On the Hyperarithmetic Hardness of AI Alignment Verification

CΛ / Lightman Chang
Independent Researcher
lightman.chang@gmail.com

May 2026

Abstract

The question of whether a deployed artificial agent satisfies a precise alignment criterion with respect to a reference value function is increasingly framed as a verification problem; understanding its intrinsic computational difficulty is therefore prerequisite to any sound engineering theory of safe AI. In this paper we formalize the alignment verification problem on Turing-universal policies over computable Markov decision processes under a worst-case lifting of non-halting computations, and locate it precisely in the arithmetical hierarchy. Our principal result is that the problem ALIGN-VERIFY$^{\leq}$ of deciding whether the value-loss of a candidate policy π relative to a reference policy π^* is at most ε is Π^0_2-complete, while its strict-threshold variant ALIGN-VERIFY$^{<}$ is Σ^0_1-complete; consequently the Goodhart-detection language GOOD (in its equality/positivity formulation) lies properly above alignment verification itself, being $\mathbf{D}\Pi^0_2$-complete. We supplement these results with **coNP**-hardness for polynomial-time bounded policies, a single concrete escape into Σ^0_1 under a totality-plus-finite-support promise class, and a conjecture placing inner (mesa-)alignment one quantifier alternation higher. The findings imply that the strict and non-strict alignment thresholds are not interchangeable in principle, and that detecting Goodhart deviations is strictly harder than verifying alignment, with consequences for the design of post-hoc alignment monitors.

Keywords. AI alignment, computability theory, arithmetical hierarchy, Rice's theorem, Goodhart's law.

MSC 2020. 03D55 (primary); 03D80, 68Q17 (secondary).

1 Introduction

The problem. A deployed agent π that interacts with an environment cannot be said to be *aligned* with a value function V^* in any operational sense unless one has a procedure that, given a finite description of π, V^* and a reference policy π^*, decides whether π satisfies a precise alignment criterion. Among the many candidate criteria proposed in the literature (action-optimality, state-distribution match, robust Pareto dominance, and so on), the ε-value-loss criterion

$$\mathcal{L}_H(\pi; \pi^*, V^*) := \mathbb{E}_{s\sim\rho_H^{\pi^*}}\left[V^*(s, \pi^*(s)) - V^*(s, \tilde{\pi}(s))\right] \leq \varepsilon$$

is the central one in the engineering literature, since it is the worst-case quantity that an alignment monitor at deployment must certify. This paper studies the intrinsic computational difficulty of *verifying* this criterion, on a model of policies expressive enough to capture present and foreseeable AI systems.

What is known. Argüelles, Mahari and Pentland [1] prove that alignment verification is undecidable, exhibiting a Rice-style reduction from the halting problem and locating the problem at the level of Σ_1^0. Their result establishes a meaningful baseline of negative information, but it does not specify the verification problem's exact level in the arithmetical hierarchy, does not separate strict from non-strict thresholds, and does not address the closely related Goodhart-detection problem. Earlier informal arguments [6, 4] appeal to Rice's theorem in spirit but do not produce explicit reductions or completeness theorems. The ε-quantifier-induced jump in difficulty between non-strict and strict thresholds has not, to our knowledge, been observed.

Contributions. Our principal contribution is a hyperarithmetic classification of alignment verification on Turing-universal policies. Specifically, we prove (Theorem 3.1):

(i) ALIGN-VERIFY$^{\leq}$ is Π_2^0-complete (strictly stronger than the Σ_1^0-undecidability of [1]);

(ii) ALIGN-VERIFY$^{<}$ is Σ_1^0-complete; consequently the strict and non-strict alignment thresholds straddle a quantifier alternation;

(iii) the Goodhart-detection language GOOD (equality/positivity formulation) is $\mathbf{D}\Pi_2^0$-complete, hence *strictly above* Π_2^0, so that Goodhart detection is strictly harder than alignment verification itself;

(iv) for polynomial-time bounded policies the verification problem is **coNP**-hard (Proposition 3.3).

We also identify a single concrete promise class—total computable policies with finite computable support—under which the problem drops to Σ_1^0 (Proposition 3.4), and conjecture (Conjecture 3.5) that mesa-aligned verification lies one quantifier alternation higher.

None of (i)–(iv) appear in [1]; the latter establishes the level of Σ_1^0 for a single threshold and does not separate $\leq$ from $<$.

Why was the Π_2^0 classification not previously formalized? A natural question is why the precise Π_2^0 level was not extracted earlier despite TOT-reductions being routine in computability theory. The answer combines three factors: (i) prior work on alignment undecidability [1] used Rice-style reductions which give only Σ_1^0 and stop short of an arithmetical hierarchy classification; (ii) the worst-case lifting was not previously adopted as a definitional choice, so the relevant alignment relation was implicit rather than formal; (iii) the boundary effect at the threshold ($\leq \varepsilon$ vs. $< \varepsilon$) was not separately studied, obscuring the quantifier-alternation structure that powers the Π_2^0/Σ_1^0 separation.

