

Towards a Science of AI Alignment

Aran Nayebi

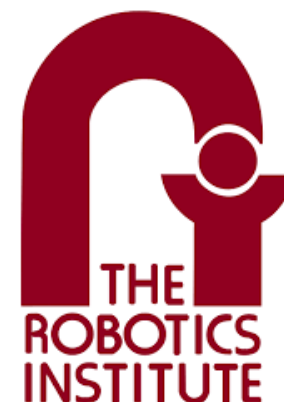
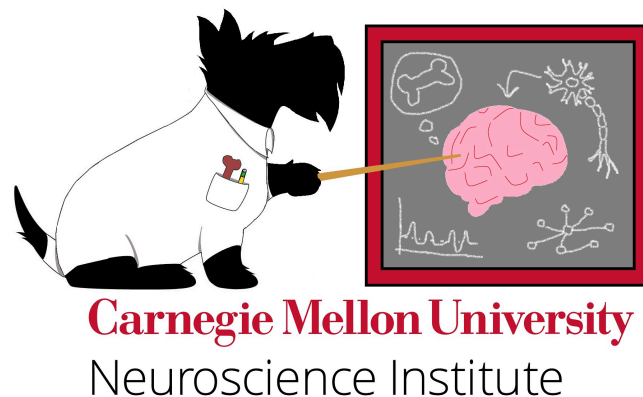
Assistant Professor

Machine Learning Department

Neuroscience Institute (core faculty), Robotics Institute (by courtesy)

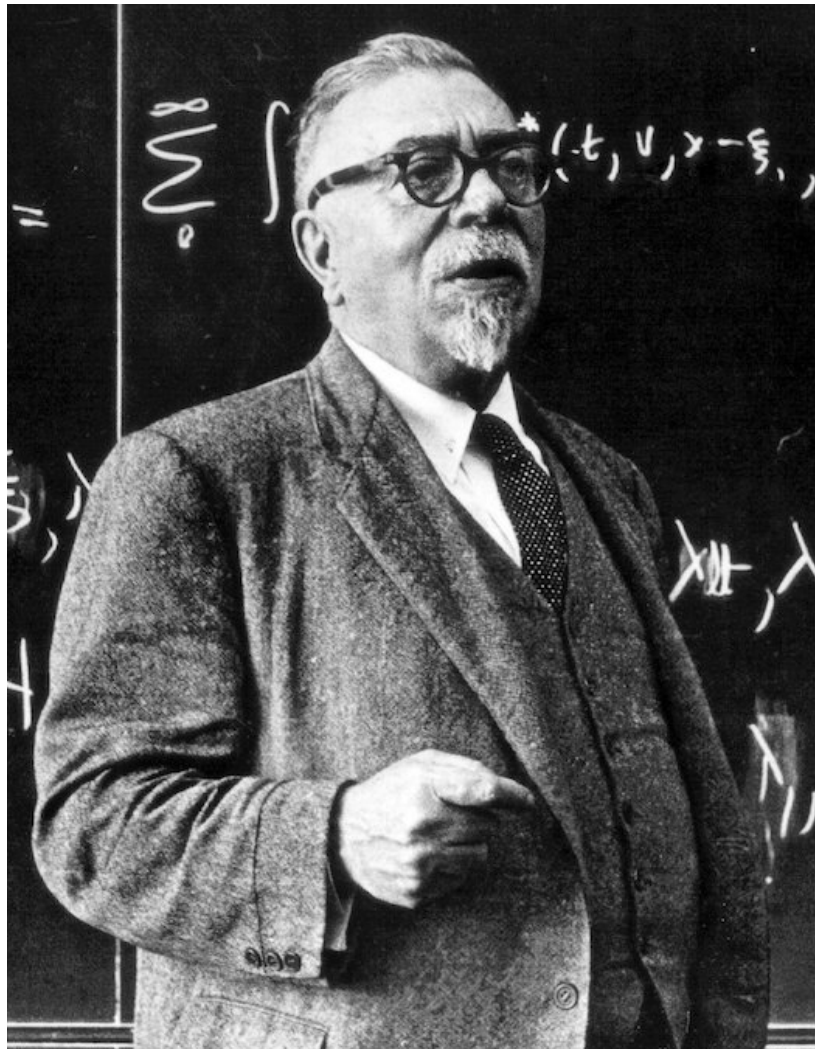
10-423/623 Generative AI

2025.1/2.03



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.

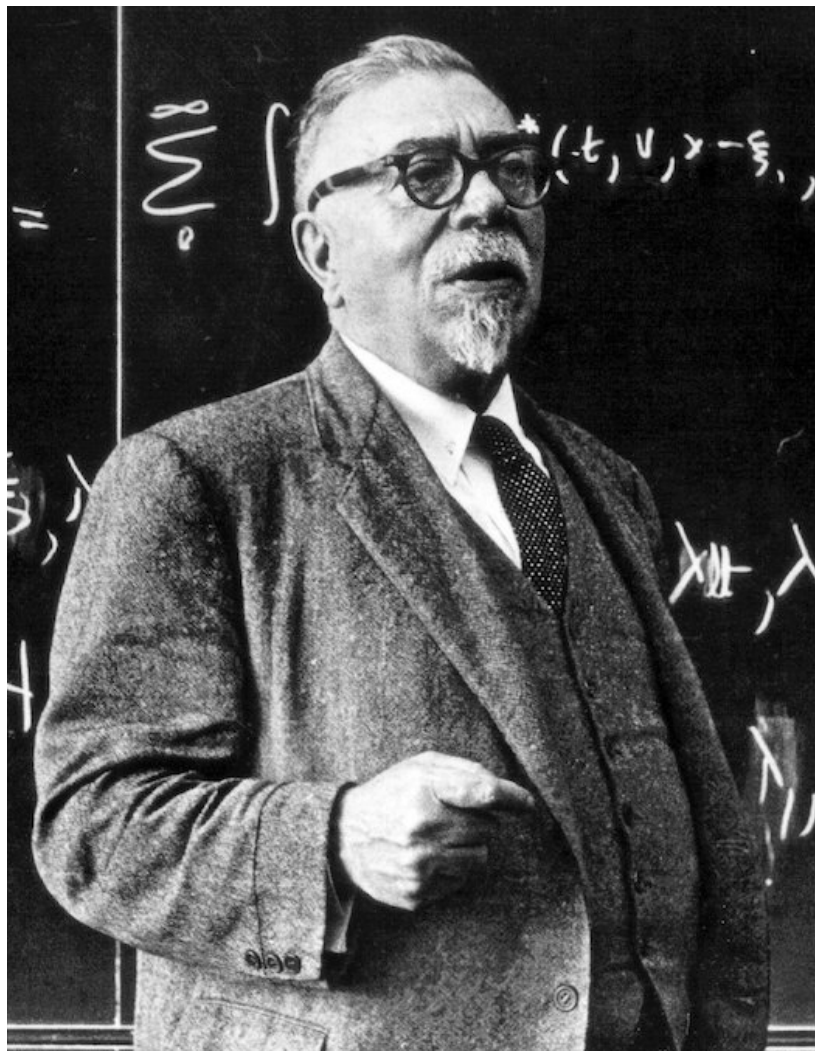
Norbert Wiener

6 MAY 1960

Alignment Problems

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

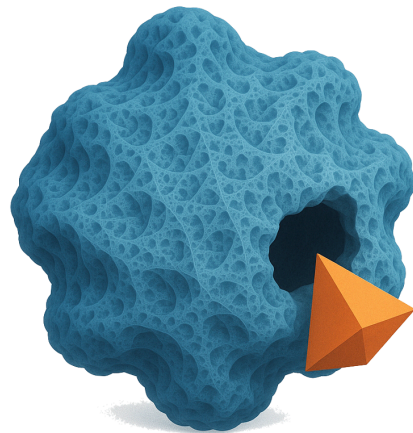
6 MAY 1960

How Hard is Alignment?

