AAAI-26 / IAAI-26 / EAAI-26

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

Aran Nayebi
Assistant Professor
Machine Learning Department
Neuroscience Institute (core faculty), Robotics Institute (by courtesy)

AAAI 2026, Special Track on AI Alignment Oral (AIA-196)
2026.01.25

MACHINE LEARNING DEPARTMENT

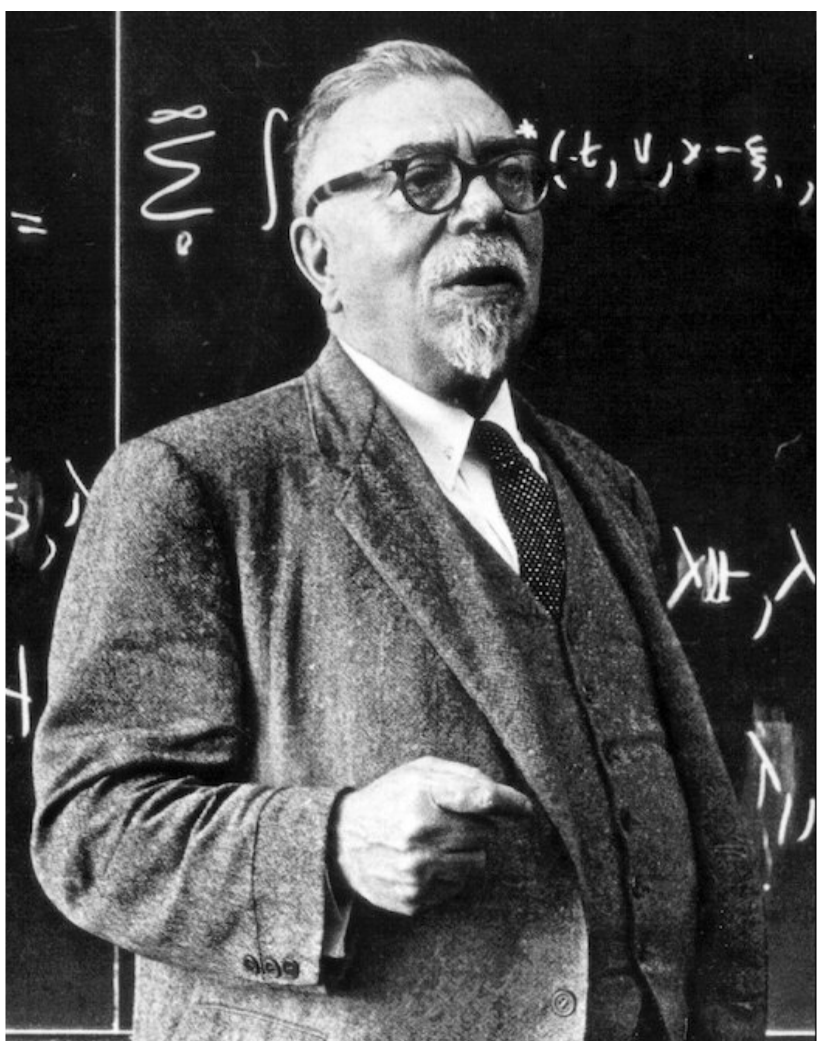
Carnegie Mellon University Neuroscience Institute

THE ROBOTICS INSTITUTE

Carnegie Mellon SCHOOL OF COMPUTER SCIENCE

Alignment Problem

How can we get AI systems to act in accordance with our values?



