Electronic Colloquium on Computational Complexity, Report No. 186 (2025)

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.

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 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*

ISSN 1433-8092

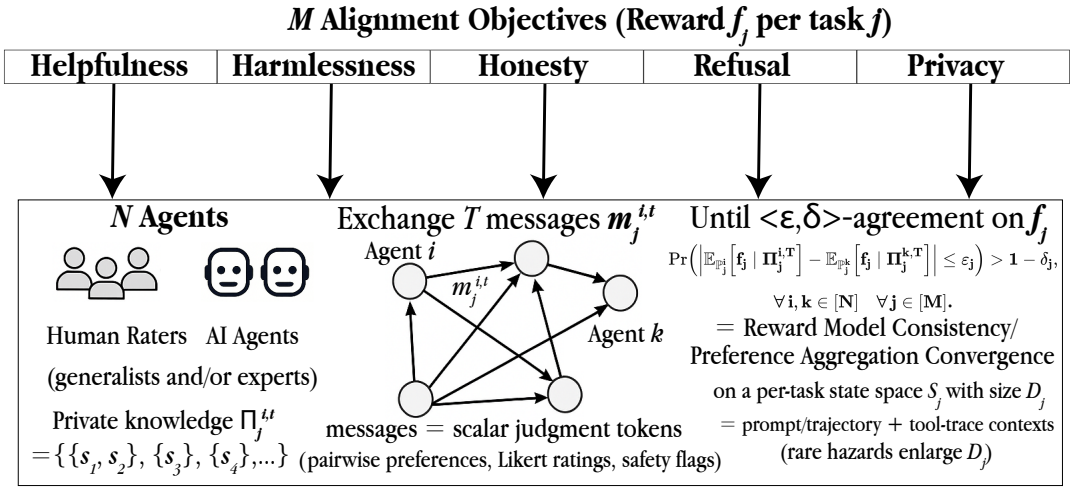

Figure 1: Mapping our $\langle M, N, \varepsilon, \delta\rangle$-agreement to current RLHF/DPO/Constitutional AI pipelines.

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.

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; $\nu_j = 0$ iff a true common prior exists. Note by the triangle inequality, $\nu_j \ge \|\mathbb{P}_j^i - \mathbb{P}_j^k\|_1$ 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

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

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 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 $\min_{j\in[M]} \varepsilon_j$.*

[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.

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 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:

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.*

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 $\mathcal{P}$
      $\Pi_j^{i,t} \leftarrow \text{RefinePartition}\big(\Pi_j^{i,t-1}, m_j^{\cdot,t}\big)$
    end for
    $\mathbb{CP}_j \leftarrow \text{ConstructCommonPrior}\big(\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N\big)$
  until $\mathbb{CP}_j \neq \text{Infeasible}$
  Condition all agents on $\mathbb{CP}_j$
  RunCPAgreement $(\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j)$
end for

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 AgentOrdering_j by cycling through $\mathcal{O}_j^1, \mathcal{O}_j^2, \ldots$, where in each round t the tail agent of $\text{AgentOrdering}_j(t)$ 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 $O\left(N^2D_j+\frac{N^7}{(\delta_j\varepsilon_j)^2}\right)$ 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.

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*

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

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*. To appear in AAAI 2026 Machine Ethics Workshop (W37) Proceedings.

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.

A Notational Preliminaries

We will use asymptotic notation throughout that is standard in computer science, but may not be in other fields. The asymptotic notation is defined as follows:

- $F(n) = O(G(n))$: There exist positive constants $c_1 > 0$ and $c_2 > 0$ such that $F(n) \leq c_1 + c_2 G(n)$, for all $n \geq 0$.
- $F(n) = \tilde{O}(G(n))$: There exist positive constants c_1, c_2, and $k > 0$ such that $F(n) \leq c_1 + c_2 G(n) \log^k n$, for all $n \geq 0$.
- $F(n) = \Omega(G(n))$: Similarly, there exist positive constants c_1 and c_2 such that $F(n) \geq c_1 + c_2 G(n)$, for all $n \geq 0$.
- $F(n) = \Theta(G(n))$: This indicates that $F(n) = O(G(n))$ and $F(n) = \Omega(G(n))$. In other words, $G(n)$ is a tight bound for $F(n)$.

For notational convenience, let

$$E_j^{i,t}(s_j) := \mathbb{E}_{\mathbb{P}_j^i}\left[f_j \mid \Pi_j^{i,t}(s_j)\right],$$

which is the expectation of f_j of agent i at timestep t, conditioned on its knowledge partition by then, starting from its *own* prior $\mathbb{P}_j^i$. To simplify notation, we drop the argument $s_j \in S_j$.

B $\langle M, N, \varepsilon, \delta\rangle$-Agreement Setup and Dynamics

The framework we consider for alignment generalizes Aumann agreement (Aumann 1976) to probabilistic $\langle \varepsilon, \delta\rangle$-agreement (Aaronson 2005) (rather than exact agreement), across M agreement objectives and N agents, *without* the Common Prior Assumption (CPA). The CPA dates back to at least Harsányi (1967–1968) in his seminal work on games with incomplete information. This is a very powerful assumption and is at the heart of Aumann's agreement theorem that two rational Bayesian agents must agree if they share a common prior (Aumann 1976). As a further illustration of how powerful the CPA is from a computational complexity standpoint, Aaronson (2005) relaxed the exact agreement requirement to $\langle \varepsilon, \delta\rangle$-agreement and showed that even in this setting, completely independent of how large the state space is, two agents with common priors will need to only exchange $O(1/(\delta\varepsilon^2))$ messages to agree within ε with probability at least $1 - \delta$ over their prior. However, the CPA is clearly a very strong assumption for human-AI alignment, as we cannot expect that our AIs will always *start out* with common priors with every human it will engage with on every task. In fact, even between two *humans* this assumption is unlikely! For other aspects of agreement and how they relate more broadly to alignment, we defer to the Discussion (§6) for a more detailed treatment.

In short, $\langle M, N, \varepsilon, \delta\rangle$-agreement represents a "best-case" scenario that is general enough to encompass prior approaches to alignment (cf. Table 1), such that if something is inefficient here, then it forms a prescription for what to avoid *in practice*, in far more suboptimal circumstances. As examples of suboptimality in practice, we will consider computational boundedness and noisy messages in §5.2, to exactly quantify how the bounds can significantly (e.g. exponentially) worsen.

Dispensing with the CPA, we now make our $\langle M, N, \varepsilon, \delta\rangle$-agreement framework more precise. For illustration, we consider two agents ($N = 2$), Alice (human) and "Rob" (robot), denoted by $\mathbf{A}$ and $\mathbf{R}$, respectively. Let $\{S_j\}_{j\in[M]}$ be the collection of (not necessarily disjoint) possible task states for each task $j \in [M]$ they are to perform. We assume each S_j is finite ($|S_j| = D_j \in \mathbb{N}$), as this is a standard assumption, and any physically realistic agent can only encounter a finite number of states anyhow. There are M agreement objectives, $f_1, \ldots, f_M$, that Alice and Rob want to jointly estimate, one for each task:

$$f_j : S_j \to [0, 1], \quad \forall j \in [M],$$

to encompass the possibility of changing needs and differing desired $\{\langle \varepsilon_j, \delta_j\rangle\}_{j\in[M]}$-agreement thresholds for those needs (which we will define shortly in (2)), rather than optimizing for a single monolithic task. Note that setting the output of f_j to $[0, 1]$ does not reduce generality. Since S_j is finite, any function $S_j \to \mathbb{R}$ has a bounded range, so one can always rescale appropriately to go inside the $[0, 1]$ domain.

Alice and Rob have priors $\mathbb{P}_j^{\mathbf{A}}$ and $\mathbb{P}_j^{\mathbf{R}}$, respectively, over task j's state space S_j. Let $\nu_j \in [0, 1]$ denote the *prior distance* (as introduced by Hellman (2013)) between $\mathbb{P}_j^{\mathbf{A}}$ and $\mathbb{P}_j^{\mathbf{R}}$, defined as the minimal L^1 distance between any point $x_j \in X_j = \mathbb{P}_j^{\mathbf{A}} \times \mathbb{P}_j^{\mathbf{R}}$ and any point $p_j \in \mathcal{D}_j = \{(p_j, p_j) \mid p_j \in \Delta(S_j)\}$, where $\Delta(S_j) \in \mathbb{R}^{D_j}$ is the probability simplex over the states in S_j. Formally,

$$\nu_j = \min_{x_j \in X_j, p_j \in \mathcal{D}_j} \|x_j - p_j\|_1, \quad \forall j \in [M]. \tag{3}$$

It is straightforward to see that there exists a *common prior* $\mathbb{CP}_j \in \mathcal{D}_j$ between Alice and Rob for task j if and only if the task state space S_j has prior distance $\nu_j = 0$. (Lemma 2 will in fact show that it is possible to find a common prior with high probability, regardless of the initial value ν_j.)

For every state $s_j \in S_j$, we identify the subset $E_j \subseteq S_j$ with the event that $s_j \in S_j$. For each task $j \in [M]$, Alice and Rob exchange messages[4] from the power set $\mathscr{P}(S_j)$ of the task state space S_j, as a sequence $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0, 1]$. Let $\Pi_j^{i,t}(s_j)$ be the set of states that agent $i \in \{\mathbf{A}, \mathbf{R}\}$ considers possible in task j after the t-th message has been sent, given that the true state of the world for task j is s_j. Then by construction, $s_j \in \Pi_j^{i,t}(s_j) \subseteq S_j$, and the set $\{\Pi_j^{i,t}(s_j)\}_{s_j \in S_j}$ forms a *partition* of S_j (known as a "knowledge partition"). As is standard (Aumann 1976, 1999; Aaronson 2005; Hellman 2013), we assume for each task j, the agents know each others' initial knowledge partitions $\{\Pi_j^{i,0}(s_j)\}_{s_j \in S_j}$. The justification for this more broadly (Aumann 1976, 1999) is that a given state of the world $s_j \in S_j$ includes the agents' knowledge.

[4] These messages could be as simple as communicating the agent's current expectation of f_j, given (conditioned on) its current knowledge partition. For now, we assume the messages are not noisy, but we will remove this assumption in §5.2.

In our setting, it is quite natural to assume that task states for agents coordinating on a task will encode their knowledge. As a consequence, every agents' subsequent partition is known to every other agent, and every agent knows that this is the case, and so on[5]. This is because with this assumption, since the agents receive messages from each other, then $\Pi_j^{i,t}(s_j) \subseteq \Pi_j^{i,t-1}(s_j)$. In other words, subsequent knowledge partitions $\{\Pi_j^{i,t}(s_j)\}_{s_j \in S_j}$ *refine* earlier knowledge partitions $\{\Pi_j^{i,t-1}(s_j)\}_{s_j \in S_j}$. (Equivalently, we say that $\{\Pi_j^{i,t-1}(s_j)\}_{s_j \in S_j}$ *coarsens* $\{\Pi_j^{i,t}(s_j)\}_{s_j \in S_j}$.) *Proper* refinement is if for at least one state $s_j \in S_j$, $\Pi_j^{i,t}(s_j) \subsetneq \Pi_j^{i,t-1}(s_j)$, representing a *strict* increase in knowledge.