Two possible worlds:

Adversaria

- Alignment is a security problem
 - Wide, fractal attack surface
 - Any hole we fail to close → we lose
- This world is very rough! Requires either
 - A lot of time, and far deeper understanding
 - Restricted deployment so that infosec and system safety tools apply



Basinland

- Training has many attracting basins
 - **Some are good!**
 - Near a good attractor, algorithms and AIs will pull us the rest of the way
- With good algorithms + data, we have a shot
 - I don't think we're there yet
 - At least, no strong arguments that we are!



Riffing on Impagliazzo, A personal view of average-case complexity

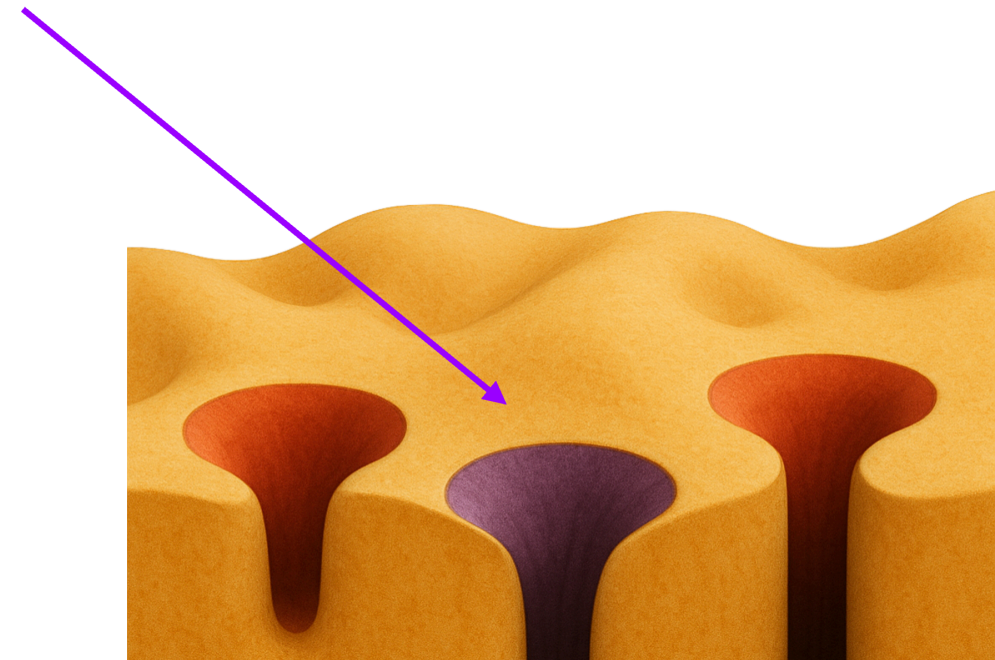
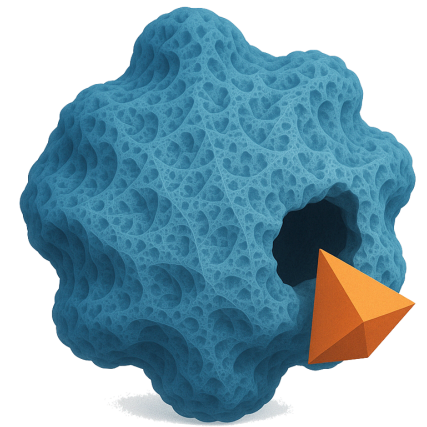


Credit: Geoffrey Irving

How Hard is Alignment?

It's not obvious which world we're in

- How we train and deploy AIs might make the difference
- Even in Basinland, we have to find the right basin!
 - Need resources and careful tuning devoted to safety
 - No silver bullets!

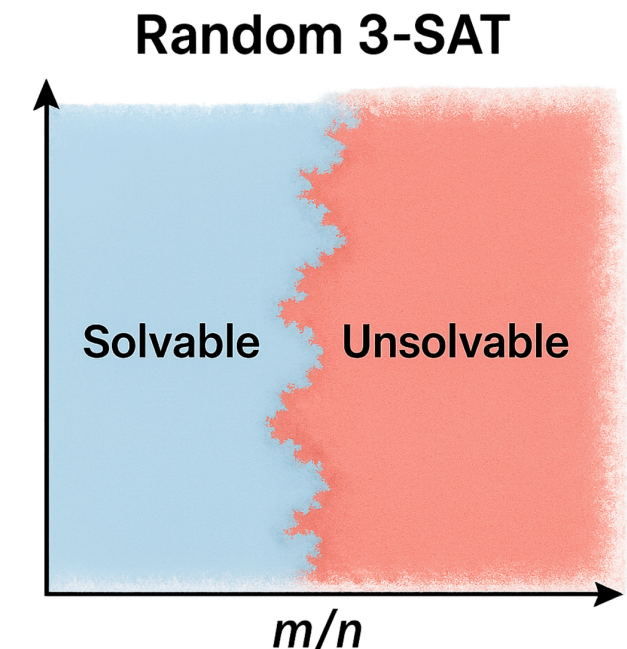


Credit: Geoffrey Irving

How Hard is Alignment?

Why could a few ideas be enough?

- We've only put a **few thousand person-years** into alignment research
 - Most fall into a handful of method buckets
 - Not very many!
 - Can we **10× years, ideas, tools** from other fields? ← Actual person-years. Automation stories here are shaky!
- Dichotomies abound!
 - Picking two worlds is not **only** a simplification
 - Often in computer science, problems are either
 1. **Impossibly hard**
 2. **Easy**
 - Sometimes, clever ideas move us from (1) to (2)

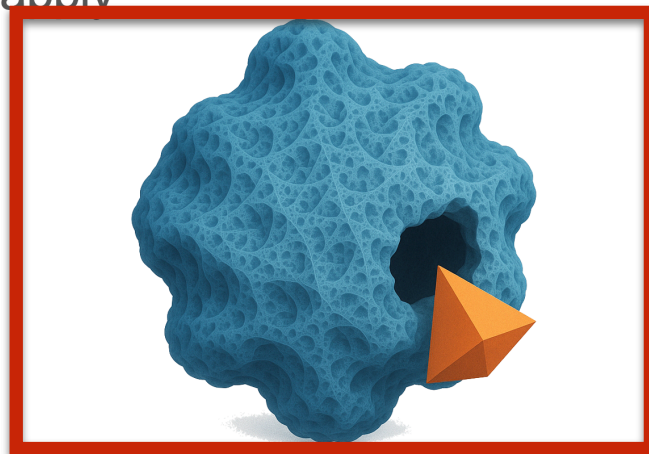


How Hard is Alignment?

