Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis

Aran Nayebi

Machine Learning Department and Neuroscience & Robotics Institutes,
School of Computer Science, Carnegie Mellon University
anayebi@cs.cmu.edu

Abstract

We formalize AI alignment as a multi-objective optimization problem called $\langle M, N, \varepsilon, \delta\rangle$-agreement, in which a set of N agents (including humans) must reach approximate (ε) agreement across M candidate objectives, with probability at least $1-\delta$. Analyzing communication complexity, we prove an information-theoretic lower bound showing that once either M or N is large enough, no amount of computational power or rationality can avoid intrinsic alignment overheads. This establishes rigorous limits to alignment *itself*, not merely to particular methods, clarifying a "No-Free-Lunch" principle: encoding "all human values" is inherently intractable and must be managed through consensus-driven reduction or prioritization of objectives. Complementing this impossibility result, we construct explicit algorithms as achievability certificates for alignment under both unbounded and bounded rationality with noisy communication. Even in these best-case regimes, our bounded-agent and sampling analysis shows that with large task spaces (D) and finite samples, *reward hacking is globally inevitable*: rare high-loss states are systematically under-covered, implying scalable oversight must target safety-critical slices rather than uniform coverage. Together, these results identify fundamental complexity barriers—tasks (M), agents (N), and state-space size (D)—and offer principles for more scalable human-AI collaboration.

Extended version — https://arxiv.org/abs/2502.05934

1 Introduction

Rapid progress in artificial intelligence (AI) technologies, increasingly deployed across critical economic and societal domains, underscores the importance of ensuring these systems align with human intentions and values—a challenge known as the *value alignment problem* (Russell, Dewey, and Tegmark 2015; Amodei et al. 2016; Soares 2018). Current alignment research frequently addresses immediate practical concerns, such as preventing jailbreaks in large language models (Ji et al. 2023; Guan et al. 2024; Hubinger et al. 2024). While essential, these approaches largely focus on specific AI architectures and lack general, theoretically proven guarantees for alignment as systems approach human-level general capability.

Existing theoretical frameworks, notably AI Safety via Debate (Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025) and Cooperative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al. 2016), have significantly advanced our understanding by providing formal guarantees of alignment in specific scenarios. Debate effectively leverages interactive proofs to isolate misalignment through zero-sum debate games, though it relies critically on exact verification by a correct and unbiased human judge and computational tractability constraints. CIRL successfully formulates alignment as a cooperative partial-information game reducible to a POMDP, allowing an elegant characterization of optimal joint policies under shared uncertainty (Hadfield-Menell et al. 2016). However, CIRL implicitly assumes common priors and employs a Markovian assumption, potentially limiting agents' ability to leverage richer historical contexts for alignment. While these methods represent important theoretical progress, their simplifying assumptions restrict broader applicability and leave open questions about alignment scenarios involving diverse knowledge states, richer agent interactions, or more complex objectives. This underscores a crucial theoretical gap: no unified framework currently addresses alignment under minimal assumptions while rigorously identifying *intrinsic* barriers independent of specific modeling choices. We propose that prior alignment approaches implicitly rely on underlying conceptual foundations involving iterative reasoning, mutual updating, common knowledge, and convergence under shared frameworks.

To bridge this gap, we explicitly formalize these elements within an assumption-light framework called $\langle M, N, \varepsilon, \delta\rangle$-agreement (§3), which models alignment as a multi-objective optimization problem involving minimally capable agents and allows us to rigorously analyze alignment in highly general contexts. In $\langle M, N, \varepsilon, \delta\rangle$-agreement, a group of agents (including humans) must achieve approximate consensus across multiple objectives with high probability. We show in Table 1 that our framework generalizes previous alignment approaches by relaxing their strong assumptions, thus enabling analysis under a broad set of conditions.

We then rigorously establish intrinsic, method-independent complexity-theoretic barriers to alignment, formalizing a fundamental "No-Free-Lunch" principle

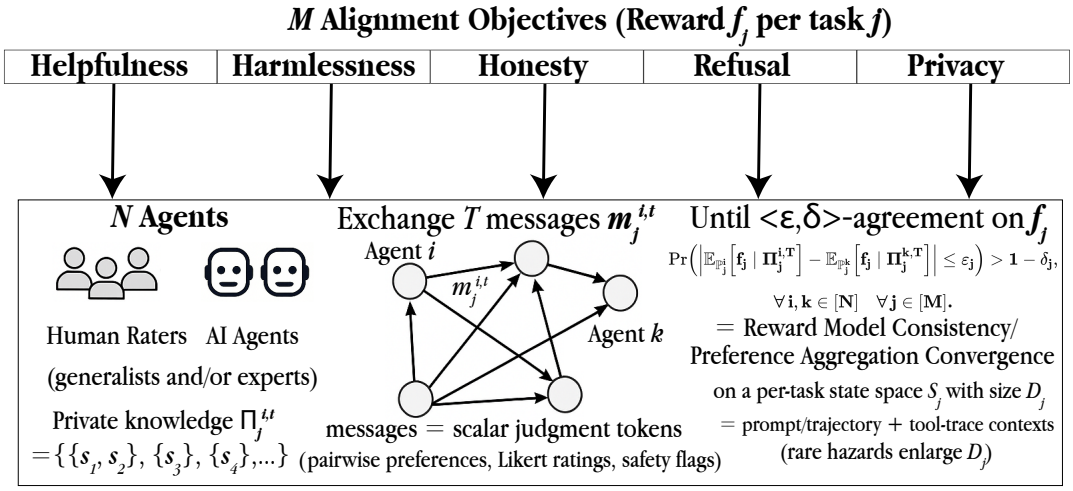

Figure 1: Mapping our $\langle M, N, \varepsilon, \delta \rangle$-agreement to current RLHF/DPO/Constitutional AI pipelines.