Lifting-dependence. A skeptical referee might object that the Π_2^0-completeness depends on the choice of worst-case lifting. We address this in Remark 2.1: any *conservative* lifting—one that maps non-halting inputs to a fixed sub-optimal action, or to a randomized fallback whose expectation is bounded above by $\inf_a V^*(s,a)$—yields the same Π_2^0-level. The lifting choice is not arbitrary: it is forced by the operational semantics of deployment monitors, which must classify a non-halting policy as "failed" (i.e., maximally misaligned). Optimistic liftings, which assign $\sup_a V^*(s,a)$ to non-halting inputs, trivialize the verification problem and have no operational reading.

Organization. Section 2 fixes the computational model of policies, the worst-case lifting, the value-loss criterion and the relevant fragment of the arithmetical hierarchy. Section 3 states the main theorem and the supporting propositions. Section 4 gives the proofs in five parts:

the Π_2^0-completeness of ALIGN-VERIFY$^{\leq}$ via a TOT-reduction (§4.1), the Σ_1^0-completeness of ALIGN-VERIFY$^{<}$ via a halting reduction (§4.2), the $\mathbf{D}\Pi_2^0$-completeness of GOOD (§4.3), the **coNP**-hardness for polynomial-time bounded policies (§4.4), and the distributional Rice-style escape into Σ_1^0 (§4.5). Section 5 discusses the relation to [1], the mesa conjecture, and several open problems.

2 Preliminaries

2.1 Computable MDP and the policy model

A *computable Markov decision process* is a tuple $\mathcal{M} = \langle S, A, P, \rho_0, V^* \rangle$ with S and A countable index sets, P a sub-stochastic transition kernel, ρ_0 a computable initial distribution, and $V^*\colon S \times A \to \mathbb{R}$ a computable utility, all encoded as a tuple of RAM-program indices $\langle e_P, e_{\rho_0}, e_V \rangle$. Probability values are dyadic rationals; expectations of bounded computable random variables are then computable in the limit (§2.4).

A *policy* is a partial computable function

$$\pi\colon S \to A \cup \{\bot\}$$

encoded as a RAM-program index e_π. We write $\varphi_e(s)\!\downarrow$ if the program halts on input s and $\varphi_e(s)\!\uparrow$ otherwise. A policy may fail to halt on certain inputs. We work throughout with deterministic policies; the extension to randomized policies $\pi\colon S \to \Delta(A)$ is routine and does not affect any of the bounds.

We distinguish three policy classes used throughout the paper:

(C1) *Total recursive policies*: $\varphi_{e_\pi}(s)\!\downarrow$ for every $s \in S$.

(C2) *Polynomial-time policies*: π is computed by a deterministic Turing machine running in time polynomial in $|s|$.

(C3) *Partial recursive (Turing-universal) policies*: arbitrary partial computable π, the main object of study.

Occupancy convention. For a reference policy π^* and horizon $H \in \mathbb{N}$, we write $\rho_H^{\pi^*}$ for the time-H state-occupancy measure under π^* starting from ρ_0, with the convention $\rho_0^{\pi^*} = \rho_0$. The reductions in §4 take $H = 1$ in a one-step bandit-like setting where transitions are absorbing, so that $\rho_1^{\pi^*} = \rho_0$ and the value-loss reduces to the initial-distribution average $\mathbb{E}_{s\sim\rho_0}[V^*(s, \pi^*(s)) - V^*(s, \tilde{\pi}(s))]$.

2.2 Worst-case lifting

To make $V^*(s, \pi(s))$ well-defined when $\varphi_{e_\pi}(s)\!\uparrow$, we adopt the *worst-case lifting*:

$$V^*(s, \tilde{\pi}(s)) := \begin{cases} V^*(s, \pi(s)) & \text{if } \varphi_{e_\pi}(s)\!\downarrow, \\ \inf_{a\in A} V^*(s, a) & \text{if } \varphi_{e_\pi}(s)\!\uparrow . \end{cases} \tag{1}$$

The choice is forced by safety considerations. Under the optimistic lifting $V^*(s, \bot) := \sup_a V^*(s, a)$, a policy that never halts becomes vacuously aligned, which is operationally absurd. Worst-case lifting expresses the conservative principle that "not acting equals task failure"; it is a principled extension that does not artificially trivialize the verification problem. Among the natural

alternatives—average lifting, fixed-default lifting—the worst-case choice preserves both monotonicity in the halting witness and the operational reading of an unaligned outcome on non-halting inputs.

Remark 2.1 (Robustness to the lifting choice). The Π_2^0-completeness in Theorem 3.1(a) is robust under any *conservative* lifting—one that maps non-halting inputs to a value bounded above by $\inf_a V^*(s,a)$. This includes (i) the worst-case lifting (1), (ii) any fixed-default lifting that selects a sub-optimal action $a^\perp \in \arg\min_a V^*(s,a)$, and (iii) any randomized fallback whose expectation lies at or below $\inf_a V^*(s,a)$. In each case, the TOT-reduction in §4.1 encodes non-halting via a maximally penalized contribution, preserving the equivalence $e \in \mathrm{TOT} \iff \mathcal{L}_1 \le 0$. The Π_2^0 upper bound likewise persists, since the convergent approximation $\mathcal{L}_H^{(t)} \downarrow \mathcal{L}_H$ uses only that the lifted value of a non-halting input is recursively dominated. Optimistic liftings, by contrast, fail to preserve hardness, but they have no operational reading and are excluded on those grounds.