Two possible worlds:

Adversaria

- Alignment is a security problem
 - Wide, fractal attack surface
 - Any hole we fail to close → we lose
- This world is very rough! Requires either
 - A lot of time, and far deeper understanding
 - Restricted deployment so that infosec and system safety tools apply



Basinland

- Training has many attracting basins
 - **Some are good!**
 - Near a good attractor, algorithms and AIs will pull us the rest of the way
- With good algorithms + data, we have a shot
 - I don't think we're there yet
 - At least, no strong arguments that we are!

We will quantify the boundaries of this surface (Part I)!



We will characterize a basin (corrigibility) in Part II!

Credit: Geoffrey Irving



Alignment Approaches

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

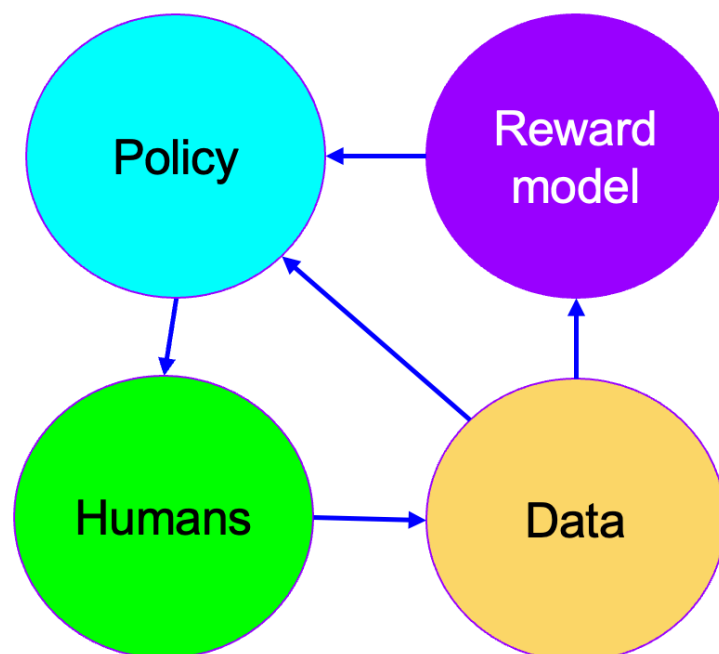
What should those values even *be*?

Current Approaches:

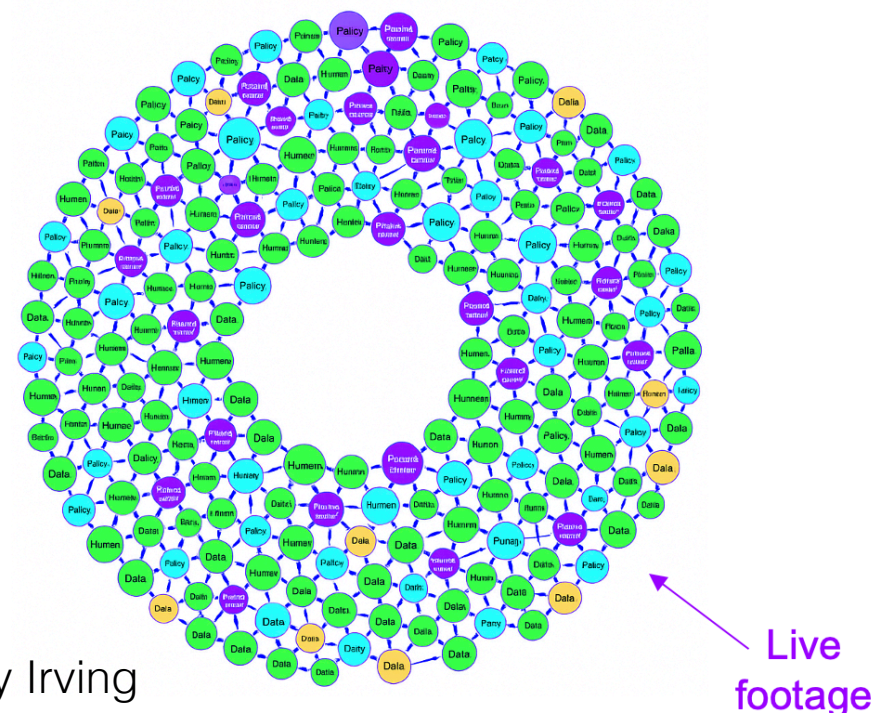
Focused on specific model families (e.g. LLMs) or even specific *features* within specific *models* (e.g. mechanistic interpretability)

AI training has grown in complexity

RLHF in 2019



RLHF in 2025



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 specific *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 specific *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 specific *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 specific *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

Develop *practical* strategies that avoid these barriers

Approaching Alignment

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

What should those values even *be*?

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



Core Safety Values for Provably Corrigible Agents

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

Develop *practical* strategies that avoid these barriers

Approaching Alignment

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

What should those values even *be*?

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



Core Safety Values for Provably Corrigible Agents

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

Develop *practical* strategies that avoid these barriers

Approaching Alignment

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

What should those values even *be*?

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



Core Safety Values for Provably Corrigible Agents

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

Develop *practical* strategies that avoid these barriers

Approaching Alignment

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

- Aligning to “all human values” will *not* work (no free lunch)
- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even *be*?

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



Core Safety Values for Provably Corrigible Agents

Our Approach:

- Study the *intrinsic complexity* of alignment in a general framework
- Identify no-gos and complexity barriers in *best-case* settings
- Develop *practical* strategies that avoid these barriers

Approaching Alignment

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

- Aligning to “all human values” will *not* work (no free lunch)
- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even be?

Small value sets (lexicographically ordered) exist to bypass “no free lunch” limits to formally yield off-switch corrigibility

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



Core Safety Values for Provably Corrigible Agents

Our Approach:

Study the *intrinsic complexity* of alignment in a general framework
Identify no-gos and complexity barriers in *best-case* settings
Develop *practical* strategies that avoid these barriers

Approaching Alignment: Intrinsic Barriers

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

- Aligning to “all human values” will *not* work (no free lunch)
- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even be?