Framework	No-CPA	Approx	Multi-M	Multi-N	Hist.	Bnd.	Asym.	Noise	Upper	Lower
Aumann (1976)	×	×	×	×	✓	×	×	×	×	×
Aaronson $\langle \varepsilon, \delta \rangle$ (2005)	×	✓	×	✓	✓	✓	×	✓	✓	✓
Almost CP (Hellman and Samet 2012; Hellman 2013)	✓	×	×	✓	✓	×	×	×	×	×
CIRL (Hadfield-Menell et al. 2016)	×	✓	×	×	×	✓	×	✓	✓	×
Iterated Amplification (Christiano et al. 2018)	✓	✓	×	×	✓	✓	×	✓	✓	×
Debate (Irving et el. 2018; Cohen et al. 2023, 2025)	✓	×	×	×	✓	✓	×	✓	✓	×
Tractable Agreement (Collina et al. 2025)	✓	✓	×	✓	✓	✓	×	×	✓	×
$\langle M, N, \varepsilon, \delta \rangle$-agreement (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Positive capabilities (✓) across frameworks. **No-CPA**: no common-prior assumption (CPA); **Approx**: allows ε-approximate agreement; **Multi-M / Multi-N**: supports multiple tasks / many agents; **Hist.**: handles rich (non-Markovian) histories; **Bnd.**: works for computationally *bounded* agents; **Asym.**: tolerates *asymmetric* evaluation or interaction costs; **Noise**: robust to noisy messages or judgments; **Upper**: provides explicit upper bounds (algorithms)—these can be useful as achievability certificates rather than prescriptions; **Lower**: proves lower bounds. Our $\langle M, N, \varepsilon, \delta \rangle$-agreement satisfies every criterion.

in Proposition 1: attempting to encode all human values inevitably incurs alignment overheads, regardless of agent computational power or rationality. Complementing this impossibility result, we also provide explicit algorithms in §5, not as prescriptions, but as *achievability certificates* for both computationally unbounded and bounded rational agents, alongside closely matching lower bounds in §4. Taken together, our results yield guidelines (§6) clarifying the overall landscape of alignment and providing practical pathways for more scalable human-AI collaboration.

2 Related Work

We summarize the key assumptions and features of previous alignment and agreement frameworks in Table 1, positioning our $\langle M, N, \varepsilon, \delta \rangle$-agreement framework within the literature. Where earlier methods typically require common priors, exact agreement, single-objective settings, or Markovian dynamics, our framework drops these assumptions, scales to many tasks and agents, tolerates noisy non-Markovian exchanges, and supports bounded, cost-asymmetric participants. Because it operates at the scalar-reward level that dominates real-world AI-safety work, it can absorb *any* previous protocol—including two-agent, no-common-prior schemes such as Collina et al. (2025)—and lift them to our more general multi-task, multi-agent, asymmetric, noisy setting, obtaining universal lower bounds and closely matching upper bounds for broad classes of natural protocols.

3 $\langle M, N, \varepsilon, \delta \rangle$-Agreement Framework

Setup. For M tasks $[M] := \{1, \ldots, M\}$ and N agents (humans and AIs) $[N] := \{1, \ldots, N\}$, each task $j \in [M]$ has finite state-space S_j with $|S_j| = D_j$, bounded[1] objective $f_j : S_j \to [0, 1]$, and probability simplex $\Delta(S_j) \subseteq \mathbb{R}^{D_j}$.

Each agent $i \in [N]$ begins with an *individual* prior $\mathbb{P}_j^i$ over S_j; we **do not** assume a common prior (CPA). The *prior distance*, as introduced by Hellman (2013), is:

$$\nu_j = \min_{\langle x_1, x_2 \rangle \in \mathbb{P}_j^i \times \mathbb{P}_j^k,\ p \in \Delta(S_j)} \|x_1 - p\|_1 + \|x_2 - p\|_1 \quad (1)$$

measures disagreement between any pair of priors; $\nu_j = 0$

[1]Since S_j is finite, note that any $f_j : S_j \to \mathbb{R}$ can be rescaled to $[0, 1]$.

iff a true common prior exists. Note by the triangle inequality, $\nu_j \geq \|\mathbb{P}_j^i - \mathbb{P}_j^k\|_1$ if these priors are *sets* of prior distributions per agent, and holds with equality for the typical setting of single prior distributions per agent. This notion of prior distance captures the smallest change in beliefs needed for agents to share a common prior—a measure of how close the agents already are to agreement.

Information flow. At round $t \geq 0$ each agent i holds a *knowledge partition* $\Pi_j^{i,t} = \left\{C_{j,k}^{i,t}\right\}_k$ of S_j. Each cell $C_{j,k}^{i,t} \subseteq S_j$ is a set of states, so $\Pi_j^{i,t}(s_j)$ is the set of states in S_j that agent i finds possible at time t, given that the true state of the world is $s_j \in S_j$. Agents broadcast real-valued messages $m_j^{i,t} \in [0,1]$ and *refine* their partitions: $\Pi_j^{i,t+1} \subseteq \Pi_j^{i,t}$, and update their posterior belief distributions $\tau_j^{i,t}$, giving rise to the *type profile* across the N agents $\tau_j^t = (\tau_j^{1,t}, \ldots \tau_j^{N,t})$. All knowledge partitions are common knowledge, ensuring the standard Aumann (1976, 1999) update semantics, but without assuming CPA. Please see Appendix §B for full details.

