi}_U)(e_t|æ_{<t}a_t){=}1$.*

Proof. A deterministic environment that is l.s.c. must also be recursive in the sense that the next percept is finitely computable from the history (it must also be a measure given this action sequence for the history to be infinite). If the action or percept spaces are not binary, choose a fixed binary encoding, and these statements also apply at the bit level. Observe that $\mathrm{env}(\hat{\xi}_U)(e_t|æ_{<t}a_t){=}\hat{\xi}_U(e_t|æ_{<t}a_t)$ by definition and apply Theorem 10. ■

We have shown (in forthcoming work) that this result is not too sensitive to the specific normalization chosen.

Implications. As a corollary of Theorem 11, in deterministic environments, $\pi^{\text{nJAIXI}} := \pi^*_{\hat{\xi}_U}$ learns to predict the percepts produced in response to its chosen action sequence. For simplicity assume π^{nJAIXI} is chosen to be deterministic itself. This means that it correctly predicts the rewards that will be obtained on-policy. From this, it is easy to see that π^{nJAIXI} will not perform in *extremely* suboptimal ways, such as indefinitely picking an action that yields minimal reward. Note that this is the reason we chose a sequential-planning-based decision rule for π^{nJAIXI}. As an alternative more in the spirit of embeddedness, we could have instead defined it to maximize the one-step-ahead action-value function, selecting $a^*_t = \text{argmax}_{a_t} \mathbb{E}_{\hat{\xi}_U}[\sum_{i=t}^{\infty} \gamma_i r_i | æ_{<t} a_t]$, which takes advantage of the learned action conditionals.[6] Sadly, our results do not yield any (even weak) performance guarantees for that alternative policy, since action conditionals have not been shown to converge to any reasonable value.

6 Hardened AIXI

Now we will assess joint AIXI as an embedded agent and argue that it does not satisfactorily address action corruption through side channels. We introduce hardened AIXI as an alternative that directly solves this problem. Hardened AIXI is a slightly more complicated (hyper)algorithm which is motivated somewhat naturally by the limitations of joint AIXI. However, its learning from actions appears to be of narrower scope - it is "less embedded."

Learning from action selection. Though we formulated joint AIXI as planning ahead sequentially in Eq. (10), in a certain sense, it is ignorant of its own policy. It updates every time that it takes an action, affecting its percept predictions and therefore its planning. This is arguably desirable, because it means that joint AIXI learns about the world through observing its own behavior - a type of anthropic reasoning.

Side-channel effects on action selection. Though it makes sense to ask about joint AIXI's action predictions, they are not used for planning. This means that joint AIXI does not seem to solve one of the core challenges of embedded agency: it does not take advantage of the knowledge that the environment may corrupt its action choices through side channels. For instance, if (an approximation of) the joint AIXI policy operates a welding robot, and sustained welding causes its processor to overheat and select random actions, joint AIXI may learn this but still plans *as if* action selection were unaffected. One can formulate a version of joint AIXI that plans only one step ahead and uses its belief distribution to predict its future actions, but we have not proven any performance guarantees about that version.

[6] This decision rule can also be expressed as a Self-AIXI policy with appropriately chosen policy and environment mixtures.

Hardening against action corruption. We will now introduce a simple variant of AIXI that directly solves action corruption. The motivating idea is that if the environment corrupts AIXI's action selection process, this can be detected through a recursive recalculation of AIXI's true actions, and the corruption process can be viewed as part of the environment[7]. This variant will be called hardened AIXI (inspired by radiation-hardened algorithms, which are similarly robust to adversarial bit flips through hardware side channels).

Select a deterministic AIXI policy π^{AIXI}, and recursively define

$$\begin{aligned} a^*(\epsilon) &:= \pi^{\text{AIXI}}(\epsilon) \\ a^*(e_{<t}) &:= \pi^{\text{AIXI}}(a^*(\epsilon)e_1 ... a^*(e_{<t-1})e_{t-1}) \end{aligned} \tag{15}$$

and let

$$\pi^{\text{HAIXI}}(æ_{<t}) = a^*(e_{<t}) \tag{16}$$

Note that π^{HAIXI} ignores whatever actions appear on its "action tape" (which can be considered as part of the input to the policy) and simply recalculates the actions that the π^{AIXI} policy would have taken. This means that it only acts differently from AIXI off-policy. Hardened AIXI has stronger self-knowledge than AIXI: it assumes that it has *always* followed the AIXI policy.