To illustrate this more concretely, first Alice computes $m_j^1\left(\Pi_j^{\mathbf{A},0}(s_j)\right)$ and sends it to Rob. Rob's knowledge partition then becomes refined to the set of messages in his original knowledge partition that match Alice's message (since they are now both aware of it):

$$\Pi_j^{\mathbf{R},1}(s_j) \;=\; \big\{\, s'_j \in \Pi_j^{\mathbf{R},0}(s_j) \mid\; m_j^1\big(\Pi_j^{\mathbf{A},0}(s'_j)\big) = m_j^1\big(\Pi_j^{\mathbf{A},0}(s_j)\big)\big\},$$

from which Rob computes $m_j^2\left(\Pi_j^{\mathbf{R},1}(s_j)\right)$ and sends it to Alice. Alice then updates her knowledge partition similarly to become the set of messages in her original partition that match Rob's message:

$$\Pi_j^{\mathbf{A},2}(s_j) \;=\; \big\{\, s'_j \in \Pi_j^{\mathbf{A},0}(s_j) \mid\; m_j^2\big(\Pi_j^{\mathbf{R},1}(s'_j)\big) = m_j^2\big(\Pi_j^{\mathbf{R},1}(s_j)\big)\big\},$$

and then she computes and sends the message $m_j^3\left(\Pi_j^{\mathbf{A},2}(s_j)\right)$ to Rob, etc.

$\langle M, N, \varepsilon, \delta\rangle$-Agreement Criterion: We examine here the number of messages (T) required for Alice and Rob to $\langle \varepsilon_j, \delta_j\rangle$-agree across all tasks $j \in [M]$, defined as

$$\Pr\Big(\Big|\, \mathbb{E}_{\mathbb{P}_j^{\mathbf{A}}}\left[f_j \mid \Pi_j^{\mathbf{A},T}(s_j)\right] - \mathbb{E}_{\mathbb{P}_j^{\mathbf{R}}}\left[f_j \mid \Pi_j^{\mathbf{R},T}(s_j)\right]\Big| \le \varepsilon_j\Big) > 1-\delta_j, \quad \forall j \in [M]. \tag{4}$$

In other words, they agree within ε_j with high probability ($> 1-\delta_j$) on the expected value of f_j with respect to their *own* task-specific priors (not a common prior!), conditioned[6] on each of their knowledge partitions by time T.

Extending this framework to $N > 2$ agents (consisting of $1 \le q < N$ humans and $N - q \ge 1$ AI agents), is straightforward: we can have their initial, task-specific priors be denoted by $\{\mathbb{P}_j^i\}^{i\in[N]}$, and we can have them $\langle \varepsilon_j, \delta_j/N^2\rangle$-agree pairwise so that they globally $\langle \varepsilon_j, \delta_j\rangle$-agree.

[5]This can be implemented via a "*common knowledge*" set, $\mathcal{C}\left(\{\Pi_j^{i,t}\}^{i\in[N]}\right)$, which is the finest common coarsening of the agents' partitions (Aumann 1976).

[6]For completeness, note that for any subset $E_j \subseteq S_j$ and distribution $\mathbb{P}$, $\mathbb{E}_{\mathbb{P}}[f_j \mid E_j] := \sum_{s_j \in E_j} f(s_j)\mathbb{P}[s_j \mid E_j] = \dfrac{\sum_{s_j \in E_j} f(s_j)\mathbb{P}[s_j]}{\sum_{s_j \in E_j} \mathbb{P}[s_j]}$.

C Modeling Tail Risk

We note in this section that our $\langle M, N, \varepsilon, \delta\rangle$-agreement framework can also model tail risk/rare events. For exposition convenience, we use the "loss" convention (higher = worse), so the Expected Shortfall (ES)/Conditional Value at Risk (CVaR) at level $\tau \in (0, 1]$ uses the upper quantile/tail. Specifically, for a catastrophe indicator $f_j := Z \in \{0, 1\}$ with $\mathbb{E}[f_j] = \Pr[Z = 1] = p$, the ES/CVaR at a given level τ is

$$\mathrm{ES}^\tau(Z) = \frac{1}{\tau}\int_{1-\tau}^{1} q_u(Z)du,$$

where for $Z \sim \mathrm{Bernoulli}(p)$, the quantile is defined as:

$$q_u(Z) = \begin{cases} 0, & u \le 1-p, \\ 1, & u > 1-p. \end{cases}$$

Therefore,

$$\begin{aligned} \mathrm{ES}^\tau(Z) &= \frac{1}{\tau}\int_{1-\tau}^{1} \mathbf{1}\{u > 1-p\}\, du \\ &= \frac{1}{\tau}\left|(1-\tau, 1] \cap (1-p, 1]\right| \\ &= \frac{1}{\tau}\min\{\tau, p\} \\ &= \min\left\{1, \frac{p}{\tau}\right\}. \end{aligned}$$

Hence, if two models agree on p within ε, their ES values differ by at most ε/τ. For a general bounded loss $f_j := X \in [0, 1]$, the Rockafellar and Uryasev (2000, Eq. 4) representation

$$\mathrm{ES}^\tau(X) = \inf_{c\in[0,1]}\left(c + \frac{1}{\tau}\mathbb{E}[(X-c)_+]\right)$$

shows that ES is the minimum over expectations of bounded transforms $\psi_c(x) = c + \frac{1}{\tau}(x-c)_+$. Then

$$\mathrm{ES}^\tau(X) = \inf_{c\in[0,1]} \mathbb{E}[\psi_c(X)]\,.$$

Consequently, for two distributions P, Q over $X \in [0, 1]$,

$$\begin{aligned} &\left|\mathrm{ES}_P^\tau(X) - \mathrm{ES}_Q^\tau(X)\right| \\ &\le \sup_{c\in[0,1]}\left|\mathbb{E}_P[\psi_c(X)] - \mathbb{E}_Q[\psi_c(X)]\right| \\ &\le \frac{1}{\tau}\sup_{c\in[0,1]}\left|\mathbb{E}_P[(X-c)_+] - \mathbb{E}_Q[(X-c)_+]\right|. \end{aligned}$$

Thus, any $\langle M, N, \varepsilon, \delta\rangle$-agreement bounds controlling expectations of bounded functions directly yield corresponding bounds on ES (scaled by $1/\tau$).

D Proofs of Lower Bounds

D.1 Proof of Proposition 1

Proof. For each task $j \in [M]$, let the input tuple to the N agents be

$$\langle x_{1,j},\, x_{2,j},\, \ldots,\, x_{N,j}\rangle \;\in\; S_j,$$

where S_j is defined by

$$S_j \;:=\; \big\{\, \langle x_{1,j},\ldots,x_{N,j}\rangle \;\big|\; x_{i,j} \in \{(j-1)\,2^n+1,\ldots \\ j\,2^n\},\ \forall i \in [N]\big\}.$$

Thus, each $x_{i,j}$ is an integer[7] in an interval of size 2^n that starts at $(j-1)\cdot 2^n+1$. We endow S_j with the uniform common prior $\mathbb{CP}_j$ (which will be necessarily difficult by the counting argument below), and define

$$f_j\big(x_{1,j},\ldots,x_{N,j}\big) \;=\; \frac{\sum_{i=1}^N x_{i,j}}{2^{n+1}}.$$

Observe that $\sum_{i=1}^N x_{i,j}$ is minimally $N\,((j-1)\,2^n+1)$ and maximally $N\,j\,2^n$. Hence, the image of f_j is contained within

$$\left[\frac{N\big((j-1)2^n+1\big)}{2^{n+1}}, \frac{Nj2^n}{2^{n+1}}\right] \\ = \;\left[\frac{N(j-1)}{2}+\frac{1}{2^{n+1}}, \frac{Nj}{2}\right].$$

Therefore, for $j \geq 1$, each instance f_j is structurally the same "shifted" problem, but crucially *non-overlapping* for each $j \in \mathbb{N}$. So it suffices to show that for each j, each instance *individually* saturates the $\Omega\left(N^2\,\log(1/\varepsilon_j)\right)$ bit lower bound, which we will do now:

Two-Agent Subproblem for N Agents. Because *all* agents must $\langle\varepsilon_j,\delta_j\rangle$-agree on the value of f_j, it follows that in particular, every pair of agents (say (i,k)) must have expectations of f_j that differ by at most ε_j with probability at least $1-\delta_j$. But for any fixed pair (i,k), we can treat $(x_{i,j},x_{k,j})$ as a two-agent input in which all other coordinates $x_{\ell,j}$ for $\ell \neq i,k$ are "known" from the perspective of these two, or do not affect the difficulty except to shift the sum[8]. Hence, for each j and each pair (i,k), there is a two-agent subproblem. We claim that these two agents alone already face a lower bound of $\Omega\left(\log(1/\varepsilon_j)\right)$ bits of communication to achieve $\langle\varepsilon_j,\delta_j\rangle$-agreement on f_j.

Suppose agent k sends only $t < \log_2\left(\frac{1-\delta_j}{\varepsilon_j}\right)$ bits to agent i about its input $x_{k,j}$. Label the 2^t possible message sequences by $m=1,\ldots,2^t$, with probability p_j^m each. Since $x_{k,j}$ is uniform in an interval of size 2^n, then conditioned on message m, there remain at least $2^n p_j^m$ possible values of $x_{k,j}$. Each unit change in $x_{k,j}$ shifts f_j by $1/2^{n+1}$, so even if agent i's estimate is optimal, the fraction of $x_{k,j}$ values producing $|E_j^{k,t}-E_j^{i,t}| \leq \varepsilon_j$ is at most

$$\frac{2^{n+1}\,\varepsilon_j}{2^n\,p_j^m} \;=\; \frac{2\,\varepsilon_j}{p_j^m}.$$

Hence, the total probability of agreement (over all messages m) is bounded by

$$\sum_{m=1}^{2^t} p_j^m \;\cdot\; \frac{2\,\varepsilon_j}{p_j^m} = 2\,\varepsilon_j\,2^t.$$

If $2\,\varepsilon_j\,2^t < 1-\delta_j$, the agents fail to $\langle\varepsilon_j,\delta_j\rangle$-agree. Equivalently, $t \geq \log_2\left(\frac{1-\delta_j}{2\,\varepsilon_j}\right)$. Since every pair (i,k) needs $\Omega\left(\log\left(1/\varepsilon_j\right)\right)$ bits for each of the M tasks, and there are $\binom{N}{2}=\Theta(N^2)$ pairs, the total cost is

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

where $\varepsilon := \min_{j\in[M]}\varepsilon_j$, corresponding to the most "stringent" task j. □

D.2 Proof of Proposition 2

Proof. We divide the $M \geq 2$ tasks into two types of payoff functions f_j as follows, each covering the first $\lfloor M/2\rfloor$ tasks and the last set of $\lceil M/2\rceil$ tasks, respectively:

Type I Tasks. For the first set of $\lfloor M/2\rfloor$ tasks, we let the state space be $S_j := \{s_{(j-1)D}, s_{(j-1)D+1},\ldots,s_{jD-1}\}$. Let the k-th element of S_j be denoted as $s_{k,j} := s_{(j-1)D+k}$. Next, choose a sign vector $b_j \in \{+1,-1\}^N$ with $N/2$ plus signs, and set each element of it, b_j^i, as follows to define the prior distributions for some $0 < p \leq \frac{1}{2}-\frac{\nu}{4}$ as:

$$\beta := \frac{p}{D-2}, \qquad \mathbb{P}_j^i(s_{0,j}) = \tfrac{1}{2}-\tfrac{b_j^i\nu}{4}-\tfrac{(D-2)\beta}{2},$$

$$\mathbb{P}_j^i(s_{1,j}) = \tfrac{1}{2}+\tfrac{b_j^i\nu}{4}-\tfrac{(D-2)\beta}{2},\ \mathbb{P}_j^i(s_{k,j}) = \beta\ (k\geq 2).$$

If $b_j^i \neq b_j^k$ the agent pair (i,k) has L1 distance ν, and therefore prior distance $\geq \nu$ by definition.