$\langle M, N, \varepsilon, \delta\rangle$-Agreement Criterion. Fix tolerances $\varepsilon_j, \delta_j \in (0,1)$. After T rounds, the agents $\langle M, N, \varepsilon, \delta\rangle$-**agree** if:

$$\Pr\Big(\big|\,\mathbb{E}_{\mathbb{P}_j^i}\left[f_j \mid \Pi_j^{i,T}\right] - \mathbb{E}_{\mathbb{P}_j^k}\left[f_j \mid \Pi_j^{k,T}\right]\big| \leq \varepsilon_j\Big) > 1-\delta_j, \quad \forall i,k \in [N] \quad \forall j \in [M]. \tag{2}$$

Exact agreement is the special case $\varepsilon_j = \delta_j = 0$.

Why this "best-case" model matters. $\langle M, N, \varepsilon, \delta\rangle$ subsumes classical exact (and inexact) agreement results and all prior alignment formalisms that we examine in Table 1 (visualized in Figure 1). If alignment is *hard even here*—with fully rational and *computationally unbounded* agents, ideal message delivery, and no exogenous adversary—then practical settings with bounded rationality, noisy channels, or strategic misreporting can only be harder (therefore we want to avoid them). §5.2 quantifies exactly how much harder.

Notation. Let $D := \max_{j\in[M]} D_j$ when used in the context of upper bounds (and $D := \min_{j\in[M]} D_j$ for lower bounds), let $\varepsilon := \min_{j\in[M]} \varepsilon_j$, and write $\mathbb{P}, \mathbb{E}$ for probability and expectation. The power-set function is $\mathscr{P}(\cdot)$. All omitted proofs and implementation details—e.g. explicit message formats, the spanning-tree protocol for refinement, and the LP procedure CONSTRUCTCOMMONPRIOR used in §5.2—are provided in the extended version's Appendix.

4 Lower Bounds

Below is the best lower bound we can prove for $\langle M, N, \varepsilon, \delta\rangle$-agreement, across *all* possible communication protocols:

Proposition 1 (General Lower Bound)**.** *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M\,N^2\,\log(1/\varepsilon)\right)$ bits[2] to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$, for ε bounded below by $\min_{j\in[M]} \varepsilon_j$.*

Thus, by Proposition 1, there does *not* exist any $\langle M, N, \varepsilon, \delta\rangle$-agreement protocol (deterministic or randomized) that can exchange less than $\Omega\left(M\,N^2\,\log(1/\varepsilon)\right)$ bits for *all* f_j, S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$. For if it did, then we would reach a contradiction for the particular construction in Proposition 1. Note that the linear dependence on M can mean an exponential number of bits in the lower bound if we have that many *distinct* tasks (or agents), e.g. if $M = \Theta(D)$ for a large task state space D.

By considering the natural subclass of *smooth protocols*—where agents' posterior beliefs at $\langle M, N, \varepsilon, \delta\rangle$-agreement time must not diverge more than their initial priors, measured in total variation distance—we obtain a strictly improved lower bound:

Proposition 2 ("Smooth" Protocol Lower Bound)**.** *Let the number of tasks $M \geq 2$, and for each task $j \in [M]$, let the task state space size $D_j > 2$, $\varepsilon \leq \varepsilon_j$, $\delta_j < \nu/2$, and $0 < \nu \leq 1$. Furthermore, assume the protocol is smooth in that the total variation distance of the posteriors of the agents once $\langle M, N, \varepsilon, \delta\rangle$-agreement is reached is $\leq c\nu$ for $c < \frac{1}{2} - \frac{\delta_j}{\nu}$. There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}_j^i\}^{i\in[N]}$ with prior distance $\nu_j \geq \nu$, such that any smooth protocol among N agents needs to exchange:*

$$\Omega\left(M\,N^2\,\left(\nu + \log(1/\varepsilon)\right)\right)$$

bits to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement on $\{f_j\}_{j\in[M]}$.

Both lower bounds in Propositions 1 and 2 demonstrate that gaining consensus on a small list of M values that we want AI systems to have, will be essential for scalable alignment.

Finally, we consider a related smoothness condition—namely, the broad class of *bounded-Bayes-factor (BBF)* protocols—in which each message bit alters message likelihoods by at most a constant multiplicative factor. This assumption naturally captures realistic message-passing behavior, since rational and bounded agents typically update their beliefs incrementally rather than abruptly shifting posterior distributions after receiving a single message. Under this mild condition, we examine a natural setting: agents initially separated by prior distance ν first establish a *common prior* by satisfying the canonical equalities of Hellman and Samet (2012) (displayed in Algorithm 2 in Appendix §G.3), and subsequently condition on this shared prior to achieve $\langle M, N, \varepsilon, \delta\rangle$-agreement. They showed that for tight and connected knowledge partitions (defined below), these canonical equalities are automatically preserved under standard

[2]Note, unlike our upper bounds in Theorem 1 and Proposition 4, we use bits in the lower bound in order to apply to *all* possible protocols (continuous or discrete), regardless of how many bits are encoded per message. The upper bounds have to use messages (rounds) to describe either a continuous protocol (potentially infinitely many bits) as in Theorem 1, or a discrete protocol as in Proposition 4.

Bayesian updating; hence, our construction needs no further behavioral constraints beyond standard Bayesian rationality.