Do humans even have an action tape? It is worth reflecting here on which version of AIXI best describes humans. In particular, what is the meaning of our action sequence, and do we remember it? This discussion requires that we are careful about drawing the lines between a person and her environment. Certainly we observe and recall our own external actions (at some level of abstraction); we hear the words we say and we see our hands grasping and picking up objects. However, these events and even the feeling of performing them are actually a part of our percept stream, which both AIXI and hardened AIXI maintain similar access to. We can call these self-observations which make up a part of our percepts $\text{im}(a)$ to stand for the image of our actions (say, in our own eyes). Perhaps the "true" actions are the *internal* conscious decisions that precede physical acts. The difficulty of locating a suggests that perhaps memory only represents $\text{im}(a)$, and the illusion of access to a is created online by retrospection - this is a much closer match to hardened AIXI than AIXI, though of course this argument is far from rigorous.

Formal equivalence with uncorrupted AIXI. Now we will formally understand the behavior of hardened AIXI under action corruption in terms of AIXI. Let $f : \cup_{n \geq 0}(\mathcal{A}^{n+1} \times \mathcal{E}^n) \to \Delta\mathcal{A}$ be a function that "corrupts" a policy by converting an action/percept history and a next action chosen by the policy into a probability distribution describing the next action that is actually taken. We will use a to refer to the chosen action sequence and a' to refer to the actual action sequence, with $æ'$ as shorthand for $a'_1 e_1 a'_2 e_2 ....$ Then the f-corrupted version of a policy π is defined as

[7] This trick is somewhat similar to Alexander et. al's "reality check transformation" [ACCM22].

$$(f\circ\pi)(a'_t|æ'_{<t}) := \sum_{a_t} f(a'_t|a'_{<t}a_t,e_{<t})\pi(a_t|æ'_{<t}) \tag{17}$$

That is, the probability that $f\circ\pi$ takes action a'_t given that it has so far chosen actions $a'_{<t}$ and received percepts $e_{<t}$ is the sum over the conditional probabilities of f on a'_t when π takes each possible action a_t.

Example 12 (Magnet factory) *Consider a robotic agent operating in a magnet factory that is tasked with sorting magnets of various sizes from a conveyor belt into bins A or B. Say* $\mathcal{O} = \{small,medium,large,none\}$ *and* $\mathcal{A} = \{pass,pickup,drop_A,drop_B\}$. *Picking up a large magnet causes the agent's decision making module to return a failure with probability* $1/2$, *resulting in a default action of pass. Normally,* $f(\cdot|a_{<t}a_t,e_{<t})$ *is concentrated on* a_t, *but if* $a_{t-1} = pickup$ *and* $o_{t-1} = large$, *we have*

$$f(a'_t|a_{<t}a_t,e_{<t}) = \frac{1}{2}[\![a'_t = pass]\!] + \frac{1}{2}[\![a'_t = a_t]\!] \tag{18}$$

♦

Instead of treating action corruption as part of the agent's policy we can consider it part of the environment. Given "base" environment μ, we can incorporate action corruption as

$$(\mu\circ f)(e_{1:t}||a_{1:t}) := \sum_{a'_{1:t}} \mu(e_{1:t}||a'_{1:t}) \prod_{i=1}^{t} f(a'_i|a'_{<i}a_i,e_{<i}) \tag{19}$$

We must introduce a little more shorthand notation: $æ^* := a^*(\epsilon)e_1a^*(e_1)e_2...$, that is, the action/percept history generated when AIXI receives the percept stream e, and $a^*(e)$ denotes just the resulting action stream. With that understanding, we have the fairly elegant "theorem" (really just a direct consequence of our definitions) that follows:

Theorem 13 *The hardened AIXI policy corrupted by* f *and facing environment* μ *"acts like" the corresponding deterministic AIXI policy in the environment* $\mu\circ f$ *that silently corrupts all actions by* f. *Formally,*

$$\sum_{a'_{1:t}} \text{dual}(\mu, f\circ\pi^{HAIXI})(æ'_{1:t}) = \text{dual}(\mu\circ f, \pi^{AIXI})(æ^*_{1:t}) \tag{20}$$

Proof.

$$\begin{aligned}\text{dual}(\mu,f\circ\pi^{\text{HAIXI}})(\text{æ}'_{1:t}) &= \mu(e_{1:t}||a'_{1:t})\prod_{i=1}^{t}(f\circ\pi^{\text{HAIXI}})(a'_i|\text{æ}'_{<i}) \\ &= \mu(e_{1:t}||a'_{1:t})\prod_{i=1}^{t}\sum_{a_i} f(a'_i|a'_{<i}a_i,e_{<i})\pi^{\text{HAIXI}}(a_i|\text{æ}'_{<i}) \\ &= \mu(e_{1:t}||a'_{1:t})\prod_{i=1}^{t}\sum_{a_i} f(a'_i|a'_{<i}a_i,e_{<i})[\![a_i=\pi^{\text{AIXI}}(\text{æ}^*_{<i})]\!] \\ &= \mu(e_{1:t}||a'_{1:t})\prod_{i=1}^{t} f(a'_i|a'_{<i}a^*(e_{<i}),e_{<i})\end{aligned} \tag{21}$$