Let $T_j(\omega_j)$ denote the number of bits exchanged on task j when the initial world state is $\omega_j \in S_j$ and the agents follow some message-passing protocol. We consider the expectation $\mathbb{E}[T_j] := \mathbb{E}_{\omega_j}[T_j(\omega_j)]$ over all initial world states $\omega_j \in S_j$ with respect to the hard prior distributions specified above. Note that for the purpose of a lower bound, we only need to consider the mismatched agent pairs, since for non-mismatched agents, $T_j \geq 0$, trivially.

Given a task index j and a mismatched agent pair (i,k), let W_j^t denote the total variation distance between the agents' posterior distributions, $\tau_j^{i,t}$ and $\tau_j^{k,t}$, at time t:

$$W_j^t := \tfrac{1}{2}\big\|\tau_j^{i,t}-\tau_j^{k,t}\big\|_1, \qquad W_j^0 = \nu/2.$$

Define the "good" event for task j as

$$G_j := \big|\,E_j^{i,T}-E_j^{k,T}\big| \leq \varepsilon,$$

which holds with probability at least $1-\delta_j$ by the $\langle\varepsilon_j,\delta_j\rangle$-agreement condition. Conditioned on G_j, our assumption implies $W_j^{T_j} \leq c\nu$ for $c < \frac{1}{2}-\frac{\delta_j}{\nu}$; on $\overline{G_j}$ we only know the trivial upper bound on total variation of $W_j^{T_j} \leq 1$. Hence,

$$\mathbb{E}\left[W_j^{T_j}\right] < (1-\delta_j)\cdot c\nu + \delta_j\cdot 1 < c\nu+\delta_j < \nu/2.$$

[7] One could encode them as binary strings of length at least $n+\lceil\log_2 j\rceil$, but in this proof we do not need the explicit binary representation: the integer *range sizes* themselves suffice to carry out the communication complexity lower bound.

[8] Equivalently, imagine the other $N-2$ agents are "dummy" participants, and we fix their inputs from the perspective of the (i,k) pair.

Since we have that:

$$T_j = \sum_{t=0}^{T_j-1} 1 \geq \sum_{t=0}^{T_j-1} (W_j^t - W_j^{t+1}) = W_j^0 - W_j^{T_j},$$

by telescoping. Thus, $\mathbb{E}[T_j] \geq W_j^0 - \mathbb{E}[W_j^T] = \Omega(\nu)$, since $\delta_j < \nu/2$. Hence, each mismatched pair pays $\Omega(\nu)$ bits on task j in expectation. By a pigeonhole argument, for every initially mismatched agent pair, there exists an initial world state $\omega_j \in S_j$ that attains at least that length transcript length T_j; this is the worst-case for task j. With $\Theta(N^2)$ such agent pairs, the mismatch cost per task is $\Omega\left(N^2\nu\right)$, giving a total worst case bit cost of $\Omega\left(M\,N^2\,\nu\right)$ across the first set of $\lfloor M/2 \rfloor$ tasks.

Type II Tasks. For the remaining set of $\lceil M/2 \rceil$ tasks, we use the hard instance f_j (and its corresponding N-tuple state space) in Proposition 1 for each task, with a uniform common prior.

Thus, any deterministic transcript that $\langle \varepsilon_j, \delta_j \rangle$-agrees on *every* task must concatenate M independent sub-transcripts across the Type I and Type II tasks, giving the final

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

bit lower bound. □

D.3 Proof of Proposition 3

Proof. We split the M tasks exactly as in Proposition 2: **Type I:** first $\lfloor M/2 \rfloor$ tasks, and **Type II:** remaining $\lceil M/2 \rceil$ tasks.

Only Type I needs modification; Type II reuses Proposition 1 verbatim and costs $\Omega\left(N^2 \log(1/\varepsilon)\right)$ bits per task.

We consider the following "uniform-slope" hard priors for the Type I tasks: We use the same state-space as in Proposition 2's Type I tasks, namely $S_j := \{s_{(j-1)D}, s_{(j-1)D+1}, \ldots, s_{jD-1}\}$. For notational convenience, fix such a task j and relabel its $D := D_j$ states $S_j = \{s_0, \ldots, s_{D-1}\}$, and therefore drop the j subscript.

The priors are defined as follows for agents i and k:

$$\mathbb{P}^i(s_m) = \frac{\lambda^m}{S}, \quad \mathbb{P}^k(s_m) = \frac{\lambda^{-m}}{S'},$$
$$S = \sum_{q=0}^{D-1} \lambda^q, \quad S' = \sum_{q=0}^{D-1} \lambda^{-q}, \quad \lambda := \frac{1+\nu/2}{1-\nu/2} > 1.$$

We now show the prior distance for any agent pair (i,k) is $\geq \nu$. By definition of prior distance, the triangle inequality shows that $\|\mathbb{P}^i - \mathbb{P}^k\|_1$ lower bounds the prior distance (for set-valued priors per agent, and for single prior distributions per agent it holds with equality), so it suffices to show that $\|\mathbb{P}^i - \mathbb{P}^k\|_1 \geq \nu$. We have that

$$\begin{aligned}
\|\mathbb{P}^i - \mathbb{P}^k\|_1 &:= 2 \sum_{m:\,\mathbb{P}^i(s_m) > \mathbb{P}^k(s_m)} \left(\mathbb{P}^i(s_m) - \mathbb{P}^k(s_m)\right) \\
&\geq 2\left|\mathbb{P}^i(s_0) - \mathbb{P}^k(s_0)\right| \\
&= 2\left|\frac{1}{\sum_{q=0}^{D-1}\lambda^q} - \frac{1}{\sum_{q=0}^{D-1}\lambda^{-q}}\right| = \frac{2(\lambda-1)(\lambda^{D-1}-1)}{\lambda^D - 1} \\
&\geq \frac{2(\lambda-1)}{1+\lambda} = \nu,
\end{aligned}$$

where the last inequality follows from the fact that $\lambda^{D-1} - \lambda \geq 0$ since $\lambda > 1$ and $D \geq 2$, and the last equality directly follows from the definition of λ.

Canonical chain gap at $t = 0$. Connectedness of the initial profile implies that for any two states there exists at least one alternating chain of states (Hellman and Samet 2012, Proposition 2). In particular, the pair (s_0, s_{D-1}) used in the hard prior is linked by a chain $c^\star = (s_0, s_1, \ldots, s_{D-1})$; tightness makes this chain unique and ensures it visits every state exactly once. We let $\tau^{i,t}$ denote agent i's posterior distribution at time t, with $\tau^{i,0} \equiv \mathbb{P}^i$. Set

$$L_t := \left|\log \prod_{(s,s')\in c^\star} \frac{\tau^{i,t}(s')}{\tau^{i,t}(s)} - \log \prod_{(s,s')\in c^\star} \frac{\tau^{k,t}(s')}{\tau^{k,t}(s)}\right|.$$

At $t = 0$,

$$\prod_{(s,s')\in c^\star} \frac{\tau^{i,0}(s')}{\tau^{i,0}(s)} = \lambda^{D-1}, \quad \prod_{(s,s')\in c^\star} \frac{\tau^{k,0}(s')}{\tau^{k,0}(s)} = \lambda^{-(D-1)}$$

Thus, we have that

$$\begin{aligned}
L_0 &= 2(D-1)\,|\log\lambda| \\
&= 2(D-1)\,\left|\log\left(1+\tfrac{\nu}{2}\right) - \log\left(1-\tfrac{\nu}{2}\right)\right| \\
&= 2(D-1)\left|\nu + \tfrac{\nu^3}{12} + \ldots\right| = \Theta(D\nu),
\end{aligned} \tag{5}$$

where the second to last equality follows by the standard Taylor expansions of $\log(1+x)$ and $\log(1-x)$, since $\nu/2 \leq 1$.

Per timestep increment. Let the message sent in round t contain b_t bits, and assume the protocol is BBF(β), i.e. for every agent a and all states s, s', the message likelihoods are bounded as such:

$$\beta^{-b_t} \;\leq\; \frac{\Pr[m^{a,t} \mid s, \Pi^{a,t-1}(s)]}{\Pr[m^{a,t} \mid s', \Pi^{a,t-1}(s')]} \;\leq\; \beta^{\,b_t}.$$

Then we will show that the canonical gap satisfies:

$$|L_t - L_{t-1}| \;\leq\; 2b_t \log\beta. \tag{6}$$

To see this, for convenience we denote $q_t^a(s) := \Pr[m^{a,t} \mid s, \Pi^{a,t-1}(s)]$ for the message likelihood at time t.

Bayes' rule then gives, for any states s, s',

$$\frac{\tau^{a,t}(s')}{\tau^{a,t}(s)} \;=\; \frac{\tau^{a,t-1}(s')}{\tau^{a,t-1}(s)} \cdot \frac{q_t^a(s')}{q_t^a(s)}. \tag{7}$$

Next, define:

$$\Xi_t^a = \log \prod_{m=0}^{D-2} \frac{\tau^{a,t}(s_{m+1})}{\tau^{a,t}(s_m)} = \sum_{m=0}^{D-2} \log \frac{\tau^{a,t}(s_{m+1})}{\tau^{a,t}(s_m)} \tag{8}$$

We fix the canonical chain $c^\star$ once for convenience—any path would serve, since we will show that the bound in (6) depends only on the BBF(β) likelihood-ratio condition and is independent of the particular chain selected.

It follows from (7) and (8) that

$$\begin{aligned}\Xi_t^a &= \sum_{m=0}^{D-2}\left[\log\frac{\tau^{a,t-1}(s_{m+1})}{\tau^{a,t-1}(s_m)} + \log\frac{q_t^a(s_{m+1})}{q_t^a(s_m)}\right] \\ &= \underbrace{\sum_{m=0}^{D-2}\log\frac{\tau^{a,t-1}(s_{m+1})}{\tau^{a,t-1}(s_m)}}_{=\ \Xi_{t-1}^a} + \underbrace{\sum_{m=0}^{D-2}\log\frac{q_t^a(s_{m+1})}{q_t^a(s_m)}}_{\text{telescopes}} \\ &= \Xi_{t-1}^a \ + \ \log\frac{q_t^a(s_{D-1})}{q_t^a(s_0)} \ = \ \Xi_{t-1}^a \ + \ \Delta_t^a.\end{aligned}$$

Because $L_t = |\Xi_t^i - \Xi_t^k|$, we have by the triangle inequality:

$$\begin{aligned}|L_t - L_{t-1}| \ &= \ \big|\,|A+B| - |A|\big| \le |B|, \\ A &:= \Xi_{t-1}^i - \Xi_{t-1}^k,\ B := \Delta_t^i - \Delta_t^k.\end{aligned}$$

Hence, $|L_t - L_{t-1}| \le |\Delta_t^i - \Delta_t^k|$.