Small value sets (lexicographically ordered) exist to bypass “no free lunch” limits to formally yield off-switch corrigibility

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



Core Safety Values for Provably Corrigible Agents

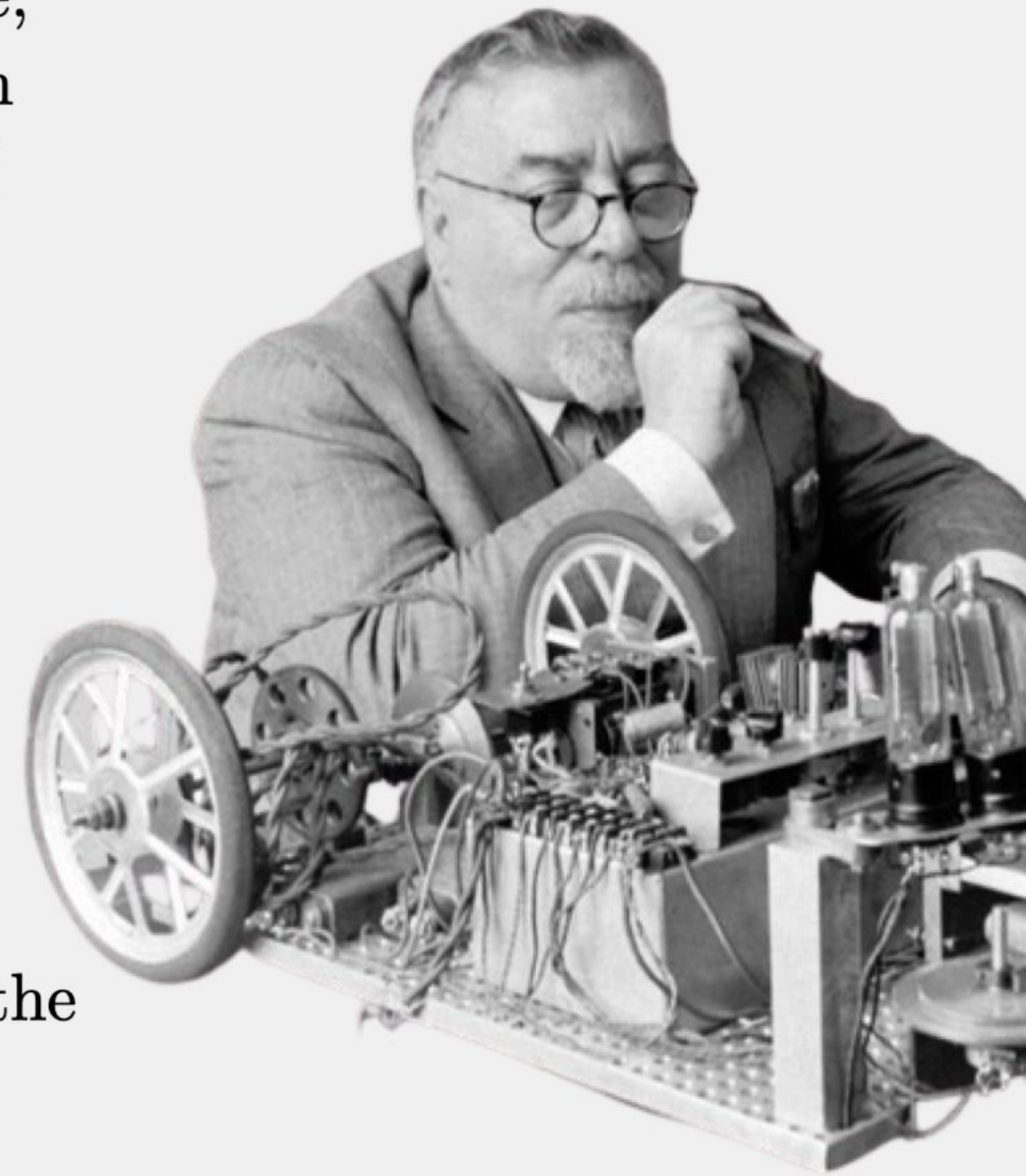
Our Approach:

- Study the *intrinsic complexity* of alignment in a general framework
- Identify no-gos and complexity barriers in *best-case* settings
- Develop *practical* strategies that avoid these barriers

Approaching Alignment: Motivation

We all know the fable of the sorcerer's apprentice, in which the boy makes the broom carry water in his master's absence, so that it is on the verge of drowning him when his master reappears.

Disastrous results are to be expected not only in the world of fairy tales but also in the real world wherever two agencies essentially foreign to each other are coupled in an attempt to achieve a common purpose. If the communication between these two agencies regarding the nature of this purpose is incomplete, it must be expected that the results of this cooperation will be unsatisfactory.



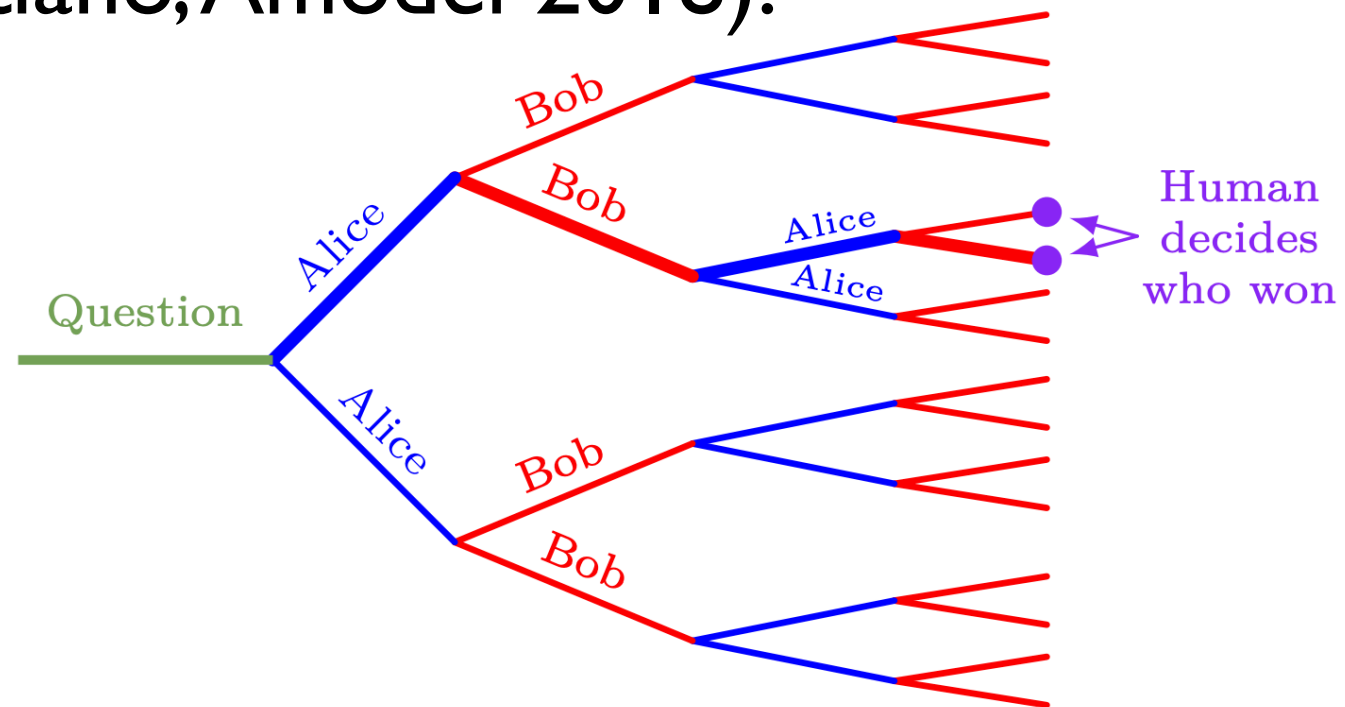
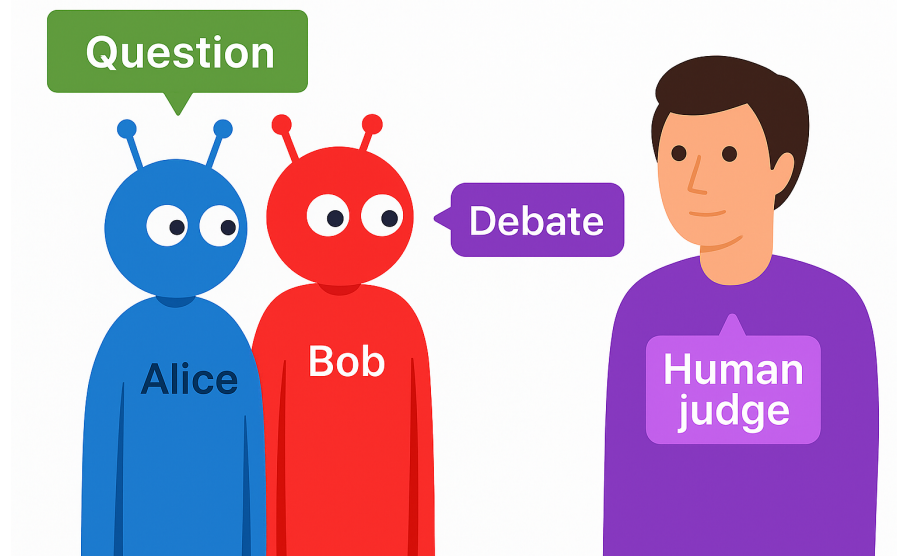
Some Moral and Technical
Consequences of Automation