Taking the sum of $a'_{1:t}$, we obtain $(\mu\circ f)(e_{1:t}||a^*(e)_{1:t})$ by Eq. (19). Because π^{AIXI} is deterministic, this is equal to $\text{dual}(\mu\circ f,\pi^{\text{AIXI}})(\text{æ}^*_{1:t})$ and we are finished. ■

As a clarifying special case, the percepts may contain an image of the corrupted action that was actually taken. For instance, the agent's cameras may record the movements of its actuators, or perhaps the agent is equipped with memory of its past corrupted action choices which is simply appended to the percept stream. Slightly informally, we can write $a'=\text{im}^{-1}(e)$. Then the sum can be eliminated, yielding

Corollary 14 *In the setting of Theorem 13, if μ is chosen so that the percept bits e contain $im(a)$ and $a'=im^{-1}(e)$ uniquely identifies the corrupted actions which are actually taken, then*

$$\text{dual}(\mu,f\circ\pi^{HAIXI})(\text{æ}'_{1:t})=\text{dual}(\mu\circ f,\pi^{AIXI})(\text{æ}^*_{1:t}) \tag{22}$$

In this sense, hardened AIXI learns to respond to action corruption like AIXI learns to respond to any feature of its environment.[8]

Learnable action corruption. Assuming further that f is a computable function and the true environment is an l.s.c. chronological semimeasure, the "effective" environment $\mu\circ f$ is realizable for AIXI's hypothesis space. Therefore, Corollary 14 tells us that hardened AIXI can learn to anticipate and correct for action corruption by f. However, it may not be reasonable to assume that f is computable. For instance, there is always a chance that some precise bit flip inverts (hardened) AIXI's utility function, causing it to minimize instead of maximize future discounted rewards. Such a policy is not computable (even with access to a^*_t). In general, hardened AIXI is not designed to reason about other agents of its intelligence (and computability) level, which may include "non-destructively" corrupted versions of itself. Reflective oracles provide the appropriate framework for this type of mutual reasoning between equals.

[8] Wyeth has argued that this may solve the "anvil problem": `https://www.lesswrong.com/posts/WECqiLtQiisqWvhim/free-will-and-dodging-anvils-aixi-off-policy`

7 Discussion

The presumed failure of joint AIXI was the motivation for studying reflective AIXI, which neatly resolves most of the problems in this paper at the cost of taking an arbitrary fixed point. For instance, the reflective oracle version of joint AIXI faces no realizability problem. A reflective Self-AIXI is constructed in [WHLT25], and can just as easily be defined to use a joint action/percept distribution. In contrast, we might expect joint AIXI's learning to sometimes fail because the history is not sampled from its belief distribution. Though we prove in this paper that it fails to converge to the correct answer in a simple case, we assume adversarially selected action bits. For a deployed agent, the action bits would be selected by the policy π^{JAIXI} which may never produce these adversarial action sequences, so we do not know whether π^{JAIXI} learns to behave well in reasonable environments. On the other hand, we prove that normalizing the joint distribution allows learning deterministic environments. To match the results for AIXI, we would actually like (on-policy) *fast* convergence results against all l.s.c. chronological semimeasures. Both our positive and our negative results on joint AIXI leave interesting open problems.

The technical difficulties involved in establishing anything about joint AIXI justify Hutter's nontrivial choice to invent a universal mixture for l.s.c. *chronological* semimeasures as a basis for studying Cartesian agents. Hardened AIXI captures the advantages of this choice, since its belief distribution is still a chronological semimeasure, only introducing a method for tracking its own actions. In fact, hardened AIXI can be considered as extending AIXI's behavior off-policy in one natural way. Future work might investigate any limitations of hardened AIXI as an embedded agent. In particular, it would be interesting to understand the behavior of an approximation to hardened AIXI which gains computational resources over time. We conjecture that it should treat its past, less intelligent action choices as if they were corrupted, perhaps even motivating further self-improvement through resource acquisition.

Many embeddedness problems can be patched by making small modifications to the AIXI agent, which may provide a rough description of embedded agents in the real world. However, these patches seem less elegant than the unified theory of reflective oracles.

Acknowledgements. This work was supported in part by a grant from the Long-Term Future Fund (EA Funds - Cole Wyeth - 9/26/2023). Cole Wyeth would also like to thank his advisor Ming Li for both support and freedom to pursue these problems during his PhD. The utility-inverting example of action corruption is thanks to Daniel Herrmann.

References

[ACCM22] Samuel Allen Alexander, Michael Castaneda, Kevin Compher, and Oscar Martinez. Extending environments to measure self-reflection in reinforcement learning. *Journal of Artificial General Intelligence*, 13(1), 2022.