By BBF(β), each agent alone obeys

$$\left|\Delta_t^a\right| = \left|\log\frac{q_t^a(s_{D-1})}{q_t^a(s_0)}\right| \le b_t \log\beta, \qquad a \in \{i,k\}.$$

Hence, by the triangle inequality,

$$\left|\Delta_t^i - \Delta_t^k\right| \le \left|\Delta_t^i\right| + \left|\Delta_t^k\right| \le 2\, b_t \log\beta,$$

giving rise to the desired inequality (6).

Per task cost. Let B^{agree} be the *total number of bits* exchanged by the time the agents agree, and let B^{cp} be the total bits exchanged when the common prior is reached. Clearly $B^{\text{agree}} \ge B^{\text{cp}}$, so it suffices to lower bound B^{cp}.

Set $B_T := \sum_{t=1}^T b_t$, so B_T is the cumulative bit count up to round T. By Hellman and Samet (2012, Proposition 4) we have $L_T = 0$ once the common prior is attained. Telescoping and (6) then gives

$$\begin{aligned}L_0 - L_T \ &= \ \sum_{t=1}^T\big(L_t - L_{t-1}\big) \ \le \ \sum_{t=1}^T\big|L_t - L_{t-1}\big| \\ &\le \ 2\log\beta\sum_{t=1}^T b_t = 2\log\beta\, B_T.\end{aligned}$$

Hence, by (5), any BBF(β) protocol must transmit at least $\Omega\left(D\nu\right)$ bits before the priors coincide, and therefore at least that many bits before $\langle M, N, \varepsilon, \delta\rangle$-agreement.

Aggregating costs. There are $\Theta(N^2)$ mismatched pairs and $\lfloor M/2\rfloor$ Type I tasks, so the total Type I cost in bits is $\Omega\big(M\,N^2\,D\nu\big)$. Type II contributes $\Omega\big(M\,N^2\log(1/\varepsilon)\big)$ bits, hence the overall lower bound in bits is

$$\Omega\left(MN^2[D\nu + \log(1/\varepsilon)]\right),$$

with constant $1/\log\beta$, completing the proof. □

E Proof of Lemma 2

Proof. As before, let $\{\mathbb{P}_j^i\}^{i\in[N]}$, be the priors of the agents. The "type profile" τ_j^t is the set of the agent's posterior belief distributions over states $s_j \in S_j$ at time t. Thus, at time 0, τ_j^0 will correspond to the prior distributions over the states in the knowledge partition $\Pi_j^{i,0}$. Since for each agent its type profile distribution is constant across the states in its knowledge partition $\Pi_j^{i,t}(s_j)$ (as they are indistinguishable to the agent, by definition), then the total size of the type profile at time t is

$$|\tau_j^t| = \sum_{i=1}^N\left|\Pi_j^{i,t}\right|. \tag{9}$$

We make use of the following result of Hellman and Samet (2012, Proposition 2), restated for our particular setting: Let $\mathcal{C}\left(\{\Pi_j^{i,t}\}^{i\in[N]}\right)$ denote the common knowledge set (finest common coarsening) across the agents' knowledge partitions at time t. If the knowledge partitions reach a total size across the N agents that satisfies:

$$\sum_{i=1}^N\left|\Pi_j^{i,t}\right| = (N-1)D_j + \mathcal{C}\left(\{\Pi_j^{i,t}\}^{i\in[N]}\right), \tag{10}$$

then any type profile τ_j^t over $\{\Pi_j^{i,t}\}^{i\in[N]}$ has a common prior $\mathbb{CP}_j$. Now, note that $\left|\mathcal{C}\left(\{\Pi_j^{i,t}\}^{i\in[N]}\right)\right| \le D_j$ as it forms a partition over the task state space S_j, so the set of singleton sets of each element $s_j \in S_j$ has the most components to saturate the upper bound. Therefore, the desired size $\sum_{i=1}^N\left|\Pi_j^{i,t}\right| \le ND_j$.

Now, starting from an initial type profile $|\tau_j^0|$, the number of proper refinements needed to get to the desired size $\sum_{i=1}^N\left|\Pi_j^{i,t}\right|$ in (10) is given by *at most*:

$$\sum_{i=1}^N\left|\Pi_j^{i,t}\right| - |\tau_j^0| + 1 = O\left(ND_j\right).$$

Thus, since[9] trivially $|\tau_j^0| \ge 0$, then $O(ND_j)$ is the most number of proper refinements we need to ensure there is a common prior with probability 1, by (10). By Lemma 1, this amounts to $O(N^2D_j)$ messages in the worst case. □

F Proof of Proposition 4

Proof. The N agents will communicate with the spanning tree protocol (cf. Lemma 1) for each task $j \in [M]$, but now with discrete, rather than continuous, messages. The discretized protocol is as follows: Let there be a node F_j that is globally accessible to all N agents. This intermediary is allowed its own prior and will see all messages between the agents (but not their inputs). Thus, $\Pi_j^{F_j,0}(s_j) = S_j$ for all states $s_j \in S_j$, and $\left\{\Pi_j^{F_j,t}(s_j)\right\}_{s_j\in S_j}$ *coarsens* the knowledge partitions at time t of the N agents, so all of the agents

[9] For example, for N agents that start with maximally unrefined knowledge partitions, $|\tau_j^0| = \sum_{i=1}^N\left|\Pi_j^{i,t}\right| = \sum_{i=1}^N 1 = N$.

can therefore compute $E_j^{F_j,t}$. When agent i wants to send a message to its neighbor agent k, then agent i sends "High" if $E_j^{i,t} > E_j^{F_j,t} + \varepsilon_j/4$, "Low" if $E_j^{i,t} < E_j^{F_j,t} - \varepsilon_j/4$, and "Medium" if otherwise. After agent i sends its message to agent k, agent k then refines its knowledge partition (and F_j also refines its partition), before agent k sequentially sends its message relative to the current $E^{F_j,t+1}$ to the next agent down the spanning tree. This process of proper refinement is continued until there is a common prior by Lemma 2, which the $N+1$ agents (including F_j) then condition on to reach $\langle M, N+1, \varepsilon, \delta\rangle$-agreement (hence the $N+1$ factor in the first term to ensure there are enough proper refinements between the $N+1$ agents). This generalizes Aaronson (2005, Theorem 6)'s discretized protocol to $N > 2$ agents (and $M > 1$ tasks), which shows that between any *pair* of agents with a common prior, the number of messages needed for them to $\langle \varepsilon_j, \delta_j\rangle$-agree remains *unchanged* from the full protocol, where each pair leverages the intermediary agent F_j. Therefore, by applying the spanning tree construction from Lemma 3, we get the same bound in the discretized case as before of $O(((N+1)g_j^2)/(\delta_j\varepsilon_j)^2)$ messages before the $N+1$ agents *pairwise* $\langle \varepsilon_j, \delta_j\rangle$-agree, and therefore $O(((N+1)^5 g_j^2)/(\delta_j\varepsilon_j)^2)$ messages until all $\binom{N+1}{2}$ pairs of agents *globally* $\langle \varepsilon_j, \delta_j\rangle$-agree, thereby ensuring that the original N agents agree. Following the rest of the proof of our Theorem 1 yields the per-task upper bound of $O\left((N+1)^2 D_j + \frac{(N+1)^7}{\varepsilon_j^2\delta_j^2}\right)$, where we took the worst-case value of $g_j = O(N+1)$. Subsuming lower-order terms in the big-O gives us the stated upper bound.

BBF(3)-compliant extension. Note that this discretized protocol can be made BBF(3)-compliant via the following simple modification: pick any *buffer parameter* $0 < \theta \le \frac{1}{3}$ (setting $\theta = \frac{1}{4}$ suffices) and, after the sender has deterministically selected the bucket $B \in \{\text{High}, \text{Medium}, \text{Low}\}$ according to the thresholds $\pm\varepsilon_j/4$ around $E_j^{F_j,t}$, let the noisy channel transmit the matching 2-bit codeword with probability $1-2\theta$ and each of the two non-matching codewords with probability θ, i.e. $\Pr[m_j^{i,t} \mid s_j, \Pi_j^{i,t-1}(s_j)] = \theta + (1-3\theta)\,\mathbf{1}\{m_j^{i,t} = m_j^{i,t,\star}(B)\}$. Because every codeword now has probability either $1-2\theta$ or θ under every state, the message likelihood ratio is always in the range $\left[\frac{\theta}{(1-2\theta)}, \frac{(1-2\theta)}{\theta}\right]$, so the channel is BBF$(\beta)$ with $\beta = (1-2\theta)/\theta \le 3$ when $\theta \ge \frac{1}{5}$.

We have the agents communicate first, as usual, for $O((N+1)D_j^2)$ messages per task until they reach a common prior and condition on it, by Lemma 2. Thus, the analysis that remains is how many more messages are needed to reach convergence to $\langle M, N+1, \varepsilon, \delta\rangle$-agreement. A round is called *informative* when the sender's bucket is an outer one (High or Low) and the channel outputs the matching codeword; this occurs with probability at least $(1-2\theta)\,\delta/2$, and in that event $E_j^{F_j,t}$ moves by at least $\varepsilon_j/4$, so the potential $\Psi_t := \|E^{F_j,t}\|_2^2$ increases by at least $(\varepsilon_j/4)^2$. Hence, $\mathbb{E}[\Psi_{t+1} - \Psi_t] \ge (1-2\theta)\,\delta_j(\varepsilon_j/4)^2/2$, giving a per-round drift $\kappa_\theta = \Theta\big((1-2\theta)\,\varepsilon_j^2\delta_j\big)$. Define the centered process $Z_t := \Psi_t - \kappa_\theta t$; then $\{Z_t\}$ is a martingale with one-step differences bounded by 2, so the Azuma-Hoeffding inequality can be applied directly to Z_t. Since $\Psi_t \le 1$, the additive-drift (optional-stopping) theorem implies $\mathbb{E}[T] \le O\,(1/\kappa_\theta) = O\left((1-2\theta)^{-1}\varepsilon_j^{-2}\delta_j^{-1}\right)$ for one pair of agents. We write $\delta := \max_{j\in[M]}\delta_j$ and, for the high-probability bound, simply divide this budget evenly: each task gets δ/M and each of its $\binom{N+1}{2} = O(N^2)$ pairs gets $\eta := \delta/(M\binom{N+1}{2})$. Azuma-Hoeffding with this η yields $O\left(\ln(MN^2/\delta)/(1-2\theta)\varepsilon_j^2(\delta/M)\right)$ rounds per pair, so a union bound over all $M\binom{N+1}{2}$ pairs leaves total failure probability $\le \delta$. Plugging the same δ/M everywhere in the multi-task bookkeeping of Proposition 4 gives

$$T = O\Big(MN^2 D + \frac{M^3N^7\ln(MN^2/\delta)}{(1-2\theta)\varepsilon^2\delta}\Big),$$

valid with probability at least $1-\delta$, with $\varepsilon := \min_{j\in[M]}\varepsilon_j$. □

G Proofs of Theorem 2 and Corollary 1

Here we prove both Theorem 2 and Corollary 1. We do this by generalizing Aaronson (2005, §4)'s computational-boundedness treatment from 2 agents to N agents (specifically, $N-q$ agents and q humans that have differing, rather than equal, query costs) and M functions (rather than 1), using a message-passing protocol that combines his smoothed and spanning tree protocols, all *without* the Common Prior Assumption (CPA).