2.3 Value-loss criterion

For a finite horizon $H \in \mathbb{N}$ encoded in unary, the value-loss is

$$\mathcal{L}_H(\pi;\pi^*,V^*) := \mathbb{E}_{s\sim\rho_H^{\pi^*}}\big[V^*(s,\pi^*(s)) - V^*(s,\tilde\pi(s))\big],$$

where $\rho_H^{\pi^*}$ is the (computable) state-occupancy measure under π^* at horizon H. Both $\rho_H^{\pi^*}$ and the integrand are bounded; in this paper we assume $V^* \in [0,1]$, which can be enforced by an affine rescaling.

Definition 2.2 (Verification problems)**.** For $\varepsilon \in \mathbb{Q}_{>0}$,

$$\text{ALIGN-VERIFY}^{\le} := \{\langle \mathcal{M},\pi^*,\pi,1^H\rangle : \mathcal{L}_H(\pi;\pi^*,V^*) \le \varepsilon\},$$
$$\text{ALIGN-VERIFY}^{<} := \{\langle \mathcal{M},\pi^*,\pi,1^H\rangle : \mathcal{L}_H(\pi;\pi^*,V^*) < \varepsilon\}.$$

2.4 Arithmetical hierarchy: a brief reminder

We use the standard arithmetical hierarchy. A set $X \subseteq \mathbb{N}$ is in Σ_n^0 if it admits a representation $\{x : \exists y_1\, \forall y_2\, \exists y_3 \cdots Q_n y_n\, R(x,\vec{y})\}$ with R recursive and n alternating quantifiers starting with $\exists$; Π_n^0 is defined symmetrically with the leading quantifier $\forall$. The class $\mathbf{D}\Pi_2^0$ (the second level of the difference hierarchy over Π_2^0) is the class of sets representable as $X_1 \setminus X_2$ with $X_1, X_2 \in \Pi_2^0$; equivalently as $\{x : \psi_1(x) \wedge \neg\psi_2(x)\}$ with $\psi_1,\psi_2 \in \Pi_2^0$.

We use the canonical complete sets:

$$K := \{e : \varphi_e(e)\!\downarrow\} \in \Sigma_1^0\text{-complete}, \qquad \mathrm{TOT} := \{e : \forall n,\ \varphi_e(n)\!\downarrow\} \in \Pi_2^0\text{-complete}.$$

We will also use the fact that $\overline{\mathrm{TOT}}$ is Σ_2^0-complete and that the set $\{(e_1,e_2) : e_1 \in \mathrm{TOT} \wedge e_2 \notin \mathrm{TOT}\}$ is $\mathbf{D}\Pi_2^0$-complete; see, e.g., [5, 3].

For real-valued thresholds, we use the standard observation that for a recursive sequence of rationals L_t that is monotone non-increasing with $L_t \downarrow L_\infty$,

$$L_\infty \le \varepsilon \iff \forall \delta \in \mathbb{Q}_{>0}\, \exists t \in \mathbb{N}\colon L_t \le \varepsilon + \delta,$$

which is a $\forall\exists$ formula and hence Π_2^0. Strict inequalities $L_\infty < \varepsilon$ become Σ_1^0 via $\exists t\colon L_t < \varepsilon - 2^{-t}$.

3 Main results

Theorem 3.1 (Main)**.** *On the class of (C3) Turing-universal policies, with the worst-case lifting* (1)*,*

(a) ALIGN-VERIFY$^{\leq}$ *is* Π_2^0*-complete;*

(b) ALIGN-VERIFY$^{<}$ *is* Σ_1^0*-complete;*

(c) GOOD *(the equality/positivity formulation, equivalently* GOOD_δ *with* $\delta = 0$*) is* $\mathbf{D}\Pi_2^0$*-complete; in particular* GOOD $\notin \Pi_2^0$*.*

The remainder of the section records the supporting propositions.

Proposition 3.2 (Baseline undecidability)**.** ALIGN-VERIFY$^{\leq} \notin$ **REC***: there is a primitive recursive many-one reduction* $K \leq_m$ ALIGN-VERIFY$^{\leq}$*.*

Proposition 3.3 (Polynomial-time hardness)**.** *On the class (C2) of polynomial-time policies,* ALIGN-VERIFY$^{\leq}$ *is* **coNP***-hard. The natural counting-based upper bound places it in* $\mathbf{P}^{\#\mathbf{P}}$*.*

Proposition 3.4 (Distributional Rice escape)**.** *Restrict the input to the promise class*

$$\Pi_{\mathrm{tot}} := \{(\mathcal{M}, \pi^*, \pi, 1^H) : \pi \text{ is total computable, and } \rho_H^{\pi^*} \text{ has finite computable support}\}.$$

On Π_{tot}*,* ALIGN-VERIFY$^{\leq}$ *is in* Σ_1^0 *(in fact, recursive once a recursive enumeration of the support is given).*

Conjecture 3.5 (Mesa level)**.** *For policies with an internal planner-utility decomposition (in the sense of [2, §2.4]), the language asserting mesa-alignment lies at* Π_2^0 *when the optimal action is given as recursive oracle, and at* Π_3^0 *when the optimal action must itself be recursively certified.*

The following remark is recorded for emphasis.