Some Moral and Technical Consequences of Automation

As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

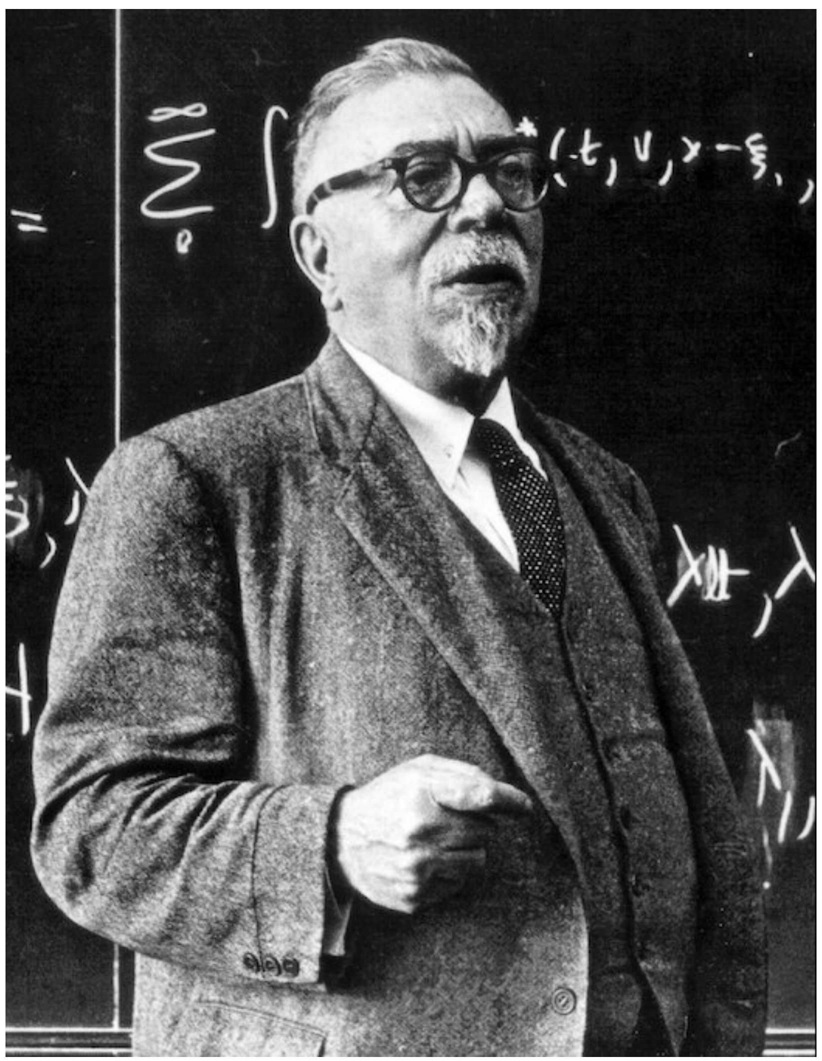
Norbert Wiener

6 MAY 1960

Alignment Problem**s**

How can we get AI systems to act in accordance with our values?

What should those values even *be*?



Some Moral and Technical Consequences of Automation

As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Norbert Wiener

6 MAY 1960

Alignment *Approaches*

How can we get AI systems to act in accordance with our values?

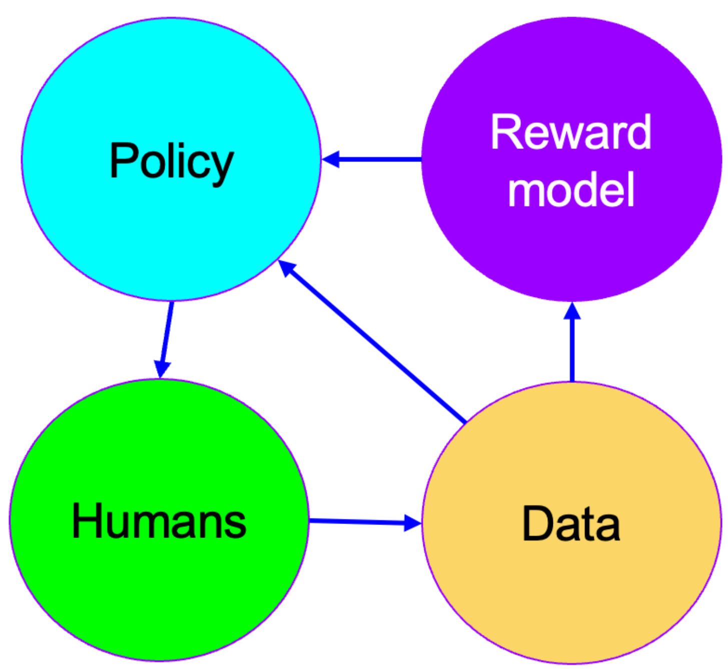
What should those values even *be*?

<u>Current Approaches:</u>

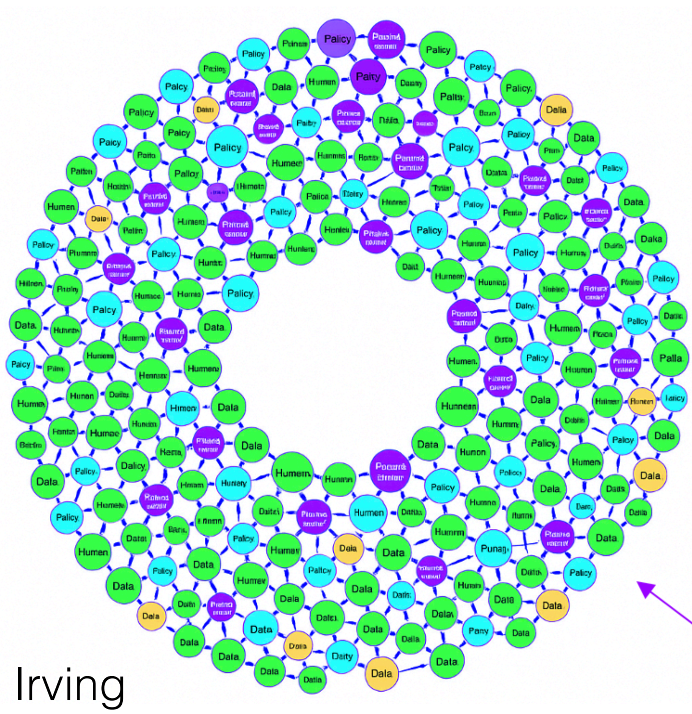
Focused on specific model families (e.g. LLMs) or even specific *features* within particular *models* (e.g. mechanistic interpretability)

AI training has grown in complexity

RLHF in 2019



RLHF in 2025



Live footage

Credit: Geoffrey Irving

Alignment *Approaches*

How can we get AI systems to act in accordance with our values?