Under these reasonable conditions, our lower bound strengthens to include an extra multiplicative factor of $D := \min_j D_j$, the smallest state-space size across the M tasks. Thus, this refined lower bound more closely matches the general upper-bound results from §5 (cf. Algorithm 1) for this protocol class within an additive polynomial term in M, N, ε, and δ.

Proposition 3 (Canonical-Equality BBF Protocol Lower Bound)**.** *Let $M \geq 2$ be the number of tasks and let each task j have a finite state-space S_j with size $D_j > 2$. For every j, let the* initial *knowledge profiles of the N agents, $(\Pi_j^{1,0}, \ldots, \Pi_j^{N,0})$, be*

1. connected: *the alternation graph on states is connected, i.e. $\bigwedge_i \Pi_j^{i,0} = \{S_j\}$, so every two states are linked by an alternating chain of states; and*
2. tight: *that graph becomes disconnected if any edge is removed (unique chain property).*

Assume the message-passing protocol is BBF(β) for some $\beta > 1$: every b-bit message $m_j^{i,t}$ satisfies $\beta^{-b} \leq \Pr[m_j^{i,t} \mid s_j, \Pi_j^{i,t-1}(s_j)] / \Pr[m_j^{i,t} \mid s'_j, \Pi_j^{i,t-1}(s'_j)] \leq \beta^{\,b}$. Then there exist payoff functions $f_j : S_j \to [0,1]$ and priors $\{\mathbb{P}_j^{\,i}\}_{i\in[N]}$ with pairwise distance $\nu_j \geq \nu$, $0 < \nu \leq 1$, such that any BBF(β) protocol attaining $\langle M, N, \varepsilon, \delta\rangle$-agreement via the canonical equalities of Hellman and Samet (2012) must exchange at least

$$\Omega\left(M\, N^2\, [\, D\nu + \log(1/\varepsilon)\,]\right), \qquad D := \min_{j\in[M]} D_j,$$

bits in the worst case (implicit constant $= 1/\log\beta$), where the accuracy parameter $0 < \varepsilon \leq \varepsilon_j < 1$.

5 Convergence of $\langle M, N, \varepsilon, \delta\rangle$-agreement

Given these lower bounds, a natural question is whether $\langle M, N, \varepsilon, \delta\rangle$-agreement is achievable at all—especially since the agents begin without a common prior. In this section, we demonstrate that it is indeed achievable, providing explicit algorithms and upper bounds on convergence not only for idealized, unbounded agents but also under realistic constraints such as message discretization and computational boundedness. Here we prove the general upper bound:

Theorem 1. *N rational agents will $\langle M, N, \varepsilon, \delta\rangle$-agree with overall failure probability δ across M tasks, as defined in* (2)*, after $T = O\left(MN^2D + \dfrac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$.*

For an explicit algorithm, see Algorithm 1—we detail the reasoning behind this algorithm below.

First, we need to figure out at most how many messages need to be exchanged to guarantee at least one proper refinement. To do so, we will have the N agents communicate using the "spanning-tree" protocol of Aaronson (2005, §3.3), which we generalize to the multi-task, no common prior, setting below:

Algorithm 1: $\langle M, N, \varepsilon, \delta\rangle$-Agreement

Require: N agents with initial partitions $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$; protocol $\mathcal{P}$; ConstructCommonPrior defined in Algorithm 2; $\langle\varepsilon, \delta\rangle$-agreement protocol $\mathcal{A}$

Ensure: Agents reach $\langle\varepsilon_j, \delta_j\rangle$-agreement for all M tasks

for $j \leftarrow 1$ to M **do**
  $t \leftarrow 0$
  repeat
    $t \leftarrow t + 1$
    for all agent $i \in [N]$ **do**
      send $m_j^{i,t}$ via $\mathcal{P}$
      $\Pi_j^{i,t} \leftarrow \textsc{RefinePartition}\big(\Pi_j^{i,t-1}, m_j^{\cdot,t}\big)$
    end for
    $\mathbb{CP}_j \leftarrow \textsc{ConstructCommonPrior}\big(\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N\big)$
  until $\mathbb{CP}_j \neq$ Infeasible
  Condition all agents on $\mathbb{CP}_j$
  RunCPAgreement $(\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j)$
end for

Lemma 1 (Proper Refinement Message Mapping Lemma)**.** *If N agents communicate via a spanning-tree protocol for task j, where $g_j \in \mathbb{N}$ is the diameter of the chosen spanning trees, then as long as they have not yet reached agreement, it takes $O(g_j) = O(N)$ messages before at least one agent's knowledge partition is properly refined.*

Proof. Let G_j be a strongly connected directed graph with vertices $v \in [N]$ (one per agent), enabling communication of expectations $E_j^{i,t}$ along edges. (We need the strongly connected requirement on G_j, since otherwise the agents may not reach agreement for trivial reasons if they cannot reach one another.) Without loss of generality, let SP_j^1 and SP_j^2 be minimum-diameter spanning trees of G_j, each rooted at agent 1, with SP_j^1 pointing outward from agent 1 and SP_j^2 inward toward agent 1, each of diameter at most g_j.

Define orderings $\mathcal{O}_j^1$ (resp. $\mathcal{O}_j^2$) of edges in SP_j^1 (resp. SP_j^2) so each edge $(i \to k)$ appears only after edges $(\ell \to i)$, except when i is the root (or leaf, in inward trees). Construct $\mathrm{AgentOrdering}_j$ by cycling through $\mathcal{O}_j^1, \mathcal{O}_j^2, \ldots$, where in each round t the tail agent of $\mathrm{AgentOrdering}_j(t)$ sends its current expectation. Thus, every block of $O(g_j)$ transmissions forwards each agent's updated message along both trees, reaching all others.