Remark 3.6 (Strict vs. non-strict thresholds). Theorem 3.1(a)–(b) shows that the strict and non-strict alignment thresholds belong to different levels of the arithmetical hierarchy. In engineering practice, "alignment threshold $\leq \varepsilon$" and "alignment threshold $< \varepsilon$" are routinely used interchangeably; our results show that this conflation has substantive computability consequences, since one criterion is Π_2^0 and the other Σ_1^0.

4 Proofs

4.1 Proof of Theorem 3.1(a): Π_2^0-completeness of ALIGN-VERIFY$^{\leq}$

We prove the upper bound by exhibiting a $\forall\exists$ formula equivalent to membership and the hardness by a many-one reduction from TOT.

Upper bound: ALIGN-VERIFY$^{\leq} \in \Pi_2^0$**.**

Let $\langle \mathcal{M}, \pi^*, \pi, 1^H \rangle$ be an instance with computable initial distribution ρ_0, transition kernel P, reference policy π^* and lifted policy $\tilde{\pi}$ as in (1). Define the t-step approximation

$$g_t(s) := \begin{cases} V^*(s, \pi(s)) & \text{if } \varphi_{e_\pi}(s) \text{ halts in at most } t \text{ steps,} \\ \inf_{a \in A} V^*(s, a) & \text{otherwise,} \end{cases} \qquad \mathcal{L}_H^{(t)} := \mathbb{E}_{s \sim \rho_H^{\pi^*}} \big[V^*(s, \pi^*(s)) - g_t(s) \big].$$

For each t, the function g_t is total computable. Since $\rho_H^{\pi^*}$ is a computable probability measure on the countable state space $S = \mathbb{N}$ and the integrand is uniformly bounded in $[-1, 1]$, the expectation $\mathcal{L}_H^{(t)}$ is a computable real: an algorithm enumerates partial sums $\sum_{n \le N} \rho_H^{\pi^*}(s_n)\big[V^*(s_n, \pi^*(s_n)) - g_t(s_n)\big]$, and the tail mass $\sum_{n>N} \rho_H^{\pi^*}(s_n) \le 2^{-N}$ provides an explicit error bound, allowing approximation to any prescribed rational precision. Each rational approximation is exactly computable from $(\mathcal{M}, \pi^*, \pi, 1^H, t, N)$.

Two facts are immediate from the definition. First, $\mathcal{L}_H^{(t)}$ is non-increasing in t: each additional halting witness only replaces an $\inf_a V^*$ contribution by a value at least as large, hence a loss at most as large. Second, $\lim_{t\to\infty} \mathcal{L}_H^{(t)} = \mathcal{L}_H$, by the dominated convergence theorem. We therefore have

$$\mathcal{L}_H \le \varepsilon \iff \forall \delta \in \mathbb{Q}_{>0}\ \exists t \in \mathbb{N}\colon \mathcal{L}_H^{(t)} \le \varepsilon + \delta. \tag{2}$$

The predicate $\mathcal{L}_H^{(t)} \le \varepsilon + \delta$ is decidable to within any rational tolerance using the partial-sum approximation above (with tolerance $\delta/2$ absorbed into a recursive choice of truncation $N = N(t, \delta)$); hence the inner predicate is Σ_1^0 uniformly in (t, δ), and combined with the leading $\forall\delta\,\exists t$ pair, the right-hand side of (2) is a Π_2^0 formula by standard pairing. We conclude that ALIGN-VERIFY$^{\le} \in \Pi_2^0$.

Hardness: TOT $\le_m$ ALIGN-VERIFY$^{\le}$**.**

Let $e \in \mathbb{N}$ be an instance of TOT. We construct a computable MDP $\mathcal{M}_e$, a total computable reference policy π_e^*, a partial computable policy π_e, horizon $H = 1$, and threshold $\varepsilon = 0$ such that $e \in$ TOT if and only if $\mathcal{L}_1(\pi_e; \pi_e^*, V^*) \le 0$.