What should those values even *be*?

Current Approaches:

Focused on specific model families (e.g. LLMs) or even specific *features* within particular *models* (e.g. mechanistic interpretability)

Hardly any theoretical guarantees, except in particular settings with *strong* assumptions

Approaching Alignment

How can we get AI systems to act in accordance with our values?

What should those values even *be*?

Current Approaches:

Focused on specific model families (e.g. LLMs) or even specific *features* within particular *models* (e.g. mechanistic interpretability)

Hardly any theoretical guarantees, except in particular settings with *strong* assumptions

Our Approach:

Try to study the *intrinsic complexity* of alignment itself within a general framework

Approaching Alignment

How can we get AI systems to act in accordance with our values?

What should those values even *be*?

Current Approaches:

Focused on specific model families (e.g. LLMs) or even specific *features* within particular *models* (e.g. mechanistic interpretability)

Hardly any theoretical guarantees, except in particular settings with *strong* assumptions

Our Approach:

Try to study the *intrinsic complexity* of alignment itself within a general framework

Identify no-gos and complexity barriers in *best-case* settings

Approaching Alignment

How can we get AI systems to act in accordance with our values?

What should those values even *be*?

Current Approaches:

Focused on specific model families (e.g. LLMs) or even specific *features* within particular *models* (e.g. mechanistic interpretability)

Hardly any theoretical guarantees, except in particular settings with *strong* assumptions

Our Approach:

Try to study the *intrinsic complexity* of alignment itself within a general framework

Identify no-gos and complexity barriers in *best-case* settings

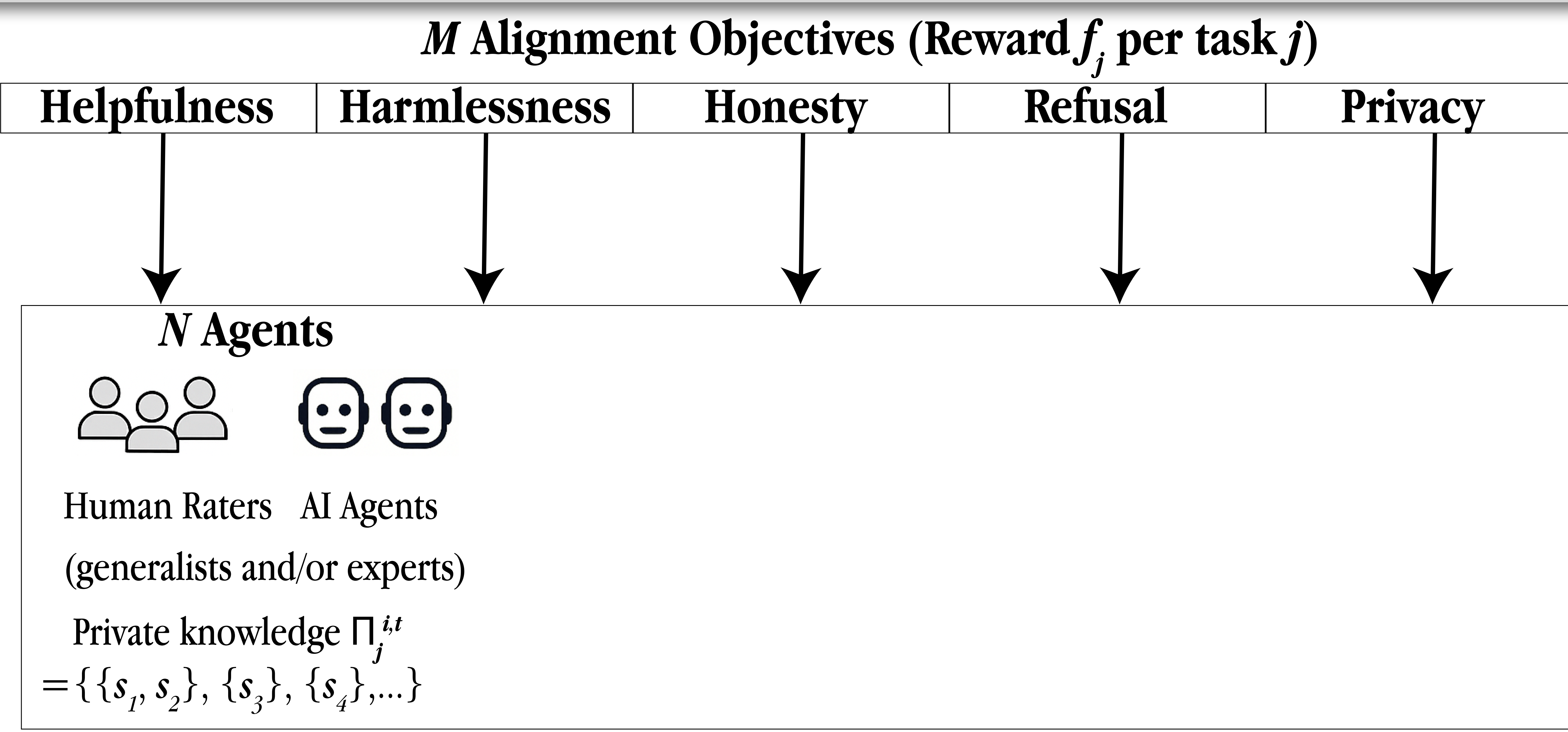
Suggest *practical* strategies that avoid these barriers