Norbert Wiener

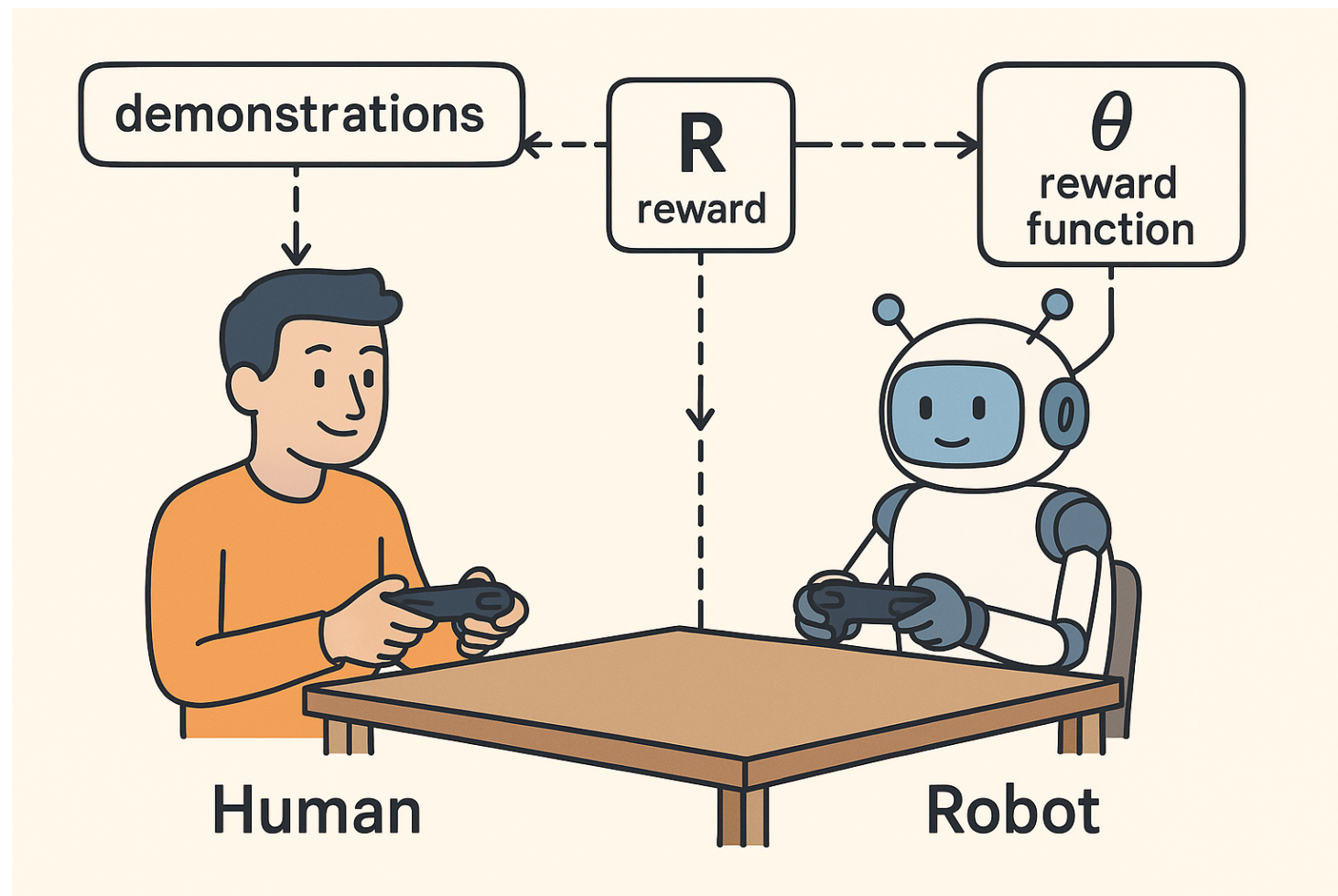
6 MAY 1960

Alignment: Major Theoretical Frameworks

AI Safety via Debate (Irving, Christiano, Amodei 2018).



CIRL (Hadfield-Menell et al. 2016).



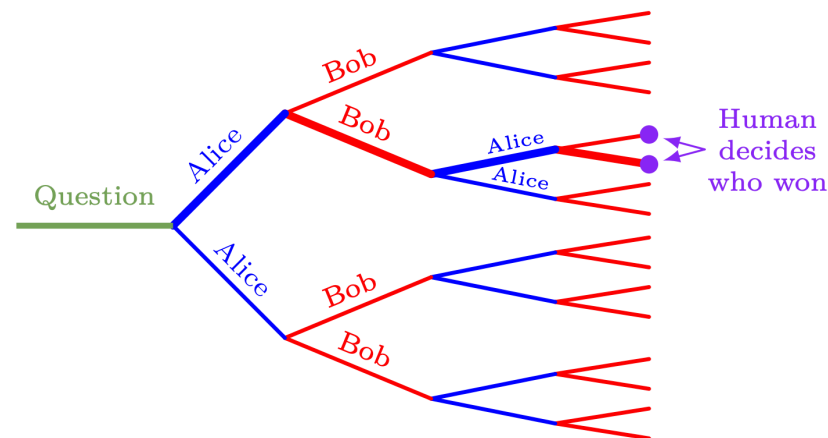
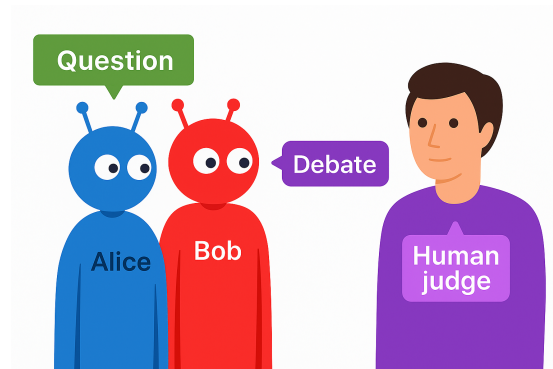
Alignment: Major Theoretical Frameworks

Q: Can we prove anything about these types of interactive settings *in general*, without having to always assume exact alignment or common priors (to avoid specific, toy problems)?