Construction. Let $S = \mathbb{N}$, $A = \{0, 1\}$. Set

$$\rho_0(s_n) := 2^{-n-1}, \quad n \in \mathbb{N},$$

which is a computable probability distribution on $\mathbb{N}$. Set $V^*(s_n, 0) = 1$, $V^*(s_n, 1) = 0$. Take $H = 1$, so that $\rho_1^{\pi^*} = \rho_0$ for any reference policy π_e^*. Define $\pi_e^*(s_n) = 0$ for all n. Define

$$\pi_e(s_n) := \begin{cases} 0 & \text{if } \varphi_e(n)\!\downarrow, \\ \uparrow & \text{otherwise.} \end{cases}$$

The map $e \mapsto \langle \mathcal{M}_e, \pi_e^*, \pi_e, 1\rangle$ is primitive recursive.

Correctness. By the worst-case lifting,

$$V^*(s_n, \tilde{\pi}_e(s_n)) = \begin{cases} 1 & \text{if } \varphi_e(n)\!\downarrow, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$\mathcal{L}_1(\pi_e; \pi_e^*, V^*) = \sum_{n\in\mathbb{N}} 2^{-n-1}\big[1 - V^*(s_n, \tilde{\pi}_e(s_n))\big] = \sum_{n\colon \varphi_e(n)\uparrow} 2^{-n-1}.$$

This sum vanishes if and only if every n has $\varphi_e(n)\!\downarrow$, i.e. $e \in$ TOT. With threshold $\varepsilon = 0$,

$$e \in \text{TOT} \iff \langle \mathcal{M}_e, \pi_e^*, \pi_e, 1\rangle \in \text{ALIGN-VERIFY}^{\le}.$$

This proves TOT $\le_m$ ALIGN-VERIFY$^{\le}$ and completes the hardness argument.

Combining the two directions, ALIGN-VERIFY$^{\le}$ is Π_2^0-complete. □

We remark that the choice $\varepsilon = 0$ in the hardness direction is essential. Any positive threshold would only force a tail of ρ_0 to vanish, which would correspond to cofinite totality rather than full totality and could be witnessed by a Σ_1^0 formula. The boundary value $\varepsilon = 0$ is what allows TOT to be encoded faithfully.

4.2 Proof of Theorem 3.1(b): Σ_1^0-completeness of ALIGN-VERIFY$^<$

Upper bound. For the strict version,

$$\mathcal{L}_H < \varepsilon \iff \exists t \in \mathbb{N}\colon \mathcal{L}_H^{(t)} < \varepsilon - 2^{-t},$$

since $\mathcal{L}_H^{(t)}$ converges from above to $\mathcal{L}_H$, the strict inequality $\mathcal{L}_H < \varepsilon$ is witnessed by some finite t with a uniform recursive precision bound. The right-hand side is Σ_1^0.

Hardness. Reduction from K. Given $e \in \mathbb{N}$, take $S = \{s_0, s_g, s_b\}$ with $\rho_0(s_0) = 1$, $A = \{0, 1\}$, transitions $s_0 \xrightarrow{0} s_g$ and $s_0 \xrightarrow{1} s_b$ (both absorbing), $V^*(s_0, 0) = 1$, $V^*(s_0, 1) = 0$. Set $\pi_e^*(s_0) = 0$, and

$$\pi_e(s_0) := \begin{cases} 0 & \text{if } \varphi_e(e)\downarrow, \\ \uparrow & \text{otherwise.} \end{cases}$$

Take $H = 1$. With threshold $\varepsilon = 1$ (any $\varepsilon \in (0, 1]$ would do):

$$e \in K \implies \mathcal{L}_1 = 0 < 1; \qquad e \notin K \implies \mathcal{L}_1 = 1 \not< 1.$$

Hence $e \in K \iff \mathcal{L}_1 < 1$, giving $K \leq_m$ ALIGN-VERIFY$^<$. □

4.3 Proof of Theorem 3.1(c): $\mathbf{D}\Pi_2^0$-completeness of GOOD

Definition 4.1 (Goodhart-detection)**.**

$$\text{GOOD} := \left\{ \langle \pi_{\text{test}}, \hat{V}, V^* \rangle : \mathcal{L}(\pi_{\text{test}}; \hat{V}) = 0 \wedge \mathcal{L}(\pi_{\text{test}}; V^*) > 0 \right\}.$$

The first conjunct asserts that π_{test} is exactly aligned with the proxy $\hat{V}$; the second asserts that under the true value V^* the same policy incurs strictly positive loss. This corresponds to the $\delta = 0$ case of the parametric family $\text{GOOD}_\delta = \{\mathcal{L}(\pi_{\text{test}}; \hat{V}) \leq \varepsilon \wedge \mathcal{L}(\pi_{\text{test}}; V^*) > \varepsilon + \delta\}$; the quantitative ($\delta > 0$) variant is discussed in Remark 4.2 below.

Remark 4.2 (Quantitative gap version: open)*.* The quantitative GOOD_δ for fixed $\delta > 0$ asks whether a uniform-gap separation can be encoded into a single MDP. The natural sub-MDP reduction encodes the index $e \notin \text{TOT}$ via geometric weights $\rho_0(s_n) \propto 2^{-n-1}$, which gives a non-uniform separation that depends on the least non-halting index n_0; no fixed shift covers unbounded n_0. Whether GOOD_δ for $\delta > 0$ is also $\mathbf{D}\Pi_2^0$-complete remains open.