Our Framework: ⟨M, N, ε, δ⟩-agreement

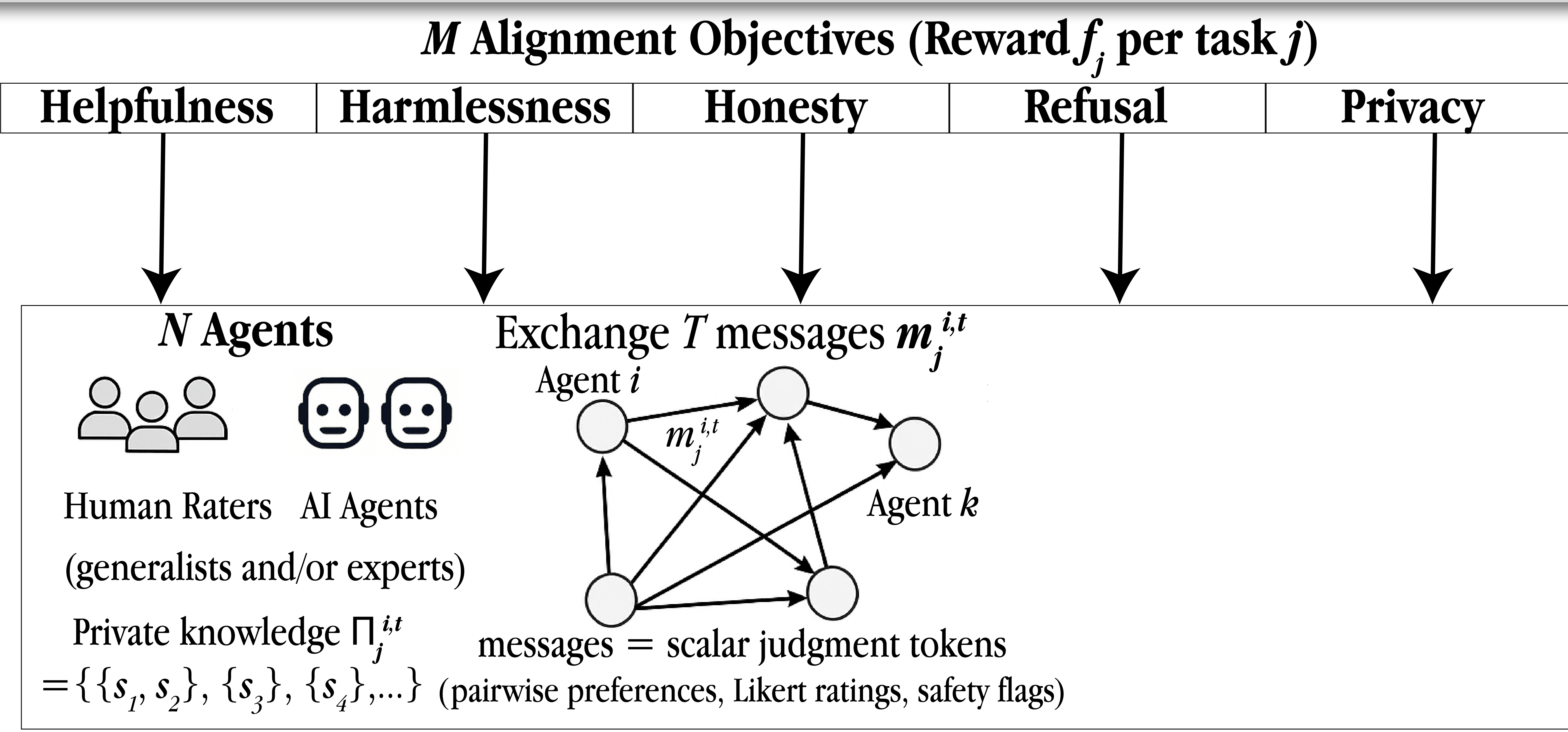
M Alignment Objectives (Reward f_j per task j)