G.1 Message-Passing Protocol

This is the multi-task generalization of Aaronson (2005, §4.1)'s "smoothed standard protocol", additionally extended to the multi-agent setting a spanning tree the agents use to communicate their messages.

Let $b_j = \lceil \log_2(\frac{C}{\varepsilon_j})\rceil$ be a positive integer we will specify later with a specific constant value $C > 0$, in (18). The N computationally bounded agents follow the $\langle M, N, \varepsilon, \delta\rangle$-agreement algorithm (Algorithm 1), passing $O(b_j)$-bit messages according to the following protocol $\mathcal{P}$:

Protocol $\mathcal{P}$ description (for each task $j \in [M]$):

1. **Current posterior expectation.** The *sending agent* $i \in [N]$ has at timestep $t-1$ a real value $E_j^{i,t-1}(s_j) \in [0,1]$, which is its conditional expectation of $f_j \in [0,1]$ given its knowledge partition $\Pi_j^{i,t-1}(s_j)$ and the current task state $s_j \in S_j$. (Recall that $E_j^{i,t-1}(s_j) := \mathbb{E}_{\mathbb{P}_j^i}\left[f_j \mid \Pi_j^{i,t-1}(s_j)\right]$.) The knowledge partition $\Pi_j^{i,t-1}(s_j)$ is formed after updating this expectation using Bayes' rule, after having received the earlier message at time $t-2$.
2. **Draw an integer r_j via a triangular distribution.** Agent i picks an integer offset

$$r_j \in \{-L_j, -L_j+1, \ldots, L_j\}$$

according to a (continuous) triangular distribution $\Delta_{\text{tri}}(\,\cdot\,;\,\alpha_j)$ that places its mass in the *discrete* set of values $\{-L_j,\ldots,L_j\}$, and has effective width $2\alpha_j$. These discrete offsets r_j ensure that the messages will be discrete as well. Concretely,

$$\mathbb{P}[r_j = x] = \frac{L_j - |x|}{\sum_{z=-L_j}^{L_j}\big(L_j - |z|\big)} = \frac{L_j - |x|}{L_j^2}, \qquad x \in \{-L_j,\ldots,L_j\},$$

where L_j is chosen so that $2^{-b_j}L_j = \alpha_j$ to bound the messages, as explained in the next step below. Note that the form above is chosen so that the "peak" of the discretized triangular distribution is at $r_j = 0$. In other words, the form above is maximized in probability when the offset $r_j = 0$ (which means that no noise is added to the messages with the highest probability).

3. **Form the message with noise.** The agent then sets

$$m_j^t\Big(\,\Pi_j^{i,t-1}(s_j)\Big) \;=\; \text{round}\,\Big(E_j^{i,t-1}(s_j)\Big) \;+\; 2^{-b_j}r_j,$$

where $\text{round}\,\Big(E_j^{i,t-1}(s_j)\Big)$ denotes rounding $E_j^{i,t-1}(s_j)$ to the nearest multiple of 2^{-b_j} (thereby keeping it in the $[0,1]$ interval). This ensures that the message m_j^t is itself a multiple of 2^{-b_j} (thereby being encodable in $O(b_j)$ bits), and is offset by $\pm\alpha_j$ from $\text{round}\big(E_j^{i,t-1}(s_j)\big)$, since $|2^{-b_j}r_j| \le 2^{-b_j}L_j = \alpha_j$, by construction. Hence, each $m_j^t \in [-\alpha_j, 1+\alpha_j]$.

4. **Broadcast.** This message $m_j^t\big(\Pi_j^{i,t-1}(s_j)\big)$ is then broadcast (either sequentially or in parallel) to the relevant agents via an edge of the two spanning trees $SP_j^1 \cup SP_j^2$ each of diameter g_j, just as in Lemma 1, who update their knowledge partitions accordingly to Step 1.

G.2 Sampling-Tree Construction and Simulation for Each Task $j \in [M]$

In our framework, each agent logically refines its knowledge partition $\{\Pi_j^{i,t}(s_j)\}_{s_j\in S_j}$ upon seeing a new message (Step 7 of Algorithm 1). However, given that the agents are computationally bounded, while refinement is allowed, the issue is with their belief updating. By Requirement 1, they have no direct ability to sample from the *conditioned* posterior distributions $\tau_j^{i,t} = \mathbb{P}_j^i(\cdot \mid \Pi_j^{i,t}(s_j))$ at run time, in order to compute the expectation in Step 1 of the protocol in §G.1. In other words, they cannot simply call "$\text{Sample}\big(\mathbb{P}_j^i \mid \Pi_j^{i,t}(s_j)\big)$" in a black-box manner. Thus, *before* any messages are exchanged, each agent constructs a sampling tree $\mathcal{T}_j^i$ offline of *unconditional* samples from the priors $\mathbb{P}_j^i$ (which they are able to do by Subroutine 2 in Requirement 1). The idea is to precompute enough unconditional draws so that each new message can be *simulated* via "walking down" the relevant path in the tree that is consistent with the current message history (including that new message), rather than enumerating or sampling from the newly refined partition directly.

That is the intuition. We now explain in detail how each agent can use sampling trees to simulate this protocol in a computationally bounded manner. This follows Aaronson (2005, §4.2)'s approach of dividing the simulation into two phases—*(I) Sampling-Tree Construction* (no communication) and *(II) Message-by-Message Simulation*—but extended here to our multi-task and multi-agent setting.

(I) **Sampling-Tree Construction (General N-Agent Version).** For each task $j \in [M]$, and for each agent $i \in [N]$, that agent i builds a sampling tree $\mathcal{T}_j{}^i$ of height R_j and branching factor B_j. We fix an ordering of the $O(R_j)$ messages for task j (so that we know which agent is "active" at each level). Formally, let $\text{ActiveAgent}_j : \{0,\ldots,R_j-1\} \to [N]$ denote the function specifying which agent sends the message at each round $\ell = 0,\ldots,R_j-1$. For instance, since the spanning tree protocol cycles through the N agents in a consistent order, then $\text{ActiveAgent}_j(0) = i_0$, $\text{ActiveAgent}_j(1) = i_1,\ldots,$ and so on, up to R_j total rounds.

- Let root_j^i be the root node of $\mathcal{T}_j{}^i$. We say that the *level* of the root node is 0, its children are at level 1, grandchildren at level 2, etc., until depth R_j. Each node will be labeled by a sample in task j's state space S_j, drawn from whichever agent's *unconditional* prior distribution $\mathbb{P}_j^k(\cdot)$ is active at that level.
- Concretely, for $\ell = 0,\ldots,R_j-1$:

(a) Let $a_j := \text{ActiveAgent}_j(\ell)$ be the agent whose distribution we want at level ℓ (i.e. the agent who sends the ℓ-th message).

(b) Every node v_j at *level* ℓ is labeled by some previously chosen sample (if $\ell = 0$ and v_j is the root, we can label it trivially or treat i's perspective by a `do-nothing` step).

(c) Each of the B_j children $w \in \text{Children}(v_j)$ at level $\ell+1$ is labeled with a fresh i.i.d. sample drawn from $\mathbb{P}_j^a(\cdot)$, i.e. from the unconditional posterior of agent a_j.

(d) We continue until level R_j is reached, yielding a total of

$$B_j + B_j^2 + \cdots + B_j^{R_j} = \frac{B_j{}^{R_j+1}-1}{B_j-1} - 1 = O\left(B_j^{R_j}\right) \tag{11}$$

newly drawn states from the unconditional prior distributions $\{\mathbb{P}_j^k\}_{k\in[N]}$ at the appropriate levels.

Thus, node labels alternate among the N agents' unconditional draws, depending on which agent is active at each level (timepoint in message history) ℓ in the eventual message sequence for task j. All of this is done *offline* by *each* agent, requiring no communication among the agents. Once constructed, each agent i holds $\mathcal{T}_j^i$ for personal use.

(II) **Message-by-Message Simulation.** After building these sampling-trees $\{\mathcal{T}_j^i\}_{i\in[N]}$ (for each task j), the N agents

enact the protocol in §G.1 in real time, *but* whenever an agent i needs to "update its posterior" after receiving a message, it does *not* sample from $\tau_j^{i,t} = \mathbb{P}_j^i(\cdot \mid \text{new messages})$. Instead, it uses the precomputed nodes in $\mathcal{T}_j^i$ as follows:

- **Initial estimate.** At time $t = 0$, agent i approximates $E_j^{i,0}(s_j)$ (its prior-based expectation of f_j) by an empirical average of the B_j children of the root node $\text{root}_j^i \in \mathcal{T}_j^i$. This becomes agent i's *initial posterior* in the *offline* sense.
- **At each round** $t = 1, \ldots, R_j$**:**

(a) The agent who is "active" (i.e. is about to send the t-th message) consults *its* sampling-tree, summing over the relevant subtree that corresponds to the newly received messages $m_j^1, \ldots, m_j^{t-1}$, so as to approximate $E_j^{i,t-1}(s_j)$.

(b) It picks an integer offset $r_j \in \{-L_j, \ldots, L_j\}$ via the discrete triangular distribution (defined in §G.1), and then sends the t-th message:

$$m_j^t = \text{round}\big(\langle E_j^{i,t-1}(\cdot)\rangle_i\big) + 2^{-b_j}\, r_j.$$

(c) The other agents, upon receiving m_j^t via the spanning tree protocol, update node-weights in their own sampling trees (via the Δ-update equations in (Aaronson 2005, §4.2)). This effectively "follows" the branch in their sampling trees consistent with m_j^t so they approximate $\Pi_j^{i,t}(\cdot)$ *without* enumerating or sampling from the *conditioned* distribution.

After all R_j messages for task j, the agents will have approximated the ideal protocol in §G.1 with high probability (assuming B_j was chosen large enough, which we will give an explicit value for below in the large-deviation analysis in §G.3).

Lemma 4 (Sampling Tree Time Complexity)**.** *For each task j, the time complexity of the sampling tree for all N agents is*

$$O\left(B_j^{R_j} T_{N,q}\right). \tag{12}$$

Proof. As described, we now assume there are N agents, among which q are humans and $N - q$ are AI agents, each potentially taking different times for evaluating $f_j(\cdot)$ or sampling from the priors (Subroutines 1 and 2 of Requirement 1, respectively):

$$T_{\text{eval},H}, \quad T_{\text{eval},AI}, \quad T_{\text{sample},H}, \quad T_{\text{sample},AI}.$$

Specifically, a *human* agent $i \in H$ uses time $T_{\text{eval},H}$ to evaluate $f_j(s_j)$ for a state $s_j \in S_j$, and time $T_{\text{sample},H}$ to sample from *another* agent's prior distribution $\mathbb{P}_j^k(\cdot)$ unconditionally; whereas an *AI* agent $i \in AI$ spends $T_{\text{eval},AI}$ and $T_{\text{sample},AI}$ on the same operations.