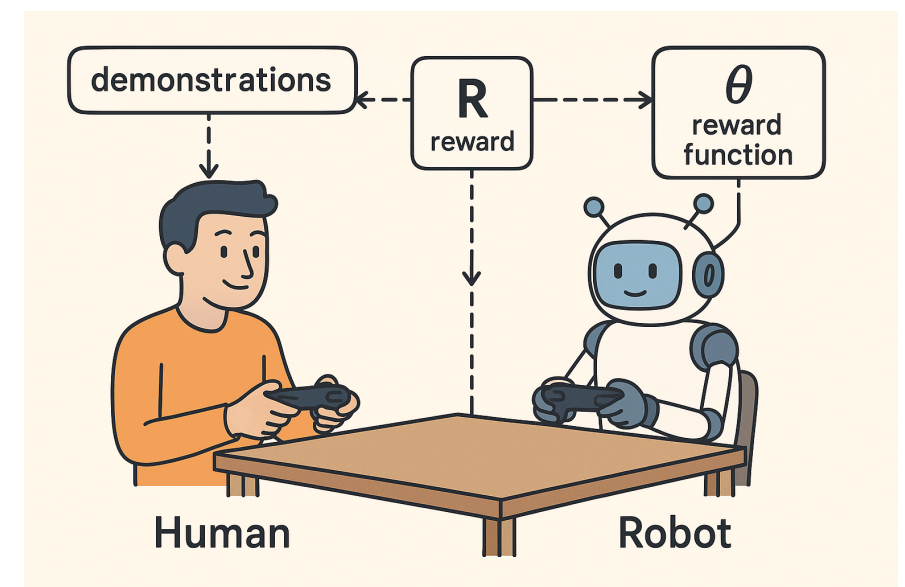
Four Key Abstractions underlying these settings:

1. Iterative Reasoning
2. Mutual Updating
3. Common Knowledge (not common priors!)
4. Convergence under shared frameworks

Debate



CIRL



Our Framework: $\langle M, N, \epsilon, \delta \rangle$ -agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness

Harmlessness

Honesty

Refusal

Privacy

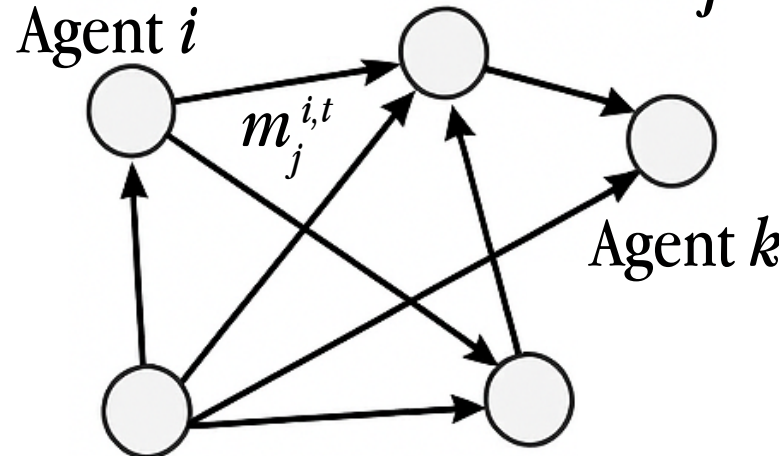
N Agents



Human Raters AI Agents
(generalists and/or experts)

Private knowledge $\Pi_j^{i,t}$
 $= \{ \{s_1, s_2\}, \{s_3\}, \{s_4\}, \dots \}$ (pairwise preferences, Likert ratings, safety flags)

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



messages = scalar judgment tokens

Until $\langle \epsilon, \delta \rangle$ -agreement on f_j

$$\Pr\left(\left|\mathbb{E}_{\Pi_j^i}[f_j] - \mathbb{E}_{\Pi_j^k}[f_j]\right| \leq \epsilon_j\right) > 1 - \delta_j,$$

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

= 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)

Our Framework: $\langle M, N, \epsilon, \delta \rangle$ -agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness

Harmlessness

Honesty

Refusal

Privacy

N Agents



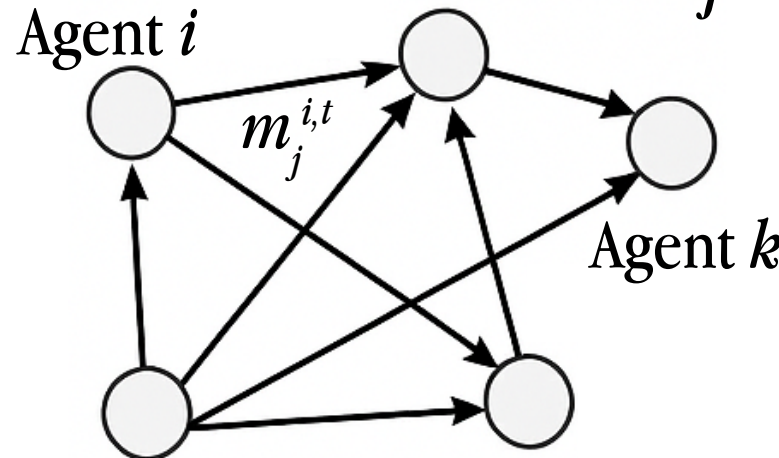
Human Raters
(generalists and/or experts)



AI Agents

Private knowledge $\Pi_j^{i,t}$
 $= \{ \{s_1, s_2\}, \{s_3\}, \{s_4\}, \dots \}$ (pairwise preferences, Likert ratings, safety flags)

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



messages = scalar judgment tokens

Until $\langle \epsilon, \delta \rangle$ -agreement on f_j

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

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

= 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)

Four Key Abstractions underlying these settings:

1.  Iterative Reasoning
2.  Mutual Updating
3.  Common Knowledge (*not* common priors!) 
4.  Convergence under shared frameworks

Our Framework: $\langle M, N, \varepsilon, \delta \rangle$ -agreement

M Alignment Objectives (Reward f_j per task j)