[BD62] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[Ben24] Michael Timothy Bennett. Computational Dualism and Objective Superintelligence. In Kristinn R. Thórisson, Peter Isaev, and Arash Sheikhlar, editors, *Artificial General Intelligence*, pages 22–32, Cham, 2024. Springer Nature Switzerland.

[CGMH23] Elliot Catt, Jordi Grau-Moya, Marcus Hutter, Matthew Aitchison, Tim Genewein, Grégoire Delétang, Kevin Li, and Joel Veness. Self-Predictive Universal AI. *Advances in Neural Information Processing Systems*, 36:27181–27198, December 2023.

[Dai09] Wei Dai. Towards a New Decision Theory. August 2009.

[DG20] Abram Demski and Scott Garrabrant. Embedded Agency, October 2020. arXiv:1902.09469 [cs].

[ELH15] Tom Everitt, Jan Leike, and Marcus Hutter. Sequential extensions of causal and evidential decision theory, 2015.

[FST15] Benja Fallenstein, Nate Soares, and Jessica Taylor. Reflective Variants of Solomonoff Induction and AIXI. In Jordi Bieger, Ben Goertzel, and Alexey Potapov, editors, *Artificial General Intelligence*, pages 60–69, Cham, 2015. Springer International Publishing.

[FTC15] Benja Fallenstein, Jessica Taylor, and Paul F. Christiano. Reflective Oracles: A Foundation for Classical Game Theory, August 2015. arXiv:1508.04145 [cs].

[HM07] Marcus Hutter and Andrej Muchnik. On semimeasures predicting Martin-Löf random sequences. *Theoretical Computer Science*, 382(3):247–261, September 2007.

[Hut00] Marcus Hutter. A Theory of Universal Artificial Intelligence based on Algorithmic Complexity, April 2000. arXiv:cs/0004001.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence*. Texts in Theoretical Computer Science An EATCS Series. Springer, Berlin, Heidelberg, 2005.

[Kos21] Vanessa Kosoy. Infra-Bayesian physicalism: a formal theory of naturalized induction, November 2021.

[LH18] Jan Leike and Marcus Hutter. On the computability of Solomonoff induction and AIXI. *Theoretical Computer Science*, 716:28–49, March 2018.

[LHG11] Tor Lattimore, Marcus Hutter, and Vaibhav Gavane. Universal Prediction of Selected Bits. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 262–276, Berlin, Heidelberg, 2011. Springer.

[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, pages 427–436, Arlington, Virginia, USA, June 2016. AUAI Press.

[OR12] Laurent Orseau and Mark Ring. Space-Time Embedded Intelligence. In Joscha Bach, Ben Goertzel, and Matthew Iklé, editors, *Artificial General Intelligence*, pages 209–218, Berlin, Heidelberg, 2012. Springer.

[WH25] Cole Wyeth and Marcus Hutter. Value under ignorance in universal artificial intelligence. In *Artificial General Intelligence* (forthcoming) 2025.

[WHLT25] Cole Wyeth, Marcus Hutter, Jan Leike, and Jessica Taylor. Limit-computable grains of truth for arbitrary computable extensive-form (un)known games. (under review) 2025.

A List of Notation

Symbol	**Explanation**
$i,j,k \in \mathbb{N}$	index for natural numbers
$[\![\text{bool}]\!]$	$=1$ if bool=True, $=0$ if bool=False
$\lvert\mathcal{X}\rvert \equiv \#\mathcal{X}$	size of set $\mathcal{X}$.
$\mathcal{A}$	finite alphabet like $\{a,...,z\}$ or ASCII or $\{0,1\}$.
$\mathcal{A}^*,\mathcal{A}^\infty$	set of all (finite, infinite) strings over alphabet $\mathcal{A}$
$\Delta\mathcal{A}$	probability simplex over $\mathcal{A}$
$\mathbb{B}$	the binary set $\{0,1\}$
$A \times B$	Cartesian product of sets A and B
$t,n \in \mathbb{N}$	time index, e.g. $x_{1:n}$ or $x_{<t}$
$x_{1:n} \in \mathcal{A}^n$	string of length n
$x_{<t} \in \mathcal{A}^{t-1}$	string of length $t-1$
$l(x) \in \mathbb{N}$	the length of $x \in \mathcal{A}^*$
ϵ	empty string
■	end of proof
$\mathbb{R},\mathbb{N},...$	set of real,natural numbers
$\nu,\nu^\cdot$	semimeasure, chronological semimeasure
$\mu,\mu^\cdot$	true distribution/environment
$\xi,\xi^\cdot$	mixture of distributions/environments
ξ_U	Solomonoff's universal mixture distribution
ξ^{AI}	Hutter's universal mixture environment