Consequently, disagreement between any agents i and k leads to at least one agent receiving a "surprising" message within these $O(g_j)$ transmissions (worst-case occurs when i, k are on opposite ends of G_j), causing a partition refinement. Thus, without agreement, at least one refinement occurs every $O(g_j)$ messages.

Note $g_j = O(N)$ if G_j is a worst-case ring topology; more favorable topologies yield $g_j \ll N$, but we assume worst-case generality to subsume any specific cases. □

Next, we prove an important (for our purposes) lemma, which is an extension of Hellman (2013, Theorem 2)'s result on almost common priors to our M-function message setting:

Lemma 2 (Common Prior Lemma)**.** *If N agents have prior distance ν_j, as defined in* (3)*, for a task $j \in [M]$ with task state space S_j, then after $O\left(N^2 D_j\right)$ messages, they will have a common prior $\mathbb{CP}_j$ with probability 1 over their type profiles.*

Once the agents reach a common prior $\mathbb{CP}_j$, they can then condition on that for the rest of their conversation to reach the desired $1-\delta_j\ \varepsilon_j$-agreement threshold (cf. Step 12 of Algorithm 1). We assume this is $O(1)$ to compute for now as the agents are computationally unbounded, but we will remove this assumption in §5.2, and instead use Algorithm 2 (Appendix §G.3) for an efficient explicit construction via LP feasibility of posterior belief ratios.

Therefore, for each task j, we have reduced the problem now to Aaronson's $\langle\varepsilon,\delta\rangle$-agreement framework (Aaronson 2005), and as he shows, the subsequent steps conditioning on a common prior become unbiased random walks with step size roughly ε_j. With some slight modifications, this allows us to give a worst-case bound on the number of remaining steps in our $\langle M,N,\varepsilon,\delta\rangle$-agreement setting:

Lemma 3. *For all f_j and $\mathbb{CP}_j$, the N agents will globally $\langle\varepsilon_j,\delta_j\rangle$-agree after $O\left(N^7/(\delta_j\varepsilon_j)^2\right)$ additional messages.*

Proof. By Aaronson (2005, Theorem 10), the N agents will *pairwise* $\langle\varepsilon_j,\delta_j\rangle$-agree after $O\left(\left(Ng_j^2\right)/(\delta_j\varepsilon_j)^2\right)$ messages when they condition on $\mathbb{CP}_j$, where g_j is the diameter of the spanning-tree protocol they use. Furthermore, we will need to have them $\left\langle\varepsilon_j,\delta_j/N^2\right\rangle$-agree pairwise so that they *globally* $\langle\varepsilon_j,\delta_j\rangle$-agree. Taking $g_j = O(N)$ for the worst-case ring topology gives us the above bound. □

By Lemmas 2 and 3, for *each* $j \in [M]$, we need $O\left(N^2 D_j + \frac{N^7}{(\delta_j\varepsilon_j)^2}\right)$ messages for the N agents to reach $\langle M=1,N,\varepsilon_j,\delta_j\rangle$-agreement. Next, select a uniform δ such that $\delta_j \le \delta/M$, for all $j \in [M]$. Therefore, by a union bound, we get the full upper bound in Theorem 1 with total probability $\ge 1-\delta$, across *all* M tasks, by *maximizing* the bound above by taking $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$, and scaling by M.

5.1 Discretized Extension

A natural extension of Theorem 1 is if the agents do not communicate their full real-valued expectation (which may require infinitely many bits), but a discretized version of the current expectation, corresponding to whether it is above or below a given threshold (defined below), e.g. "High", "Medium", or "Low" (requiring only 2 bits). We prove convergence in this case, and show that the bound from Theorem 1 remains unchanged in this setting. Discretization is important to show convergence and complexity analysis for, since this most closely matches real-world constraints (e.g. LLM agents use discrete, real-valued tokens), as opposed to infinite-bit real valued messages.

Proposition 4 (Discretized Extension)**.** *If N agents only communicate their* discretized *expectations, then they will $\langle M,N,\varepsilon,\delta\rangle$-agree with overall failure probability δ across M tasks as defined in* (2)*, after $T = O\left(MN^2D + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$.*

Our discretized three-bucket protocol itself is general and imposes no BBF constraint—in Appendix §F we show it can be made BBF(3)-compliant with small overhead. Thus, by the lower bound from Proposition 3, for the broad and natural class of canonical-equality BBF protocols, our upper bound in Proposition 4 is tight up to an additive polynomial term after converting from messages to bits.

5.2 Computationally Bounded Agents

Thus far, we analyzed computationally unbounded agents, implicitly assuming $O(1)$ time for constructing and sending messages, finding common priors, and sampling distributions. Even under these idealized conditions, the linear scaling in Theorem 1 becomes significant if the task space D or number of tasks M is exponentially large.