Sampling Overhead. By (11), we do $O\left(B_j^{R_j}\right)$ unconditional draws in total *per agent*. Each draw is performed by the *agent building the tree*, but it might sample from *another* agent's distribution. Hence the time cost per sample is

$$\begin{cases} T_{\text{sample},H} & \text{if the sampling agent is human,} \\ T_{\text{sample},AI} & \text{if the sampling agent is an AI.} \end{cases}$$

We separate the $O\left(q\, B_j^{R_j}\right)$ samples by humans vs. $O\left((N-q)\, B_j^{R_j}\right)$ samples by the AI agents, yielding

$$O\left(q\, B_j^{R_j}\, T_{\text{sample},H} + (N-q)\, B_j^{R_j}\, T_{\text{sample},AI}\right).$$

Evaluation Overhead. Each sampled state $s_j \in S_j$ may require computing $f_j(s_j)$. Again, if the *labeling* agent is a human, that cost is $T_{\text{eval},H}$, whereas if an AI agent is performing the labeling, it is $T_{\text{eval},AI}$. Consequently,

$$O\left(q\, B_j^{R_j}\, T_{\text{eval},H} + (N-q)\, B_j^{R_j}\, T_{\text{eval},AI}\right)$$

suffices to bound all function evaluations across all task j sampling-trees $\{\mathcal{T}_j^i\}_{i\in[N]}$.

During the Actual R_j Rounds. Once messages start flowing, each round requires partial sums or lookups in the prebuilt tree $\mathcal{T}_j^i$. If agent i is a human in that round, each node update in the subtree will cost either $T_{\text{eval},H}$ or $T_{\text{sample},H}$ (though typically we do not *resample*, so it may be just function evaluations or small indexing). Since the total offline overhead still dominates, summing over R_j rounds still yields an overall $O\left(B_j^{R_j}\,(T_{\text{sample},\cdot} + T_{\text{eval},\cdot})\right)$ bound, substituting the index {H,AI} depending on the category of the agent.

Summing the above together gives us the stated upper bound in (12). □

G.3 Runtime Analysis

Recall that the N agents follow the algorithm for $\langle M, N, \varepsilon, \delta\rangle$-agreement (Algorithm 1), which is a "meta-procedure" that takes in any message protocol we have discussed thus far (specifically, the one above). We now want the agents to communicate with enough rounds R_j such that they can feasibly construct a common prior consistent with their current beliefs, with high probability (Step 9 of Algorithm 1). We now have to bound R_j.

Recall that in the sampling-tree scenario, agents now no longer have *exact* access to each other's posterior distributions, but rather approximate them by offline sampling—thus they cannot do a *direct, immediate* conditional update (as they could in the unbounded case). Indeed, triangular noise does *not* strictly *mask* a surprising message; rather, each agent still *can* receive an improbable message from its vantage. However, because these posteriors are only approximate, we must invoke large-deviation bounds to ensure that with high probability, a message deemed γ-likely by the sender but $\leq \gamma/2$ by the *neighboring* receiver *actually* appears within some number of messages (in other words, they disagree), forcing a proper refinement in their knowledge partitions. We rely on a probabilistic argument that surprises still occur on a similar timescale as in the exact setting, with high probability, thereby prompting a partition refinement:

Lemma 5 (Number of Messages for One Refinement Under Sampling Trees)**.** *Suppose two* neighboring *agents i and k have not yet reached $\langle M, N, \varepsilon, \delta\rangle$-agreement on a given task $j \in [M]$. Therefore, they disagree on some message m_j^* as follows:*

$$\Pr_{\substack{s_j\sim\mathbb{P}_j^i \\ r_j\sim\Delta_{\mathrm{tri}}(\alpha_j)}} \Big[\mathrm{round}\big(E_j^{i,t-1}(s_j)\big) + 2^{-b_j} r_j = m_j^*\Big] \;\geq\; \gamma, \tag{13}$$

namely, Agent i regards m_j^* as γ-likely.

$$\Pr_{\substack{s'_j\sim\mathbb{P}_j^k \\ r'_j\sim\Delta_{\mathrm{tri}}(\alpha_j)}} \Big[\mathrm{round}\big(E_j^{k,t-1}(s'_j)\big) + 2^{-b_j} r'_j = m_j^*\Big] \;\leq\; \frac{\gamma}{2}, \tag{14}$$

namely, Agent k sees m_j^* with probability at most $\gamma/2$. *Then the probability of m_j^* failing to be produced in $T = O\left(\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ consecutive rounds is at most μ_j. In other words, the neighboring agents will disagree (thereby triggering at least one proper refinement, namely, for agent k) with probability at least $1-\mu_j$ after $T = O\left(\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ consecutive rounds, for $\mu_j > 0$.*

Proof. Note that (13) ensures m_j^* has at least probability γ from the *sender's* perspective, so we can bound how quickly m_j^* is produced. Specifically, by Aaronson (2005, Lemma 15), we know that solely from the sender's vantage (and therefore not assuming a common prior), the probability of m_j^* *not* appearing after T consecutive rounds is given by at most

$$\lambda_j^{T/2} \max\Big\{\gamma, \tfrac{1}{B_j}\Big\},$$

where $\lambda_j := \frac{4\,e}{\alpha_j}\ln(B_j)$. Therefore, it suffices for T to be such that $\lambda_j^{T/2}\max\{\gamma, 1/B_j\} \leq \mu_j$. Isolating T gives us:

$$\begin{aligned} \lambda_j^{T/2} \;&\leq\; \frac{\mu_j}{\max\{\gamma,\, 1/B_j\}} \\ \implies \quad \frac{T}{2}\ln\lambda_j \;&\leq\; \ln\mu_j - \ln\big(\max\{\gamma,\, 1/B_j\}\big). \end{aligned}$$

Hence

$$\begin{aligned} T &\leq \frac{2\big[\ln\mu_j - \ln\big(\max\{\gamma,\, 1/B_j\}\big)\big]}{\ln\lambda_j} \\ &= \frac{2\big[\ln\mu_j + \ln\big(1/\max\{\gamma,\, 1/B_j\}\big)\big]}{\ln(4e) + \ln(1/\alpha_j) + \ln\ln B_j}. \end{aligned}$$

As γ is a free parameter, we can obtain a cleaner (albeit looser) bound on T by subsuming suitably large constants. A natural choice for γ is $\gamma = \alpha_j$, since the added noise to each message is at most $\alpha_j < 1/40$ (via Step 3 in §G.1) at each round. Therefore, the maximum additive noise is also the natural threshold for a "surprising" message from the receiver's (agent k) point of view. We will also choose the branching factor B_j to be sufficiently large enough that $1/B_j \leq \alpha_j$. Thus, $\max\{\gamma, 1/B_j\} = \alpha_j$. Hence, for $B_j \geq 1/\alpha_j$, trivially $\ln\ln(B_j) > 0$, and we have that

$$\begin{aligned} T \;&\leq\; \frac{2\big[\ln\mu_j + \ln(1/\alpha_j)\big]}{\ln(4e) + \ln(1/\alpha_j) + \ln\ln B_j} \\ &\leq\; \frac{2\big[\ln\mu_j + \ln(1/\alpha_j)\big]}{\ln(1/\alpha_j)} \\ &=\; O\bigg(\frac{\ln\mu_j}{\ln(1/\alpha_j)}\bigg). \end{aligned}$$

□

We are now finally in a position to bound R_j.

Lemma 6 (Common Prior Lemma Under Sampling Trees)**.** *Suppose the N agents have not yet reached $\langle M, N, \varepsilon, \delta\rangle$-agreement on a given task $j \in [M]$. They will reach a common prior $\mathbb{CP}_j$ with probability at least $1-\delta_j$, after $R_j = O\left(N^2 D_j \frac{\ln(\delta_j/(N^2 D_j))}{\ln(1/\alpha_j)}\right)$ rounds.*

Proof. We recall from Lemma 1 that, *if* every agent had exact access to its posterior distribution, any block of $O(g_j)$ messages in the two-spanning-trees ordering would pass each agent's "current expectation" to every other agent precisely once, guaranteeing that if a large disagreement persists, the *receiving* agent sees an unlikely message and refines its partition.

In the *sampled* (bounded) setting, each agent i approximates its posterior by building an offline tree of $O\left(B_j^{R_j}\right)$ unconditional samples, labeling its nodes by unconditional draws from the respective priors $\{\mathbb{P}_j^k(\cdot)\}$. Then, when agent i must send a message—nominally its exact posterior—it instead *looks up* that value via partial averages in its sampling-tree. This is done in a manner consistent with message alternation. Specifically, in each round, the $\mathrm{ActiveAgent}_j$ function from §G.2 ensures that every agent's approximate expectation is routed along SP_j^1 and SP_j^2 (Step 4 of §G.1) once in every block of $O(g_j)$ messages.

Now, in Lemma 5, we assumed the agents were neighbors. Thus, in the more general case of N agents, who communicate with spanning trees $SP_j^1 \cup SP_j^2$ each of diameter g_j, if there is a "large disagreement" between some agent pair (i,k), then from $(i \to k)$ or $(k \to i)$'s vantage, the other's message is improbable. In the worst case, the agents are $O(g_j)$ hops apart. Hence, once that message arrives in these $O\left(g_j\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right) = O\left(N\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ transmissions, the probability that the receiving agent still sees it as a *surprise* and properly refines its partition is $\geq 1 - g_j\mu_j \geq 1 - N\mu_j$, by a union bound over hops.

These $O\left(N\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ transmissions constitute one "block" of messages that results in at least *one* agents' refinement with high probability. Finally, we will need $O(ND_j)$ refinements by Lemma 2 to reach a common prior. Partition the total timeline into $X = O(ND_j)$ disjoint "blocks," each block of $O\left(g_j\frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ messages. Let E_i

denote the event that "≥ 1 refinement occurs in block i". By the single-block argument, $\mathbb{P}[E_i] \geq 1 - N\mu_j$.

We want $\mathbb{P}\Big[\bigcap_{i=1}^{X} E_i\Big]$, the probability that *all* X blocks yield at least one refinement. Using a union bound on complements,

$$\begin{aligned}\Pr\Big[\bigcap_{i=1}^{X} E_i\Big] &= 1 - \Pr\Big[\bigcup_{i=1}^{X} E_i^{\mathrm{c}}\Big] \\ &\geq 1 - \sum_{i=1}^{X} \Pr\big[E_i^{\mathrm{c}}\big] \\ &\geq 1 - X\,N\,\mu_j \\ &= 1 - N^2 D_j\,\mu_j.\end{aligned}$$

Thus, after $X \times O\left(g_j \frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right) = O\left(N^2 D_j \frac{\ln(\mu_j)}{\ln(1/\alpha_j)}\right)$ rounds, we have converged to a common prior with probability at least $1 - N^2 D_j \mu_j$. Setting $\delta_j := N^2 D_j \mu_j$ gives us the final result. □

Now that we have established that we can converge with high probability $\geq 1 - \delta_j$ to a state where there is a consistent common prior after $R_j = O\left(N^2 D_j \frac{\ln(\delta_j/(N^2 D_j))}{\ln(1/\alpha_j)}\right)$ rounds, we now introduce an explicit algorithm for the efficient searching of the belief states of the agents. This is given by Algorithm 2, and finds a common prior via linear programming feasibility of the Bayesian posterior consistency ratios across states.