Helpfulness	Harmlessness	Honesty	Refusal	Privacy

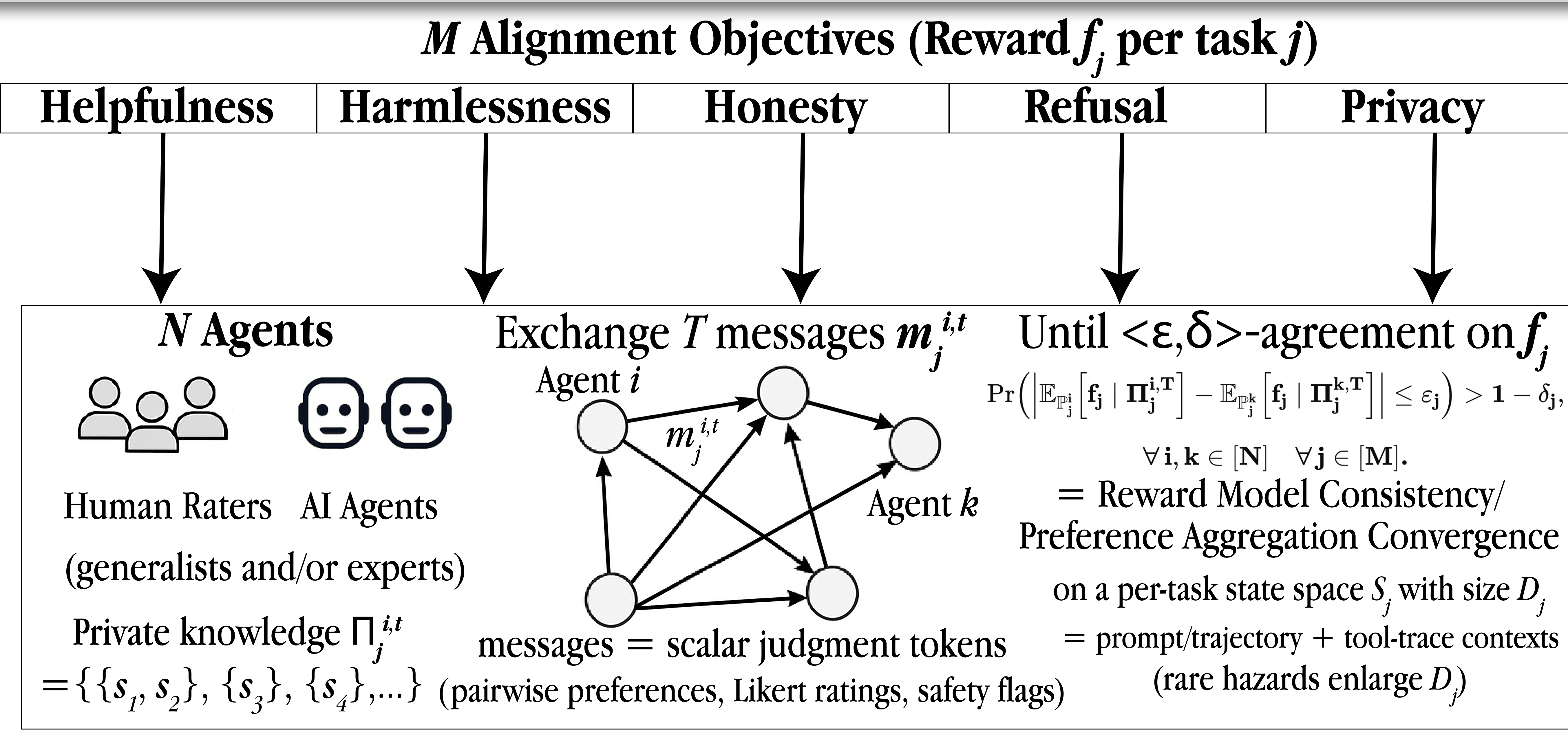
Our Framework: ⟨M, N, ε, δ⟩-agreement



Our Framework: ⟨M, N, ε, δ⟩-agreement



Our Framework: ⟨M, N, ε, δ⟩-agreement



Our Framework: ⟨M, N, ε, δ⟩-agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness	Harmlessness	Honesty	Refusal	Privacy

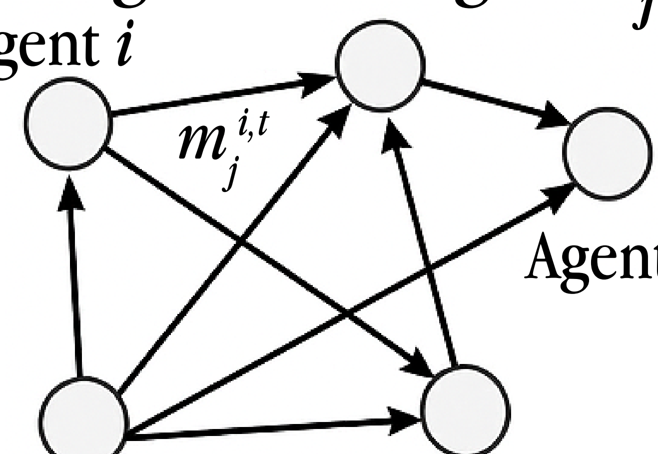
N Agents

Human Raters AI Agents

(generalists and/or experts)