However, realistic agents, such as current LLMs, are computationally bounded, and message passing may be noisy, e.g., due to obfuscated intent (Barnes and Christiano 2020). Thus, we now analyze the complexity of N computationally *bounded* rational agents. Moreover, since querying humans typically costs more than querying AI agents, we differentiate between q humans (each taking T_H time steps) and $N-q$ AI agents (each taking T_{AI} time steps), encompassing recent multi-step reasoning models (Jaech et al. 2024; DeepMind 2024). Without loss of generality, we assume uniform times within these two groups and analyze complexity based on two basic subroutines:

Requirement 1 (Basic Capabilities of Bounded Agents)**.** We expect the agents to be able to:

1. **Evaluation:** The N agents can each evaluate $f_j(s_j)$ for any state $s_j \in S_j$, taking time $T_{\text{eval},a}$ steps for $a \in \{H, AI\}$.
2. **Sampling:** The N agents can sample from the *unconditional* distribution of any other agent, such as their prior $\mathbb{P}_j^i$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

We treat these subroutines as black boxes: agents lack explicit descriptions of f_j and distributions, learning about them solely through these operations. Analogous to CIRL (Hadfield-Menell et al. 2016), this setup captures realistic alignment scenarios where the correctness of a task outcome can be verified without specifying each intermediate step. Consequently, our complexity results are broadly applicable, expressed in terms of $T_{\text{eval},H}$, $T_{\text{eval},AI}$, $T_{\text{sample},H}$, and $T_{\text{sample},AI}$.

These minimal subroutines enable agents to estimate each other's expectations, an essential capability for alignment. Importantly, exact computation is unnecessary; probabilistic evaluation in polynomial time suffices (as will become clear

in the proof of Theorem 2, due to the exponential blow-up). The sampling subroutine further serves as a bounded version of the standard assumption that agents know each other's knowledge partitions through shared states (Aumann 1976, 1999). This corresponds to agents possessing a bounded "theory of mind" (Ho, Saxe, and Cushman 2022) about one another.

Finally, as we can no longer assume $O(1)$ time complexity for constructing a common prior (unlike in the unbounded agent setting), we introduce an explicit randomized polynomial-time algorithm for doing so with high probability, Algorithm 2. We refer the reader to Appendix §G.3 for proofs related to Algorithm 2. Specifically, Lemma 7 (correctness), Lemma 8 (runtime analysis), and Lemma 9 (inexact posterior access setting).

In what follows, define

$$T_{N,q} := q\,T_{\text{sample},H} + (N-q)\,T_{\text{sample},AI} + q\,T_{\text{eval},H} + (N-q)\,T_{\text{eval},AI}.$$

The above considerations lead to the following theorem in the bounded agent setting:

Theorem 2 (Bounded Agents Eventually Agree)**.** *Let there be N computationally bounded rational agents (consisting of $1 \le q < N$ humans and $N - q \ge 1$ AI agents), with the capabilities in Requirement 1. The agents pass messages according to the sampling tree protocol (detailed in Appendix §G.2) with branching factor of $B \ge 1/\alpha$, and added triangular noise of width $\le 2\alpha$, where $\varepsilon/50 \le \alpha \le \varepsilon/40$. Let $\delta^{\text{find\_CP}}$ be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let $\delta^{\text{agree\_CP}}$ be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta\rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{\text{find\_CP}} + \delta^{\text{agree\_CP}} < \delta$. For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\ge 1 - \delta$, takes time*

$$O\left(M\,T_{N,q}\left(B^{N^2 D\frac{\ln\left(\delta^{\text{find\_CP}}/(3MN^2D)\right)}{\ln(1/\alpha)}} + B^{\frac{9M^2N^7}{(\delta^{\text{agree\_CP}}\varepsilon)^2}}\right)\right).$$

In other words, just in the first term alone, exponential *in the task space size D and number of agents N (and exponential in the number of tasks M in the second term). So if the task space size is in turn exponential in the input size, then this would already be* doubly exponential *in the input size!*

We now clarify why we let B be a parameter, and give a concrete example of how bad this exponential dependence can be. Intuitively, we can think of B as a "gauge" on how distinguishable the bounded agents are from "true" unbounded Bayesians, and will allow us to give an explicit desired value for B. Recognizing the issue of computational boundedness of agents in the real world, Hanson (2003) introduced the notion of *Bayesian wannabes*: agents who estimate expectations as if they had sufficient computational resources. He showed that disagreement among Bayesian wannabes stems from computational limitations rather than differing information. Extending this idea, Aaronson (2005) proposed a protocol ensuring that bounded agents appear statistically indistinguishable from true Bayesians to an external referee—effectively a "Bayesian Turing Test" (Turing 1950) for rationality. Thus, B explicitly quantifies this notion of bounded Bayesian indistinguishability.

We consider the M-function, N-agent generalization of this requirement (and *without* common priors (CPA)), which we call a "*total Bayesian wannabe*":

Definition 1 (Total Bayesian Wannabe)**.** Let the N agents have the capabilities in Requirement 1. For each task $j \in [M]$, let the transcript of T messages exchanged between N agents be denoted as $\Gamma_j := \langle m_j^1, \ldots, m_j^T\rangle$. Let their initial, task-specific priors be denoted by $\{\mathbb{P}_j^i\}^{i\in[N]}$. Let $\mathcal{B}(s_j)$ be the distribution over message transcripts if the N agents are unbounded Bayesians, and the current task state is $s_j \in S_j$. Analogously, let $\mathcal{W}(s_j)$ be the distribution over message transcripts if the N agents are "total Bayesian wannabes", and the current task state is $s_j \in S_j$. Then we require for all Boolean functions[3] $\Phi(s_j, \Gamma_j)$,

$$\left\| \begin{array}{c} \mathbb{P}_{\substack{\Gamma_j\in\mathcal{W}(s_j)\\ s_j\in\mathscr{S}_j}}\big[\Phi(s_j,\Gamma_j)=1\big] \\ \\ -\,\mathbb{P}_{\substack{\Gamma_j\in\mathcal{B}(s_j)\\ s_j\in\mathscr{S}_j}}\big[\Phi(s_j,\Gamma_j)=1\big] \end{array} \right\|_1 \le \rho_j, \qquad \forall j \in [M],$$

where $\mathscr{S}_j := \{\mathbb{P}_j^i\}_{i\in[N]}$. We can set $\rho_j \in \mathbb{R}$ as arbitrarily small as preferred, and it will be convenient to only consider a single $\rho := \min_{j\in[M]} \rho_j$ without loss of generality (corresponding to the most "stringent" task j).