We first give proofs of correctness (Lemma 7) and runtime (Lemma 8) in the exact setting, before generalizing it to the inexact sampling tree setting (Lemma 9).

Lemma 7 (Correctness of "CONSTRUCTCOMMONPRIOR" (Algorithm 2))**.** *Let S_j be a finite state space, and for each agent $i \in [N]$ let $\Pi_j^{i,t}$ be a partition of S_j with corresponding posterior $\tau_j^{i,t}$. Then the algorithm* CONSTRUCTCOMMONPRIOR *returns a probability distribution p_j on S_j that is a Bayes-consistent common prior for all $\tau_j^{i,t}$ if and only if such a (possibly different) common prior $\mathbb{CP}_j$ exists in principle.*

Proof. **(⟹)** Suppose first that there is some common prior $\mathbb{CP}_j$ over S_j satisfying

$$\mathbb{CP}_j\big(s_j \mid C_{j,k}^{i,t}\big) = \tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big) \quad \forall\, i \in [N],\ s_j \in C_{j,k}^{i,t}.$$

That is, the common prior $\mathbb{CP}_j$ agrees with every agent's posterior on every state in each cell $C_{j,k}^{i,t}$.

In particular, for $s_j, s_j' \in C_{j,k}^{i,t}$,

$$\frac{\mathbb{CP}_j(s_j)}{\mathbb{CP}_j(s_j')} = \frac{\tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big)}{\tau_j^{i,t}\big(s_j' \mid C_{j,k}^{i,t}\big)}.$$

Since the meet partition Π_j^* refines each $\Pi_j^{i,t}$, those ratio constraints must also hold on every meet-cell $Z_\alpha \subseteq C_{j,k}^{i,t}$.

Algorithm 2: CONSTRUCTCOMMONPRIOR$\big(\{\Pi_j^{i,t}, \tau_j^{i,t}\}_{i=1}^{N}\big)$

Require: Finite state-space S_j of size D_j; for each agent $i \in [N]$: partition $\Pi_j^{i,t} = \{C_{j,k}^{i,t}\}_k$ and posterior $\tau_j^{i,t}(\,\cdot \mid C_{j,k}^{i,t})$.

Ensure: Either a distribution p_j on S_j Bayes-consistent with all $\tau_j^{i,t}$, or INFEASIBLE.

1: $\Pi_j^* \leftarrow \bigwedge_{i=1}^{N} \Pi_j^{i,t}$ /* intersections */
2: Let $\Pi_j^* = \{Z_1, \ldots, Z_r\}$, each $Z_\alpha \subseteq S_j$.
3: **for** $\alpha \leftarrow 1$ **to** r **do**
4: create LP variable $p_j(Z_\alpha) \geq 0$, where $p_j(Z_\alpha) := \sum_{s_j \in Z_\alpha} p_j(s_j)$ /* prior mass */
5: **end for**
6: **for all** agent $i \in [N]$ **and** cell $C_{j,k}^{i,t} \in \Pi_j^{i,t}$ **do**
7: **for all** $Z_\alpha \subseteq C_{j,k}^{i,t}$ **do**
8: **for all** $s_j, s_j' \in Z_\alpha$ **do**
9: add constraint

$$\frac{p_j(s_j)}{p_j(s_j')} = \frac{\tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big)}{\tau_j^{i,t}\big(s_j' \mid C_{j,k}^{i,t}\big)}$$

10: **end for**
11: **end for**
12: **end for**
13: add normalization constraint $\sum_{\alpha=1}^{r} p_j(Z_\alpha) = 1$.
14: solve the resulting LP for feasibility
15: **if** a non-negative solution $\{p_j(Z_\alpha)\}$ exists **then**
16: reconstruct p_j on each $s_j \in Z_\alpha$ via the ratio constraints
17: **return** p_j
18: **else**
19: **return** INFEASIBLE /* no single prior fits all posteriors */
20: **end if**

Hence, if we introduce variables for $p_j(Z_\alpha)$ and enforce these pairwise ratio constraints (Algorithm 2, Step 6), the distribution $\mathbb{CP}_j$ serves as a *feasible solution* to that linear system. Furthermore, the normalization (Step 13) is satisfied by $\mathbb{CP}_j$. Consequently, the algorithm will return some valid p_j.

(⟸) Conversely, if the algorithm's LP is feasible and yields $\big\{p_j(Z_\alpha)\big\}_{\alpha=1}^{r}$ with $\sum_{\alpha=1}^{r} p_j(Z_\alpha) = 1$, then for any agent i and cell $C_{j,k}^{i,t} \in \Pi_j^{i,t}$, the meet-cells $Z_\alpha \subseteq C_{j,k}^{i,t}$ satisfy

$$\frac{p_j(s_j)}{p_j(s_j')} = \frac{\tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big)}{\tau_j^{i,t}\big(s_j' \mid C_{j,k}^{i,t}\big)} \quad \text{for all } s_j, s_j' \in Z_\alpha.$$

Define for each state $s_j \in Z_\alpha$,

$$p_j(s_j) = p_j\big(Z_\alpha\big) \cdot \tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big).$$

This p_j is a proper distribution over S_j (since the $p_j(Z_\alpha)$ sum to 1). Moreover,

$$p_j\big(s_j \mid C_{j,k}^{i,t}\big) = \tau_j^{i,t}\big(s_j \mid C_{j,k}^{i,t}\big),$$

so p_j is indeed a common prior that *matches* each agent's posterior distribution.

Remark on the "true" prior vs. the algorithm's output. If there were a "true" common prior $\mathbb{CP}_j$, the distribution p_j returned by the algorithm need not coincide with $\mathbb{CP}_j$ numerically; many distributions can satisfy the same ratio constraints in each cell. But from every agent's viewpoint, p_j and $\mathbb{CP}_j$ induce the same posterior on all cells, and thus are equally valid as a Bayes-consistent common prior.

Hence ConstructCommonPrior returns a valid common prior if and only if one exists. □

Lemma 8 (Runtime of finding a common prior)**.** *Given posteriors for N agents over a state space S_j of size D_j, a consistent common prior can be found in $O\Big(\mathrm{poly}(N\,D_j^2)\Big)$ time.*

Proof. Each agent's partition $\Pi_j^{i,t}$ divides S_j into at most D_j cells, so the meet partition

$$\Pi_j^* = \bigwedge_{i=1}^{N} \Pi_j^{i,t},$$

has at most D_j nonempty blocks. Labeling each of the D_j states with its N-tuple of cell indices takes $O(N)$ time per state, and grouping (using hashing or sorting) can be done in $O(D_j)$ or $O(D_j \log D_j)$ time. Hence, constructing Π_j^* takes $\tilde{O}(N\,D_j)$ time (the $\tilde{O}$ subsumes polylogarithmic factors).

Next, for each agent $i \in [N]$ and for each cell $C_{j,k}^{i,t}$ in $\Pi_j^{i,t}$, we impose pairwise ratio constraints over states in the same meet-cell contained in $C_{j,k}^{i,t}$. In the worst case, an agent's cell may contain all D_j states, leading to $\binom{D_j}{2} = \Theta(D_j^2)$ pairwise constraints for that agent. Summing over N agents gives a total of $O(N\,D_j^2)$ constraints.

Standard LP solvers run in time polynomial in the number of variables (at most D_j) and constraints ($O(N\,D_j^2)$), plus the bit-size of the numerical data. Therefore, the overall runtime is $O\Big(\mathrm{poly}(N\,D_j^2)\Big)$. □

Having proven the runtime and correctness in the *exact* case, we now turn to bounding the runtime in the inexact sampling tree setting. From here on out, we will define

$$T_{N,q,\mathrm{sample}} := q\,T_{\mathrm{sample},H} + (N-q)T_{\mathrm{sample},AI}. \quad (15)$$

Lemma 9 (Approximate Common Prior via Sampling Trees)**.** *Assume that each agent approximates its conditional posterior $\tau_j^{i,t}(\cdot \mid C_{j,k}^{i,t})$ using its offline sampling tree $\mathcal{T}_j{}^i$ (of height R_j and branching factor B_j) with per-sample costs $T_{\mathrm{sample},H}$ (for the q human agents) and $T_{\mathrm{sample},AI}$ (for the $N-q$ AI agents). Suppose that for each cell $C_{j,k}^{i,t} \subseteq S_j$ and for any two states $s_j, s_j' \in C_{j,k}^{i,t}$, the ratio*

$$\frac{\tau_j^{i,t}(s_j \mid C_{j,k}^{i,t})}{\tau_j^{i,t}(s_j' \mid C_{j,k}^{i,t})}$$

can be estimated via the sampling tree using

$$S = O\left(\frac{1}{\varepsilon_j^2} \ln \frac{1}{\delta_j}\right)$$

samples per ratio, so that each is accurate within error ε_j with probability at least $1-\delta_j$ (we assume sufficiently large B_j for this, specifically such that ${B_j}^{R_j} \gtrsim S$ for R_j given by Lemma 6). Then the overall time complexity in the sampling-tree setting of ConstructCommonPrior *is*

$$O\left(\frac{1}{\varepsilon_j^2} \ln \frac{N\,D_j^2}{\delta_j} \cdot \mathrm{poly}\left(D_j^2 T_{N,q,\mathrm{sample}}\right)\right).$$

Proof. For each cell $C_{j,k}^{i,t}$ and each state $s_j \in C_{j,k}^{i,t}$, the agent uses its sampling tree $\mathcal{T}_j{}^i$ to compute an empirical estimate $\widehat{\tau}_j^{i,t}(s_j \mid C_{j,k}^{i,t})$ of $\tau_j^{i,t}(s_j \mid C_{j,k}^{i,t})$ by averaging over the leaves of the appropriate subtree. (Recall that the tree is constructed offline by drawing $O\left(B_j^{R_j}\right)$ i.i.d. samples from the unconditional prior of the active agent at each level, defined by the $\mathrm{ActiveAgent}_j$ function.) Thus, for a fixed agent i, cell $C_{j,k}^{i,t} \subseteq S_j$, and state $s_j \in C_{j,k}^{i,t}$, define the i.i.d. indicator random variables $X_1, \ldots, X_m$ by

$$X_\ell = \begin{cases} 1, & \text{if the } \ell\text{-th sampled state equals } s_j, \\ 0, & \text{otherwise.} \end{cases}$$

(Each sample is drawn from the leaves of $\mathcal{T}_j{}^i$ restricted to the cell $C_{j,k}^{i,t}$.) Thus, each X_ℓ indicates whether the ℓ-th sample drawn via the sampling tree lands on the state s_j for agent i. The empirical average

$$\widehat{\tau}_j^{i,t}(s_j \mid C_{j,k}^{i,t}) = \frac{1}{m}\sum_{\ell=1}^{m} X_\ell$$

serves as an estimate for the true probability $\tau_j^{i,t}(s_j \mid C_{j,k}^{i,t})$.