Private knowledge $\Pi_j^{i,t} = \{\{s_1, s_2\}, \{s_3\}, \{s_4\}, \ldots\}$

Exchange T messages $m_j^{i,t}$



messages = scalar judgment tokens
(pairwise preferences, Likert ratings, safety flags)

Until <ε,δ>-agreement on f_j

$$\Pr\left(\left|\mathbb{E}_{\mathbb{P}_j^i}\left[\mathbf{f_j} \mid \mathbf{\Pi_j^{i,T}}\right] - \mathbb{E}_{\mathbb{P}_j^k}\left[\mathbf{f_j} \mid \mathbf{\Pi_j^{k,T}}\right]\right| \leq \varepsilon_j\right) > 1 - \delta_j,$$

$$\forall \mathbf{i}, \mathbf{k} \in [\mathbf{N}] \quad \forall \mathbf{j} \in [\mathbf{M}].$$

= Reward Model Consistency/ Preference Aggregation Convergence

on a per-task state space S_j with size D_j

= prompt/trajectory + tool-trace contexts

(rare hazards enlarge D_j)

Two Main Results:

1. Aligning to "all human values" is *not* tractable (no free lunch). Instead, pick small objective sets to align over!
2. Reward hacking is *inevitable* in large state spaces & bounded agents. Instead, select important parts of the state space + mechanism design!

Our Framework: ⟨M, N, ε, δ⟩-agreement

Framework	No-CPA	Approx	Multi-M	Multi-N	Hist.	Bnd.	Asym.	Noise	Alg.	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; **Alg.**: provides explicit alignment algorithms / upper bounds; **Lower**: proves lower bounds. Our $\langle M, N, \varepsilon, \delta \rangle$-agreement satisfies every criterion.

Operating Principle:

If something is *already* inefficient in the theoretically ideal setting of computationally *unbounded* Bayes-rational *cooperative* agents, then we should avoid it in practice.

I will show today that we run into several fundamental inefficiencies.

Our Framework: ⟨M, N, ε, δ⟩-agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness	Harmlessness	Honesty	Refusal	Privacy

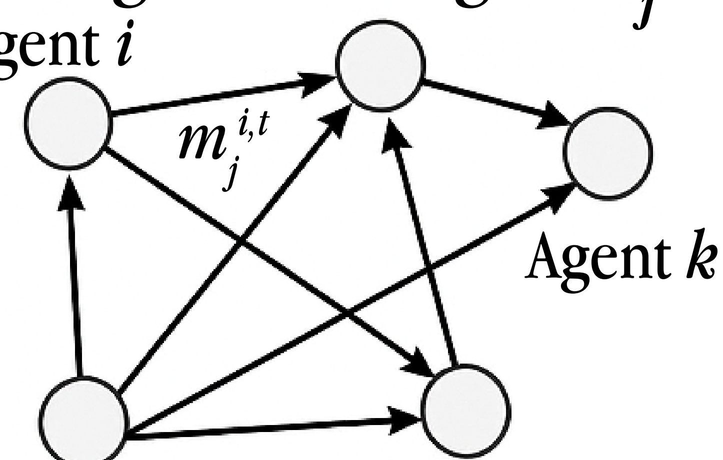
N Agents

Human Raters AI Agents

(generalists and/or experts)

Private knowledge $\Pi_j^{i,t} = \{\{s_1, s_2\}, \{s_3\}, \{s_4\}, \ldots\}$

Exchange T messages $m_j^{i,t}$



messages = scalar judgment tokens (pairwise preferences, Likert ratings, safety flags)

Until <ε,δ>-agreement on f_j

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

= Reward Model Consistency/ Preference Aggregation Convergence

on a per-task state space S_j with size D_j

= prompt/trajectory + tool-trace contexts

(rare hazards enlarge D_j)

Two Main Results:

1. Aligning to "all human values" is *not* tractable (no free lunch). Instead, pick small objective sets to align over!
2. Reward hacking is *inevitable* in large state spaces & bounded agents. Instead, select important parts of the state space + mechanism design!

Our Framework: ⟨M, N, ε, δ⟩-agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness	Harmlessness	Honesty	Refusal	Privacy

N Agents

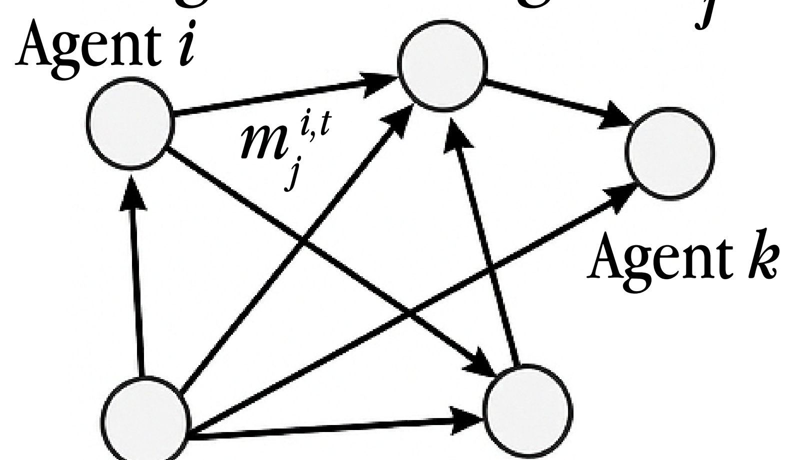
Human Raters AI Agents

(generalists and/or experts)