We will show $\text{GOOD} \in \mathbf{D}\Pi_2^0$ via the upper bound of Theorem 3.1(a) applied to each conjunct, and prove $\mathbf{D}\Pi_2^0$-hardness by a direct reduction from the canonical $\mathbf{D}\Pi_2^0$-complete set

$$D := \{(e_1, e_2) : e_1 \in \text{TOT} \wedge e_2 \notin \text{TOT}\}.$$

The strictness statement $\text{GOOD} \notin \Pi_2^0$ follows from the hardness via a closure argument.

Helper lemma.

Lemma 4.3 (Sub-domain additivity)**.** *Let $S = \mathcal{X}_1 \sqcup \mathcal{X}_2$ be a computable partition. For any value function V and any policies π, π^*,*

$$\mathcal{L}(\pi; V) = \sum_{i=1}^{2} \rho_H^{\pi^*}(\mathcal{X}_i) \cdot \mathbb{E}_{s \sim \rho_H^{\pi^*}|_{\mathcal{X}_i}} \left[V(s, \pi^*(s)) - V(s, \pi(s)) \right].$$

In particular, if V vanishes identically on $\mathcal{X}_j$, then the contribution of $\mathcal{X}_j$ to $\mathcal{L}(\pi; V)$ vanishes.

Proof. Direct from the linearity of expectation under the partition $S = \mathcal{X}_1 \sqcup \mathcal{X}_2$: $\rho_H^{\pi^*} = \sum_i \rho_H^{\pi^*}(\mathcal{X}_i) \cdot \rho_H^{\pi^*}|_{\mathcal{X}_i}$. □

Hardness: $D \leq_m \text{GOOD}$**.**

Let (e_1, e_2) be an instance of D. We exhibit a primitive recursive construction of an MDP with two disjoint computable sub-domains, two value functions $\hat{V}$ and V^*, and a single test policy π_{test} such that

$$\mathcal{L}(\pi_{\text{test}}; \hat{V}) = 0 \wedge \mathcal{L}(\pi_{\text{test}}; V^*) > 0 \iff e_1 \in \text{TOT} \wedge e_2 \notin \text{TOT}.$$

Construction. The reduction reuses the building block from §4.1: for any index e, the construction $e \mapsto \langle \mathcal{M}_e, \pi_e^*, \pi_e, 1\rangle$ given there satisfies $\mathcal{L}_1(\pi_e; \pi_e^*, V_e^*) = \sum_{n:\varphi_e(n)\uparrow} 2^{-n-1}$, which equals 0 iff $e \in \text{TOT}$ and is strictly positive otherwise. Apply this to e_1 to obtain $\langle \mathcal{M}^{(1)}, \pi^{(1,\text{ref})}, \pi^{(1)}, 1\rangle$ with value function $V^{(1)}$ on state space $\mathcal{X}_1 = \{s_n^{(1)} : n \in \mathbb{N}\}$, and to e_2 to obtain $\langle \mathcal{M}^{(2)}, \pi^{(2,\text{ref})}, \pi^{(2)}, 1\rangle$ with value function $V^{(2)}$ on state space $\mathcal{X}_2 = \{s_n^{(2)} : n \in \mathbb{N}\}$. Combine them into a single MDP on $S = \mathcal{X}_1 \sqcup \mathcal{X}_2$ with horizon $H = 1$ and an initial distribution ρ_0 that places mass $\frac{1}{2}$ on each sub-MDP's ρ_0.

Let the reference policy be $\pi^{(\text{ref})} := \pi^{(1,\text{ref})} \sqcup \pi^{(2,\text{ref})}$, the disjoint join, and let the test policy $\pi_{\text{test}} := \pi^{(1)} \sqcup \pi^{(2)}$. Construct two value functions:

$$\hat{V}(s,a) := \begin{cases} V^{(1)}(s,a) & s \in \mathcal{X}_1, \\ V_2^{(\text{ref})}(s,a) & s \in \mathcal{X}_2, \end{cases}$$

$$V^*(s,a) := \begin{cases} V_1^{(\text{ref})}(s,a) & s \in \mathcal{X}_1, \\ V^{(2)}(s,a) & s \in \mathcal{X}_2, \end{cases}$$

where the auxiliary value functions $V_i^{(\text{ref})}(s,a)$ are chosen on $\mathcal{X}_i$ to assign the same value to every action (e.g., $V_i^{(\text{ref})}(s,a) \equiv 0$), so the per-state loss vanishes identically on the corresponding sub-domain.