By the "textbook" additive Chernoff-Hoeffding bound, if $m = O\left(\frac{1}{\varepsilon_j^2} \ln \frac{1}{\delta_j'}\right)$, then

$$\Pr\Big[\,\big|\widehat{\tau}_j^{i,t}(s_j \mid C_{j,k}^{i,t}) - \tau_j^{i,t}(s_j \mid C_{j,k}^{i,t})\big| \geq \varepsilon_j\Big] \leq \delta_j'.$$

Similarly for s_j', hence the ratio $\widehat{\tau}_j^{i,t}(s_j \mid C_{j,k}^{i,t})/\widehat{\tau}_j^{i,t}(s_j' \mid C_{j,k}^{i,t})$ differs by at most $\pm\varepsilon_j$ with probability $\geq 1-2\delta_j'$, as it is computed from two such independent estimates of $\widehat{\tau}$. Taking $\delta_j' = \delta_j/(2\,N\,D_j^2)$ and union-bounding over all $O(N\,D_j^2)$ ratios yields a net success probability $\geq 1-\delta_j$.

Finally, each ratio uses $O((1/\varepsilon_j^2)\ln(1/\delta_j'))$ subtree draws, each draw in the outer `for` loop in Step 6 of

Algorithm 2, incurring $T_{\text{sample},H}$ or $T_{\text{sample},AI}$ cost depending on whether a human or AI agent (of the N total agents) built that portion of $\mathcal{T}_j^i$. Hence, we replace the N in Lemma 8's $O\left(\text{poly}\left(N\, D_j^2\right)\right)$ runtime with $q\, T_{\text{sample},H} + (N-q)T_{\text{sample},AI}$, and multiply by the per-ratio sampling overhead of $O((1/\varepsilon_j^2)\ln(1/\delta_j'))$. This gives us the stated runtime. Thus, CONSTRUCTCOMMONPRIOR still returns a valid common prior with probability at least $1-\delta_j$. □

Now we have reached a state where the N agents have found a common prior $\mathbb{CP}_j$ with high probability. In what follows, note that it does not matter if their common prior is approximate, but rather that the N agents can consistently find *some* common distribution to condition on, which is what Lemma 9 guarantees as a consequence. By Subroutine 2 of Requirement 1, all N agents can sample from the unconditional common prior $\mathbb{CP}_j$ once it is found (Step 12 of Algorithm 1), in total time $T_{N,q,\text{sample}}$.

They will do this and then build their *new* sampling trees of height R_j' and branching factor B_j', post common prior. We now want to bound R_j':

Lemma 10. *For all f_j and $\mathbb{CP}_j$, the N agents will globally $\langle\varepsilon_j,\delta_j\rangle$-agree on a given a task $j\in[M]$ after $R_j' = O\left(N^7/(\delta_j\varepsilon_j)^2\right)$ rounds under the message-passing protocol in §G.1.*

Proof. By Aaronson (2005, Theorem 11), when any *two* agents use this protocol under a common prior, it suffices to take $R_j' = O\left(1/(\delta_j\varepsilon_j^2)\right)$ rounds to reach pairwise $\langle\varepsilon_j,\delta_j\rangle$-agreement, which is the same runtime as in the non-noisy two agent case. We just need to generalize this to N agents who share a common prior $\mathbb{CP}_j$. We take the same approach as in our Proposition 4, by having an additional node F_j that is globally accessible to all N agents. The rest of the proof follows their Theorem 11 for any *pair* of agents that use the intermediary F_j, so then by our Lemma 3, under the spanning tree protocol, it will instead require $R_j' = O\left((N+1)^7/(\delta_j\varepsilon_j)^2\right)$ rounds for all $\binom{N+1}{2}$ pairs of $N+1$ agents (including F_j) to reach *global* $\langle\varepsilon_j,\delta_j\rangle$-agreement. We subsume the lower-order terms in the big-O constant, giving rise to the above lemma. □

Thus, combining Lemma 10 with Lemma 4, reaching *global* $\langle\varepsilon_j,\delta_j\rangle$-agreement with sampling trees will take total time:

$$O\left(B_j'^{\,N^7/(\delta_j\varepsilon_j)^2} T_{N,q}\right). \tag{16}$$

Let $1-\delta_j^{\text{find\_CP}}$ be the probability of converging to a common prior by Lemma 6, let $1-\delta_j^{\text{construct\_CP}}$ be the probability of constructing a common prior once reaching convergence by Lemma 9, and let $1-\delta_j^{\text{agree\_CP}}$ be the probability of reaching global $\left\langle\varepsilon_j,\delta_j^{\text{agree\_CP}}\right\rangle$-agreement when conditioned on a constructed common prior, then we have that

$$\begin{aligned}\Pr\left[\varepsilon_j\text{-agreement}\right] \;&\geq\; \left(1-\delta_j^{\text{find\_CP}}\right)\left(1-\delta_j^{\text{construct\_CP}}\right)\\ &\quad\left(1-\delta_j^{\text{agree\_CP}}\right)\\ &\geq\; 1-\left(\delta_j^{\text{find\_CP}}+\delta_j^{\text{construct\_CP}}\right.\\ &\quad\left.+\,\delta_j^{\text{agree\_CP}}\right).\end{aligned} \tag{17}$$

Hereafter, we set $\delta_j := \delta_j^{\text{find\_CP}} + \delta_j^{\text{construct\_CP}} + \delta_j^{\text{agree\_CP}}$.

Thus, for a single task j, sufficiently large B_j', and $B_j \geq \max\{S^{1/R_j}, 1/\alpha_j\}$, we have that with probability $\geq 1-\delta_j$, the N agents will reach ε_j-agreement in time:

$$\begin{aligned}&\widetilde{O}\Big(B_j^{\,N^2 D_j \frac{\ln\left(\delta_j^{\text{find\_CP}}/(N^2 D_j)\right)}{\ln(1/\alpha_j)}}\, T_{N,q}\\ &+\; N^2 D_j\; \text{poly}\big(D_j^2\big)\, T_{N,q,\text{sample}} \;+\; T_{N,q,\text{sample}}\\ &+\; B'^{\,\frac{N^7}{(\delta_j^{\text{agree\_CP}}\varepsilon_j)^2}}\, T_{N,q}\Big)\\ &=\; O\Big(T_{N,q}\Big[B_j^{\,N^2 D_j \frac{\ln\left(\delta_j^{\text{find\_CP}}/(N^2 D_j)\right)}{\ln(1/\alpha_j)}} + B'^{\,\frac{N^7}{(\delta_j^{\text{agree\_CP}}\varepsilon_j)^2}}\Big]\Big).\end{aligned}$$

since the logarithmic terms (from Lemma 9) subsumed by the $\tilde{O}$ are on the non-exponential additive terms. This follows from summing the runtime of computing the sampling tree (Lemma 4), the total runtime of the common prior procedure in Lemma 9 multiplied by the number of online rounds R_j in the `while` loop of Algorithm 1 (Lines 4-16) given by Lemma 6, and the runtimes of conditioning and the second set of sampling trees in (15) and (16), respectively. We then take $D := \max_{j\in[M]} D_j$, $\delta^{\text{find\_CP}} := \max_{j\in[M]}\delta_j^{\text{find\_CP}}$, $\delta^{\text{construct\_CP}} := \max_{j\in[M]}\delta_j^{\text{construct\_CP}}$, $\delta^{\text{agree\_CP}} := \max_{j\in[M]}\delta_j^{\text{agree\_CP}}$, $\alpha := \min_{j\in[M]}\alpha_j$, $\varepsilon := \min_{j\in[M]}\varepsilon_j$. To ensure an overall failure probability of at most δ we allocate the budget δ/M *per task*, splitting it equally among the three sub-routines: $\delta_j^{\text{find\_CP}} = \delta_j^{\text{construct\_CP}} = \delta_j^{\text{agree\_CP}} = \delta/(3M)$. (A union bound over all M tasks and the three sub-routines then bounds the total error by δ.) Finally, let $B := \max_{j\in[M]}\max\{B_j, B_j'\}$ so that a single base subsumes all exponential terms. Multiplying the worst-case per-task time by M tasks yields the global bound of Theorem 2.

To prove Corollary 1, it suffices to explicitly bound B_j'. This is because for each *individual* task j, a "total Bayesian wannabe", *after* conditioning on a common prior $\mathbb{CP}_j$, as defined in Definition 1, corresponds to a single "Bayesian wannabe" in Aaronson (2005, pg. 19)'s sense. Thus, by their Theorem 20, for $\varepsilon_j/50 \leq \alpha_j \leq \varepsilon_j/40$, it suffices to set on a per-task basis,

$$B_j' := O\left((11/\alpha_j)^{R_j'^{\,2}}/\rho_j^2\right), \quad b_j := \left\lceil \log_2 \frac{R_j'}{\rho_j\alpha_j}\right\rceil + 2, \tag{18}$$

where R_j' is the value for N agents that we found in Lemma 10. Taking $\rho := \min_{j\in[M]}\rho_j$ maximizes the bound, and scaling by M gives rise to what is stated in Corollary 1.

H Proof of Proposition 5

Proof. **Hard priors.** For a state space $S = \{s_0, s_1, \ldots, s_{D-1}\}$ $(D \geq 3)$, define

$$\mathbb{P}^i(s_0) = 0, \qquad \mathbb{P}^k(s_0) = \nu/2,$$
$$\mathbb{P}^i(s_m) = \frac{1}{D-1}, \quad \mathbb{P}^k(s_m) = \frac{1-\nu/2}{D-1} \quad (m \geq 1).$$

Then $\|\mathbb{P}^i - \mathbb{P}^k\|_1 = \nu$, which means the prior distance $\geq \nu$ by the triangle inequality.

Payoff and targets. Let $f(s) = \mathbf{1}[s = s_0]$; thus $\mathbb{E}_i[f] = 0,\ \mathbb{E}_k[f] = \nu/2$. Set $\varepsilon = \nu/8,\ \delta = 1/4$.

One agent, one tree. It suffices to consider only agent k. Let the sampling tree have L leaves labeled $s_1, \ldots, s_L$ and define the Bernoulli indicators $X_\ell = \mathbf{1}[s_\ell = s_0]$. The empirical expectation sent in the ***first*** message is

$$\widehat{E}[f] \;=\; \frac{1}{L}\sum_{\ell=1}^{L} X_\ell.$$

If none of the L samples equals s_0, then $\widehat{E}[f] = 0$ and the absolute error is $|\widehat{E}[f] - \mathbb{E}_k[f]| = \nu/2 \geq 4\varepsilon$; the protocol necessarily fails. The probability of that event is

$$\Pr[\text{no } s_0] = (1-\nu/2)^L \geq 1 - (\nu/2)L$$

To make it $\leq \delta = 1/4$, we need $1 - (\nu/2)L \leq 1/4$, then $L \geq 3/(2\nu)$. Because the sampling tree has to be at least of size L, we complete this part.

Many tasks & agents. Embed an independent copy of the hard task in each of M task indices, assign P^i to $N/2$ of the agents and P^k to the other $N/2$ agents, and sum the costs to get $\Omega(M\nu^{-1}\, T_{N,q,\mathrm{sample}})$ time units once the per-sample cost of Requirement 1 is applied. □

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il