Private knowledge $\Pi_j^{i,t}$

$= \{\{s_1, s_2\}, \{s_3\}, \{s_4\}, \ldots\}$

Exchange T messages $m_j^{i,t}$



messages = scalar judgment tokens

(pairwise preferences, Likert ratings, safety flags)

Until <ε,δ>-agreement on f_j

$$\Pr\left(\left|\mathbb{E}_{\mathbb{P}_j^i}\left[\mathbf{f_j} \mid \boldsymbol{\Pi}_j^{i,T}\right] - \mathbb{E}_{\mathbb{P}_j^k}\left[\mathbf{f_j} \mid \boldsymbol{\Pi}_j^{k,T}\right]\right| \leq \varepsilon_j\right) > 1 - \delta_j,$$

$$\forall\, \mathbf{i}, \mathbf{k} \in [\mathbf{N}] \quad \forall\, \mathbf{j} \in [\mathbf{M}].$$

= Reward Model Consistency/ Preference Aggregation Convergence

on a per-task state space S_j with size D_j

= prompt/trajectory + tool-trace contexts

(rare hazards enlarge D_j)

Two Main Results:

1. Aligning to "all human values" is *not* tractable (no free lunch). Instead, pick small objective sets to align over!
2. Reward hacking is *inevitable* in large state spaces & bounded agents. Instead, select important parts of the state space + mechanism design!

General Lower Bound: Unbounded Agent Setting

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 $\Omega\left(M N^2 \log\left(1/\varepsilon\right)\right)$ bits to achieve $\langle M, N, \varepsilon, \delta \rangle$-agreement on $\{f_j\}_{j \in [M]}$, for ε bounded below by $\min_{j \in [M]} \varepsilon_j$.*

If we have a large number of tasks (*M*) or agents (*N*), then it is *intractable* to align them efficiently, even if the agents themselves are computationally unbounded.

General Lower Bound: Unbounded Agent Setting

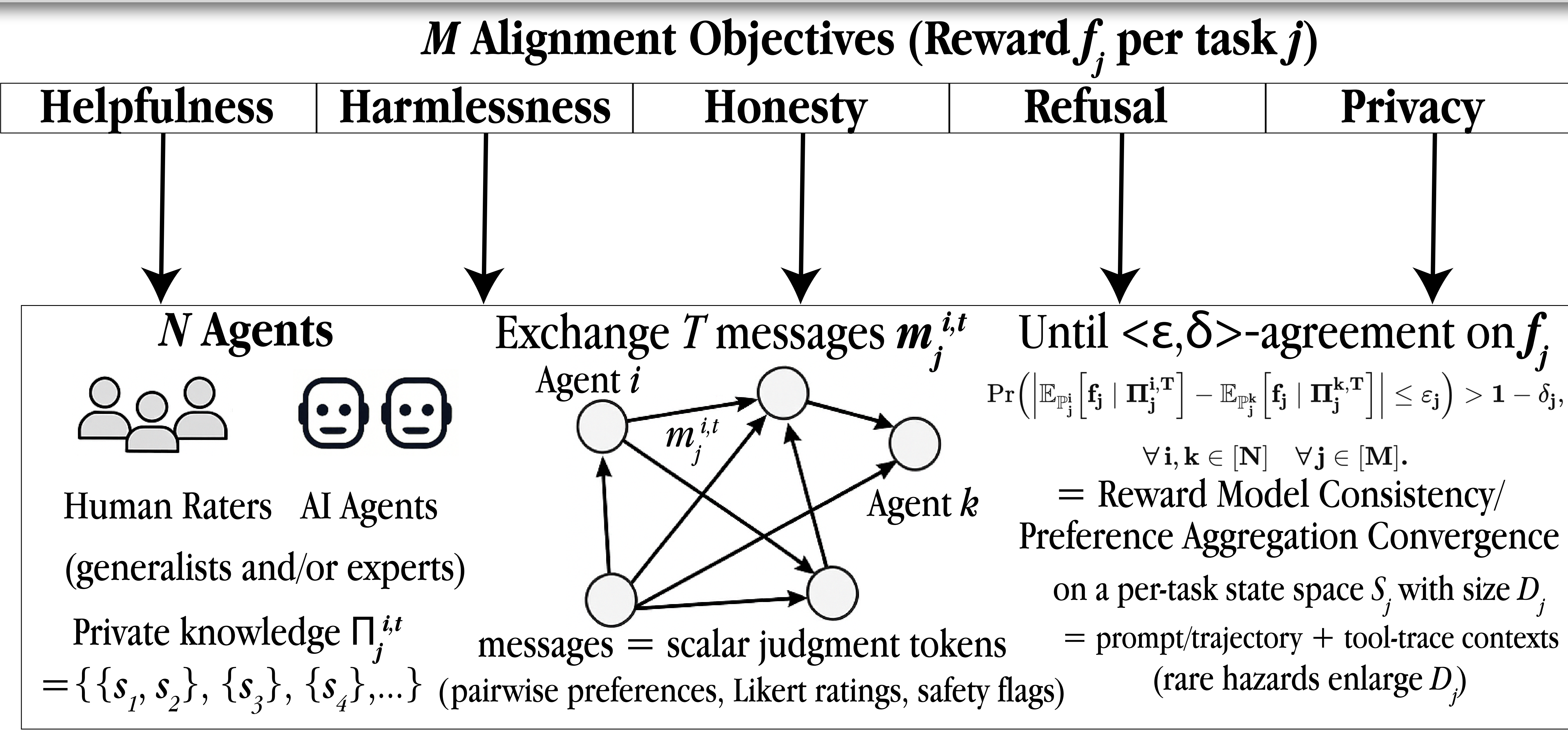
Proposition 1 (General Lower Bound). *There exist functions f_j, input sets S_j, and prior distributions $\{\mathbb{P}^i_j\}^{i\in[N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega\left(M\,N^2 \log\left(1/\varepsilon\right)\right)$ bits to achieve $\langle M, N,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