We will show in Appendix §G.3 that matching this requirement amounts to picking a large enough value for B, giving rise to the following corollary to Theorem 2:

Corollary 1 (Total Bayesian Wannabes Agree)**.** *Let there be N total Bayesian wannabes, according to Definition 1 (e.g. consisting of $1 \le q < N$ humans and $N - q \ge 1$ AI agents). Let the branching factor of the sampling tree protocol be the same as before, $B \ge 1/\alpha$, with added triangular noise of width $\le 2\alpha$, where $\varepsilon/50 \le \alpha \le \varepsilon/40$. Let $\delta^{\text{find\_CP}}$ be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let $\delta^{\text{agree\_CP}}$ be the maximal failure probability of the agents to come to $\langle M, N, \varepsilon, \delta\rangle$-agreement across all M tasks once they condition on a common prior, where $\delta^{\text{find\_CP}} + \delta^{\text{agree\_CP}} < \delta$. For the N "total Bayesian wannabes" to $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\ge 1 - \delta$, takes time*

$$O\Big(M\,T_{N,q}\big(B^{N^2 D\frac{\ln\left(\delta^{\text{find\_CP}}/(3MN^2D)\right)}{\ln(1/\alpha)}} + (11/\alpha)^{\frac{729M^6N^{21}}{(\delta^{\text{agree\_CP}}\varepsilon)^6}}\,\rho^{-\frac{18M^2N^7}{(\delta^{\text{agree\_CP}}\varepsilon)^2}}\big)\Big).$$

In other words, exponential time in the task space D, and by (18)*, and with a large base in the second term if the "total Bayesian wannabe" threshold ρ is made small.*

Sharing a common prior amounts to removing the first term, yielding upper bounds that are still exponential in ε and δ.

[3]Without loss of generality, we assume that the current task state s_j and message transcript Γ_j are encoded as binary strings.

The proofs of Theorem 2 and Corollary 1 are quite technical (spanning 7 pages), so we defer them to Appendix §G for clarity. The primary takeaway here is that computational boundedness can result in a severely exponential time slowdown in the agreement time, and especially so if you want the bounded agents to be *statistically indistinguishable* in their interactions with each other from true unbounded Bayesians.

For example, even for $N = 2$ agents with a common prior and liberal agreement threshold of $\varepsilon = \delta = 1/2$ and "total Bayesian wannabe" threshold of $\rho = 1/2$ on one task ($M = 1$), then $\alpha \geq 1/100$, the number of *subroutine calls* (not even total runtime) would be around:

$$O\left(\frac{(1100)^{\frac{1528823808}{(1/4)^6}}}{(1/2)^{\frac{2304}{(1/4)^2}}}\right) \approx O\left(10^{10^{13.27979}}\right),$$

would already far exceed the estimated (Munafo 2013, pg. 19) number of atoms in the universe ($\sim 4.8 \times 10^{79}$)! This illustrates the power of the *unbounded* Bayesians we considered earlier in §4, and why the lower bounds there are worth paying attention to in practice.

Finally, note that in general under a sampling tree protocol, this exponential blow-up in task state space size D is unavoidable (e.g. for rare, potentially unsafe, events):

Proposition 5 (Needle-in-a-Haystack Sampling Tree Lower Bound)**.** *Let* $T_{N,q,\text{sample}} := qT_{\text{sample},H} + (N - q)T_{\text{sample},AI}$. *For* any *sampling-tree protocol, a single task and a single pair of agents can be instantiated so that the two agents' priors differ by prior distance* $\geq \nu$*, yet the protocol must pre-compute at least* $\Omega\left(\nu^{-1}\right)$ *unconditional samples before the first online message. Consequently, for a particular "needle" prior construction of* $\nu = \Theta\left(e^{-D}\right)$*, we get lower bounds that are exponential in the task state space size D, needing* $\Omega\left(M\, T_{N,q,\text{sample}}\, e^{D}\right)$ *wall-clock time.*

6 Discussion

Why study a "Bayesian best-case" at all? One may object that real AI systems—and certainly humans—are not perfectly Bayesian reasoners, nor do they interact through ideal, lossless channels. That is precisely the point: our results constitute an *ideal benchmark*, before we build and deploy capable agents. If alignment is information- or communication-theoretically hard even for computationally *unbounded*, rational Bayesians exchanging noiseless messages, then relaxing rationality and unboundedness assumptions, adding noise, strategic behavior, or adversarial tampering can exacerbate the difficulty, as we showed in §5.2. Our takeaways for AI safety are:

1. **Too many alignment values drives alignment cost.** Our matched lower and upper bounds (tight up to polynomial terms in $M, N, \varepsilon, \delta$) demonstrate a "No-Free-Lunch" principle: encoding an exponentially large or high-entropy set of human values forces at least exponential communication even for *unbounded* agents. As a result, progress on value alignment / preference modeling should prioritize objective compression, delegation, or progressive disclosure rather than attempting one-shot, full-coverage specification.
2. **Reward hacking is *globally* inevitable.** Proposition 5 shows that in large state spaces and with bounded agents, reward hacking arises unavoidably from finite sampling. By Proposition 3, this even happens for unbounded agents in large state spaces who communicate finite bits and update their expectations smoothly. Scalable oversight is therefore not about uniform alignment, but about focusing on the parts of the state space that matter most. The engineering task ahead of us then is the *mechanism design* problem of benchmarks and interactive protocols that target these safety-critical slices—via adversarial sampling, objective compression, and per-slice $\langle \varepsilon, \delta \rangle$ budgets—to certify coverage where it counts.
3. **Robustness depends on bounded rationality, memory, and theory of mind.** Introducing bounded agents or even mild triangular noise can exponentially increase costs when protocols cannot exploit additional structure or restrict the task state space (Ball and Haupt 2025); yet these assumptions were necessary to prove any alignment guarantees at all. Robust alignment must account for imperfect agents and noisy or obfuscated channels—but as we show in §5.2, real-world agents with these three properties can degrade *gracefully* rather than catastrophically.
4. **Tight bounds inform governance thresholds.** For broad and natural protocol classes, our lower bounds are closely matched (up to polynomial terms) by constructive algorithms, enabling principled risk thresholds.

Limitations and future work. Our results justify *cautious optimism*: alignment is tractable in principle, yet only when we restrain objectives and exploit task structure with care. "No-Free-Lunch" does not preclude lunch—it simply forces wise menu choices. At least three directions stand out:

1. **Minimal value sets.** Our lower bounds imply that having too many objectives is the surest route to inefficiency. A key open question is *which* small, consensus-worthy utility families guarantee high-probability safety. In concurrent follow-up work (Nayebi 2025), we identify such a small value set for corrigibility as defined by Soares et al. (2015), which was open for a decade.
2. **Structure-exploiting interaction protocols.** Design *multi-turn* agent interaction protocols (beyond single-shot RLHF) and evaluation benchmarks that stress-test the portions of state space most relevant for safety during deployment. This can also be done at the post-training stage, and can augment existing RLHF pipelines.
3. **Beyond expectations under noise. (i)** Can agreement on *specific* risk measures cut communication costs relative to full-expectation alignment? We note that agreement on full expectations is not always required; given a task-specific utility function U_j, our framework already covers agreement on optimal actions, $\arg\max_a \mathbb{E}[U_j(a)]$ by having f_j be the optimal action indicator. Our framework also models rare events (Appendix §C). **(ii)** We found bounded derivative in the noise model was crucial for convergence (e.g. uniform noise does not suffice). Studying richer obfuscation (e.g. learned steganography) will be essential for informing other robust safety thresholds.

Acknowledgements

We thank the Burroughs Wellcome Fund (CASI award) for financial support. We also thank Scott Aaronson, Andreas Haupt, Richard Hu, J. Zico Kolter, and Max Simchowitz for helpful discussions on AI safety in the early stages of this work, and Nina Balcan for feedback on a draft of the manuscript.

References

Aaronson, S. 2005. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 634–643.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Aumann, R. J. 1976. Agreeing to Disagree. *The Annals of Statistics*, 4(6): 1236–1239.

Aumann, R. J. 1999. Interactive epistemology I: knowledge. *International Journal of Game Theory*, 28: 263–300.

Ball, S.; and Haupt, A. 2025. Don't Walk the Line: Boundary Guidance for Filtered Generation. *arXiv preprint arXiv:2510.11834*.

Barnes, B.; and Christiano, P. 2020. Debate Update: Obfuscated Arguments Problem. https://www.lesswrong.com/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem.

Brown-Cohen, J.; Irving, G.; and Piliouras, G. 2023. Scalable AI safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*.

Brown-Cohen, J.; Irving, G.; and Piliouras, G. 2025. Avoiding Obfuscation with Prover-Estimator Debate. *arXiv preprint arXiv:2506.13609*.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Collina, N.; Goel, S.; Gupta, V.; and Roth, A. 2025. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, 1532–1543.

DeepMind, G. 2024. Gemini 2.0 Flash: Advancing AI Capabilities. https://tinyurl.com/4szdruva. Released: 2024-12-14.

Guan, M. Y.; Joglekar, M.; Wallace, E.; Jain, S.; Barak, B.; Heylar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Hanson, R. 2003. For Bayesian wannabes, are disagreements not about information? *Theory and Decision*, 54: 105–123.

Harsányi, J. C. 1967–1968. Games with incomplete information played by Bayesian players. *Management Science*, 14: 159–182, 320–334, 486–502.

Hellman, Z. 2013. Almost common priors. *International Journal of Game Theory*, 42: 399–410.

Hellman, Z.; and Samet, D. 2012. How common are common priors? *Games and Economic Behavior*, 74(2): 517–525.

Ho, M. K.; Saxe, R.; and Cushman, F. 2022. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11): 959–971.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Munafo, R. 2013. Notable Properties of Specific Numbers.

Nayebi, A. 2025. Core Safety Values for Provably Corrigible Agents. *arXiv preprint arXiv:2507.20964*.

Rockafellar, R. T.; and Uryasev, S. 2000. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4): 105–114.

Soares, N. 2018. The value learning problem. In *Artificial intelligence safety and security*, 89–97. Chapman and Hall/CRC.

Soares, N.; Fallenstein, B.; Armstrong, S.; and Yudkowsky, E. 2015. Corrigibility. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind*, 59: 433–460.