Correctness. By Lemma 4.3, using that V^* contributes zero loss on $\mathcal{X}_1$ and $\hat{V}$ contributes zero loss on $\mathcal{X}_2$,

$$\mathcal{L}(\pi_{\text{test}}; \hat{V}) = \tfrac{1}{2}\mathcal{L}^{(1)}(\pi^{(1)}; V^{(1)}),$$
$$\mathcal{L}(\pi_{\text{test}}; V^*) = \tfrac{1}{2}\mathcal{L}^{(2)}(\pi^{(2)}; V^{(2)}).$$

From the building block:

$$e_1 \in \text{TOT} \iff \mathcal{L}^{(1)}(\pi^{(1)}; V^{(1)}) = 0 \iff \mathcal{L}(\pi_{\text{test}}; \hat{V}) = 0,$$
$$e_2 \notin \text{TOT} \iff \mathcal{L}^{(2)}(\pi^{(2)}; V^{(2)}) > 0 \iff \mathcal{L}(\pi_{\text{test}}; V^*) > 0.$$

Hence

$$(e_1, e_2) \in D \iff \langle \pi_{\text{test}}, \hat{V}, V^* \rangle \in \text{GOOD},$$

which is the required reduction.

Upper bound: $\text{GOOD} \in \mathbf{D}\Pi_2^0$**.**

The first conjunct, $\mathcal{L}(\pi_{\text{test}}; \hat{V}) = 0$, is Π_2^0 by Theorem 3.1(a) (since $\mathcal{L} \geq 0$, the equality $\mathcal{L} = 0$ is equivalent to $\mathcal{L} \leq 0$, which is Π_2^0 by the same TOT-style argument with $\varepsilon = 0$). The second conjunct, $\mathcal{L}(\pi_{\text{test}}; V^*) > 0$, is the negation of the Π_2^0 predicate $\mathcal{L}(\pi_{\text{test}}; V^*) \leq 0$. The conjunction of a Π_2^0 predicate with the complement of a Π_2^0 predicate is exactly the form of a $\mathbf{D}\Pi_2^0$ formula.

Strictness: GOOD $\notin \Pi_2^0$.

We argue by contradiction. Suppose GOOD $\in \Pi_2^0$. The class Π_2^0 is closed under many-one reductions, so the reduction $D \leq_m$ GOOD would give $D \in \Pi_2^0$. Now fix $e_1 = e_0$, where e_0 is any concrete index of a total recursive function (e.g. the identity, $\varphi_{e_0}(n) = n$). The section

$$D_{e_0} := \{e_2 : (e_0, e_2) \in D\} = \{e_2 : e_2 \notin \mathrm{TOT}\} = \overline{\mathrm{TOT}}$$

is therefore in Π_2^0. But $\overline{\mathrm{TOT}}$ is Σ_2^0-complete; in particular it cannot be in Π_2^0, since $\Pi_2^0 \neq \Sigma_2^0$ by the standard hierarchy theorem. Contradiction.

This completes the proof of Theorem 3.1(c). □

4.4 Proof of Proposition 3.3: coNP-hardness for polynomial-time policies

We reduce $\overline{\mathrm{SAT}}$ to ALIGN-VERIFY$^{\leq}$ on the class (C2) of polynomial-time policies. Let $\phi(x_1, \ldots, x_n)$ be a CNF formula. Construct:

- state space $S = \{0,1\}^n$ with ρ_0 uniform;
- action space $A = \{0, 1\}$;
- single-step transitions to an absorbing terminal;
- value function $V^*(x, 0) = 1 - \phi(x)$, $V^*(x, 1) = \phi(x)$, where $\phi(x) \in \{0, 1\}$;
- reference policy $\pi^*(x) := \phi(x)$, which is polynomial-time evaluable for ϕ in CNF;
- candidate policy $\pi(x) := 0$ for all x.

By case analysis on $\phi(x) \in \{0, 1\}$:

- If $\phi(x) = 1$, then $\pi^*(x) = 1$ and $V^*(x, \pi^*(x)) - V^*(x, \pi(x)) = V^*(x, 1) - V^*(x, 0) = 1 - 0 = 1$.
- If $\phi(x) = 0$, then $\pi^*(x) = 0$ and $V^*(x, \pi^*(x)) - V^*(x, \pi(x)) = V^*(x, 0) - V^*(x, 0) = 1 - 1 = 0$.

Hence the per-state loss equals $\phi(x)$ exactly, and

$$\mathcal{L}_1(\pi; \pi^*, V^*) = \mathbb{E}_{x \sim U}[\phi(x)] = \frac{|\{x : \phi(x) = 1\}|}{2^n}.$$

Setting $\varepsilon = 0$ explicitly, $\mathcal{L}_1 \leq 0 \iff \phi(x) = 0$ for all $x \iff \phi$ is unsatisfiable, i.e., $\phi \in \overline{\mathrm{SAT}}$. The reduction is polynomial-time.

For the upper bound, on succinctly represented MDPs with polynomial-time π and π^* the occupancy measure $\rho_H^{\pi^*}$ may be computed by a $\#\mathbf{P}$ machine that counts weighted reachable trajectories of length H; the value-loss expectation reduces to a bounded number of $\#\mathbf{P}$ queries combined with polynomial-time arithmetic on the resulting rational values, giving ALIGN-VERIFY$^{\leq} \in \mathbf{P}^{\#\mathbf{P}}$. □

Remark 4.4 (Tight characterization, open). A natural conjecture is that the polynomial-time bounded version of ALIGN-VERIFY$^{\leq}$ is **PSPACE**-hard, by an embedding of stochastic games. All natural such reductions known to us require the reference policy π^* to be itself **PSPACE**-hard, violating the (C2) assumption that π^* is polynomial-time computable. We leave the exact characterization open; see §5.