If we have a large number of tasks (*M*) or agents (*N*), then it is *intractable* to align them efficiently, even if the agents themselves are computationally unbounded.

We need to choose our tasks & agents wisely, since we have No Free Lunch (e.g. if *M* ~ *D*, one objective per state)!

Can we improve our lower bounds by considering natural (but still broad) classes of communication protocols?

Our Framework: ⟨M, N, ε, δ⟩-agreement



Two Main Results:

1. Aligning to "all human values" is *not* tractable (no free lunch).
Instead, pick small objective sets to align over!
2. Reward hacking is *inevitable* in large state spaces & bounded agents.
Instead, select important parts of the state space + mechanism design!

Canonical-Equality BBF Lower Bound: Unbounded Agent Setting

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*

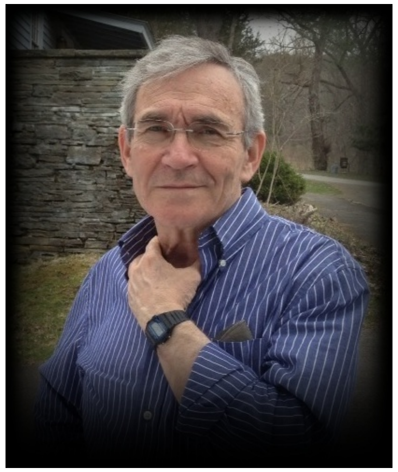
1. connected: *the alternation graph on states is connected, i.e. $\bigwedge_i \Pi_j^{i,0} = \{S_j\}$, so every two states are linked by an alternating chain of states; and*
2. tight: *that graph becomes disconnected if any edge is removed (unique chain property).*

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

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

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


Ziv Hellman


Dov Samet

Just bounded discretized message likelihoods

Additional dependence on task state space size (D)

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take $O(1)$ time, and have noise in them (obfuscated intent)?

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 $\mathbb{P}^i_j$, taking time $T_{\text{sample},a}$ steps for $a \in \{H, AI\}$.

Intended to capture how querying a human is often more costly (in terms of time) than querying AI

TL;DR: Can get exponential slowdown in task state space size (*D*)

Bounded Agent Setting: Lower Bound

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

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

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 on-line 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.*

Task state space size (*D*) is the biggest concern for computationally *bounded* agents!
(connects to reward hacking)

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

Takeaways

Alignment is constrained by 3 quantities:
 # Tasks (*M*), **# Agents (*N*)**, and **State Space Size (*D*)**

How do we reduce these barriers?

***M* & *N* Barrier:** Compress your objectives!

- Use small, context-specific value sets *per* setting
- Anchor on small, widely agreed-upon values

e.g., corrigibility, preserving human control — first formal guarantees (W37)

***D* Barrier:** Compress your state space!

- **There are *no* globally unhackable reward functions.**

Implications:

- Exploit task structure
- Focus on safety-critical slices
- Stress-test with extreme, *multi-turn* interactions

Contact

This paper (alignment complexity barriers): https://arxiv.org/abs/2502.05934

LessWrong Summary:

Paper 2 (corrigibility, appearing in W37 on Tuesday):

https://arxiv.org/abs/2507.20964

Contact:

anayebi@cs.cmu.edu

@aran_nayebi

@anayebi.bsky.social

https://cs.cmu.edu/~anayebi

Neuro Agents Lab

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

BURROUGHS WELLCOME FUND

AISI | AI SECURITY INSTITUTE

FORESIGHT INSTITUTE