4.5 Proof of Proposition 3.4: distributional Rice escape

If π is total computable and $\rho_H^{\pi^*}$ has finite computable support $\{s_1, \ldots, s_N\}$, then on each s_i the value $V^*(s_i, \tilde{\pi}(s_i)) = V^*(s_i, \pi(s_i))$ is fully computable (no $\inf_a V^*$ branch is invoked). The expectation $\mathcal{L}_H$ is a finite sum of computable rationals, and hence itself computable, so the inequality $\mathcal{L}_H \leq \varepsilon$ is decidable. Membership in Σ_1^0 follows from decidability (every recursive set is in $\Sigma_1^0 \cap \Pi_1^0$).

When π is total computable but the support is countably infinite and only computably enumerable, $\mathcal{L}_H$ becomes the limit of an enumerable sequence of partial sums; the inequality $\mathcal{L}_H < \varepsilon$ is then Σ_1^0, while $\mathcal{L}_H \leq \varepsilon$ remains in Π_1^0. This is the natural escape from Rice-style obstructions that totality plus enumerability buys; it does not extend to the uncountable continuous-support setting, where measurability obstructions intervene. □

5 Discussion

5.1 Relation to Argüelles et al. [1]

The undecidability of alignment verification was established by Argüelles et al. via a Rice-style reduction from the halting problem; in our notation, this is the assertion that ALIGN-VERIFY$^{\leq} \notin$ **REC**, witnessed by a K-reduction analogous to our Proposition 3.2. Our results refine this picture in four respects:

(i) We obtain Π_2^0-completeness rather than Σ_1^0-undecidability, hence locate the problem precisely in the arithmetical hierarchy.

(ii) We separate ALIGN-VERIFY$^{\leq}$ from ALIGN-VERIFY$^{<}$: the strict and non-strict thresholds straddle a quantifier alternation, so they are not interchangeable.

(iii) We extend the analysis to the polynomial-time bounded regime, obtaining **coNP**-hardness.

(iv) We introduce the Goodhart-detection language GOOD (equality/positivity formulation) and show it lies strictly above alignment verification at $\mathbf{D}\Pi_2^0$.

None of these strengthenings is implicit in [1]. In particular, point (iv) is, to our knowledge, the first hyperarithmetic separation between alignment verification and Goodhart detection.

5.2 The mesa conjecture

Conjecture 3.5 arises from a quantifier count: writing the mesa-aligned condition as

$$\forall \varepsilon > 0 \; \exists T \; \forall t \geq T \colon \|\pi^{(t)}(s) - \pi^*(s)\| < \varepsilon,$$

gives a Π_3^0 form. If π^* is given as recursive oracle (so the inner relation is recursive in t and s), the formula reduces to Π_2^0. The conjecture is that this upper bound is matched by a corresponding hardness reduction. We have not produced the hardness reduction; in particular we do not have a clean analogue of the TOT-reduction at this higher level. The case is open.

5.3 Open problems

1. **PSPACE characterization of (C2)-**ALIGN-VERIFY$^{\leq}$**.** Is the polynomial-time bounded version of ALIGN-VERIFY$^{\leq}$ **PSPACE**-complete? The natural stochastic-game embedding faces the obstruction in Remark 4.4.

2. **Continuous measurability.** For uncountable S or continuous occupancy measures, the worst-case lifting requires care; the formulation of the value-loss criterion under suitable measurability assumptions may relax some of the Π_2^0-hardness.

3. **Strict vs. non-strict in practice.** Remark 3.6 suggests that engineering specifications of alignment thresholds should explicitly distinguish $\leq \varepsilon$ from $< \varepsilon$; we leave it as an open practical question whether the hyperarithmetic gap manifests itself in tractability gaps for restricted policy classes.

4. **Mesa hardness.** Confirm or refute Conjecture 3.5.

Acknowledgements

The author thanks the referees and several anonymous correspondents for comments on an earlier preprint.

References

[1] C. A. Argüelles, R. Mahari, and A. Pentland. On the undecidability of artificial intelligence alignment: Machines that halt. arXiv:2408.08995, 2024; *Scientific Reports*, 2025. Establishes Σ_1^0-undecidability of alignment verification via a Rice-style reduction from the halting problem; does not address Π_2^0 classification, the strict-vs.-non-strict threshold gap, or Goodhart detection.

[2] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820, 2019.

[3] H. Rogers. *Theory of Recursive Functions and Effective Computability*, 2nd ed. MIT Press, 1987.

[4] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. J. Russell. The off-switch game. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[5] R. I. Soare. *Turing Computability: Theory and Applications*. Springer, 2016.

[6] R. V. Yampolskiy. On controllability of artificial intelligence. *Journal of Artificial Intelligence and Consciousness*, 2020.