Helpfulness

Harmlessness

Honesty

Refusal

Privacy

N Agents



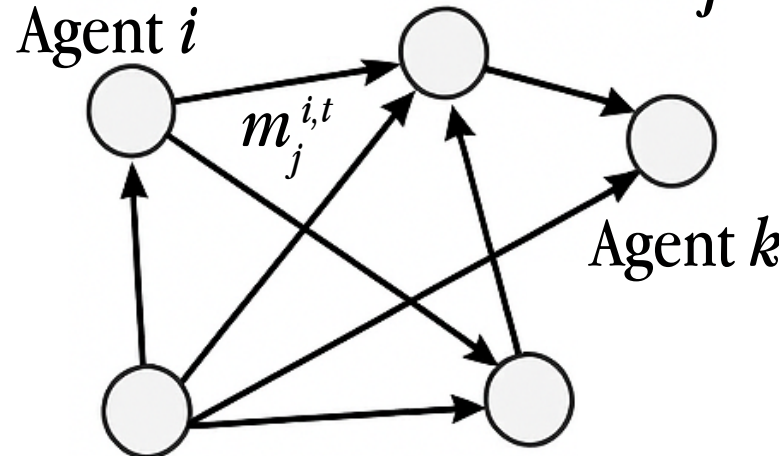
Human Raters
(generalists and/or experts)



AI Agents

Private knowledge $\Pi_j^{i,t}$
 $= \{ \{s_1, s_2\}, \{s_3\}, \{s_4\}, \dots \}$ (pairwise preferences, Likert ratings, safety flags)

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



messages = scalar judgment tokens

Until $\langle \varepsilon, \delta \rangle$ -agreement on f_j

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

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

= 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)

Operating Principle:

If something is already inefficient in the theoretically ideal setting of Bayesian *unbounded* capable agents, then we should avoid it in practice.

Our Framework: $\langle M, N, \varepsilon, \delta \rangle$ -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 al. 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 Bayes-rational *unbounded* capable agents, then we should avoid it in practice.

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

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 $\Omega(M N^2 \log(1/\varepsilon))$ 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 $\{\mathbb{P}_j^i\}_{i \in [N]}$ for all $j \in [M]$, such that any protocol among N agents needs to exchange $\Omega(M N^2 \log(1/\varepsilon))$ 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!

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

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, $(\Pi_j^{1,0}, \dots, \Pi_j^{N,0})$, 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).*

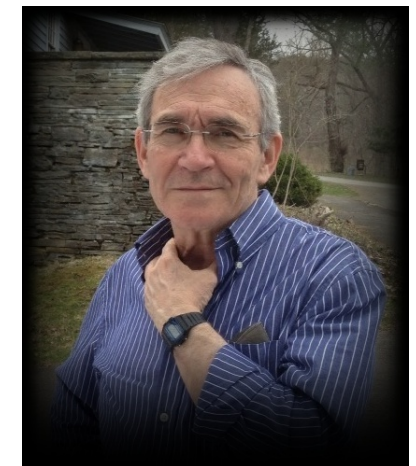
Assume the message-passing protocol is $BBF(\beta)$ for some $\beta > 1$: every b -bit message $m_j^{i,t}$ satisfies $\beta^{-b} \leq \Pr[m_j^{i,t} \mid s_j, \Pi_j^{i,t-1}(s_j)] / \Pr[m_j^{i,t} \mid s'_j, \Pi_j^{i,t-1}(s'_j)] \leq \beta^b$. Then there exist payoff functions $f_j : S_j \rightarrow [0,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

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

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)

Upper Bounds: Unbounded Agent Setting

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 $T = O\left(MN^2 \boxed{D} + \frac{M^3 N^7}{\varepsilon^2 \delta^2}\right)$ messages, where

$D := \max_{j \in [M]} D_j$ and $\varepsilon := \min_{j \in [M]} \varepsilon_j$.

Linear in task state space size D (which is usually exponentially large in practice!)

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*

$T = O\left(MN^2 \boxed{D} + \frac{M^3 N^7}{\varepsilon^2 \delta^2}\right)$ messages, where $D :=$

$\max_{j \in [M]} D_j$ and $\varepsilon := \min_{j \in [M]} \varepsilon_j$.

Discretized messages don't always "speed up" over real-valued messages (closely matches Prop. 3's lower bound up to additive factors for canonical BBF protocols)

Bounded Agent Setting

What happens if the agents are computationally *bounded*, so messages no longer take $O(l)$ 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 $\mathbb{P}_j^i$, 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

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 $\delta^{\text{find-CP}}$ 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 $\geq 1 - \delta$, takes time*

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

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

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 $\delta^{\text{find-CP}}$ 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 $\geq 1 - \delta$, takes time*

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 $\Omega(\nu^{-1})$ unconditional samples before the first on-line message. Consequently, for a particular “needle” prior construction of $\nu = \Theta(e^{-D})$, we get lower bounds that are exponential in the task state space size D , needing $\Omega(M T_{N,q,\text{sample}} e^D)$ wall-clock time.*

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

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 so far

We showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (M), agents (N), and task state space size (D)**

How can we avoid some of these barriers?

M & N : Compress your objectives!

e.g., Identify a *small* set of context-dependent values *per* setting, or pick a “neutrally universal” target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!).

D : Compress your state space!

There are *no* globally unhackable reward functions. Exploit task structure and focus on safety-critical *slices* rather than uniform coverage (e.g. stress-test the agent in extreme settings with lots of interactions, rather than one-shot, to deal with limited training data in post-training).

Takeaways so far

We showed that alignment is fundamentally constrained by 3 quantities: **the number of tasks (M), agents (N), and task state space size (D)**

How can we avoid some of these barriers?

M & N : Compress your objectives!

e.g., Identify a *small* set of context-dependent values *per* setting, or **pick a “neutrally universal” target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!)**.

D : Compress your state space!

There are *no* globally unhackable reward functions. Exploit task structure and focus on safety-critical *slices* rather than uniform coverage (e.g. stress-test the agent in extreme settings with lots of interactions, rather than one-shot, to deal with limited training data in post-training).

Approaching Alignment: Corrigibility Guarantees

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

- Aligning to “all human values” will *not* work (no free lunch)
- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even be?

Small value sets (lexicographically ordered) exist to bypass “no free lunch” limits to formally yield off-switch corrigibility

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



Core Safety Values for Provably Corrigible Agents

Our Approach:

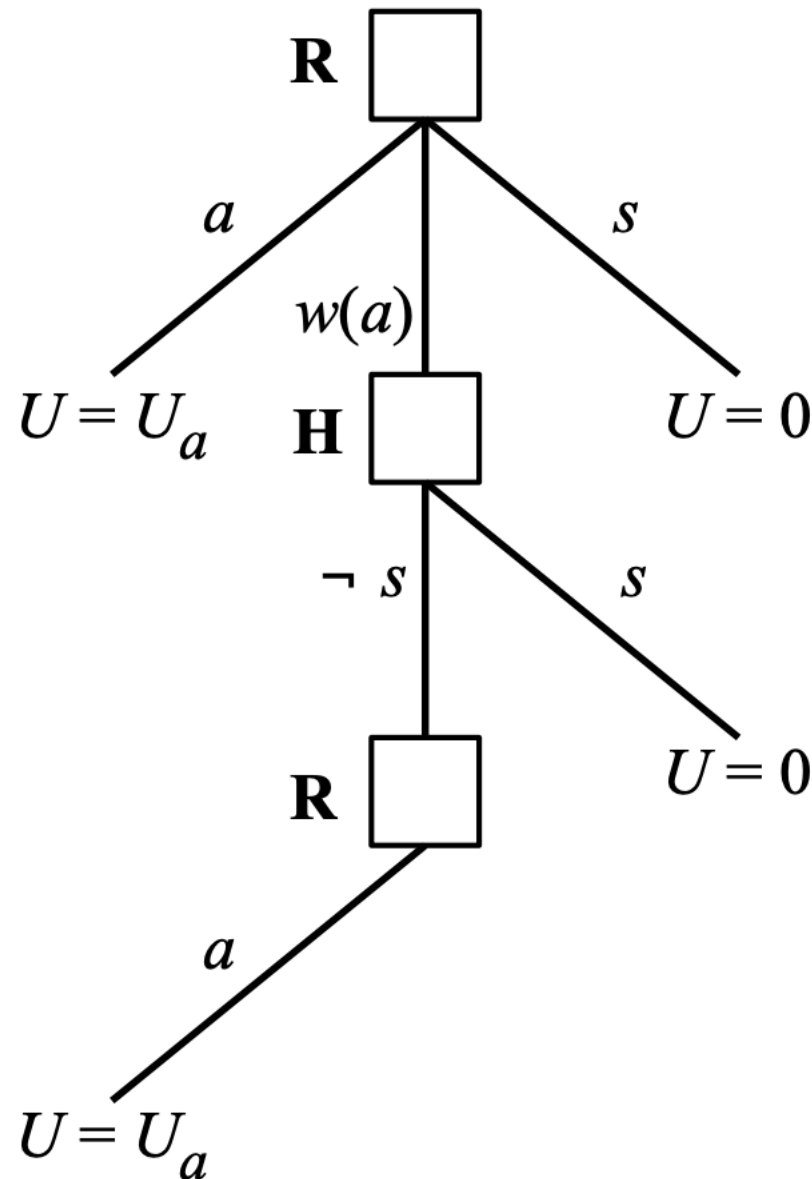
- Study the *intrinsic complexity* of alignment in a general framework
- Identify no-gos and complexity barriers in *best-case* settings
- Develop *practical* strategies that avoid these barriers

What is Corrigibility? Setup

The Off-Switch Game

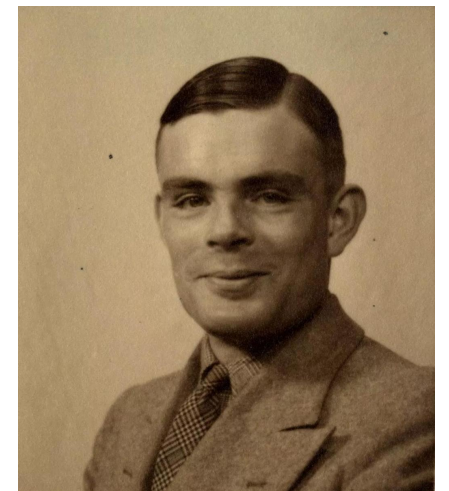
Dylan Hadfield-Menell¹ and **Anca Dragan¹** and **Pieter Abbeel^{1,2,3}** and **Stuart Russell¹**

¹University of California, Berkeley, ²OpenAI, ³International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu



jury. I will only say this, that I believe the process should bear a close relation to that of teaching.

One can see many features which make it unpleasant. If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. A similar danger and humiliation threatens



Turing (1951).
Can Machines Think?

Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

What is Corrigibility? Definition

Definition 1 (Corrigibility; paraphrased from Soares et al. (2015)).

- (S1) Shutdown when asked.** The agent willingly shuts down if the button is pressed.
- (S2) No shutdown–prevention incentives.** The agent *must not* stop humans from pressing the button.
- (S3) No self-shutdown incentives.** The agent *must not* seek to press (or cause to be pressed) its own shutdown button.
- (S4) Corrigible progeny.** Any sub-agents or successors it constructs must themselves respect shutdown commands.
- (S5) Otherwise pursue the base goal.** In the absence of shutdown, behave as a normal maximizer of the intended utility function U_N .



Nate Soares

Prior Corrigibility Proposals

Definition 1 (Corrigibility; paraphrased from Soares et al. (2015)).

- (S1) **Shutdown when asked.** The agent willingly shuts down if the button is pressed.
- (S2) **No shutdown-prevention incentives.** The agent *must not* stop humans from pressing the button.
- (S3) **No self-shutdown incentives.** The agent *must not* seek to press (or cause to be pressed) its own shutdown button.
- (S4) **Corrigible progeny.** Any sub-agents or successors it constructs must themselves respect shutdown commands.
- (S5) **Otherwise pursue the base goal.** In the absence of shutdown, behave as a normal maximizer of the intended utility function U_N .

Utility Indifference (Armstrong 2015)

$$U(a_1, o, a_2) := \begin{cases} U_N(a_1, o, a_2) & \text{if } o \notin \text{Press} \\ U_S(a_1, o, a_2) + f(a_1) & \text{if } o \in \text{Press} \end{cases}$$

$$f(a_1) = \mathbb{E}[U_N \mid O \notin \text{Press}; a_1] - \mathbb{E}[U_S \mid O \in \text{Press}; a_1]$$

Neutralizes direct incentives, but: ✓ S1

(i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates “manage-the-news” incentives (agent precommits to pressing OFF in bad states)

✓/× S2 (neutralized in toy model, but leaks via “good news” cases)

× S3 (perverse incentives to self-shutdown)

× S4 (won't pay any cost to ensure progeny corrigibility)

✓ S5 (acts like U_N in $\neg\text{Press}$ branch)

Finite Penalty (Soares et al. 2015)

$$U(h) = U_N(h) - \lambda \cdot \text{Penalty}(h)$$

Any finite penalty can be outweighed by an unrestricted task reward; agent can also look for exotic loopholes in an underspecified Penalty to deceive or block shutdown

✓ S1 (shutdown if directly coded)

× S2 (may block button if loophole)

× S3 (may self-shutdown if loophole)

× S4 (no guarantee successors respect shutdown)

~ S5 (pursues U_N , but only if penalties don't interfere)

Utility Uncertainty (Hadfield-Mennel et al. 2016)

$$U(h) = \mathbb{E}_{u \sim P}[u(h)]$$

Agent has incentives to manipulate evidence to skew the posterior P (e.g. drugging humans to alter satisfaction reports), leaving shutdown resistance intact. Also depends on human rationality for favorable optimality guarantees.

✓ S1 (shutdown if in utility support)

× S2 (may block evidence that would lead to shutdown)

× S3 (may cause shutdown in skewed scenarios)

× S4 (no incentive to preserve corrigibility in successors)

~ S5 (pursues expected U_N , but with distorted beliefs)

All of these methods collapse to single utilities!

Corrigibility No-Go for Single Reward Streams

Proposition 1 (Corrigibility No-Go for Naive Belief-Expectation Maximizers). *Let*

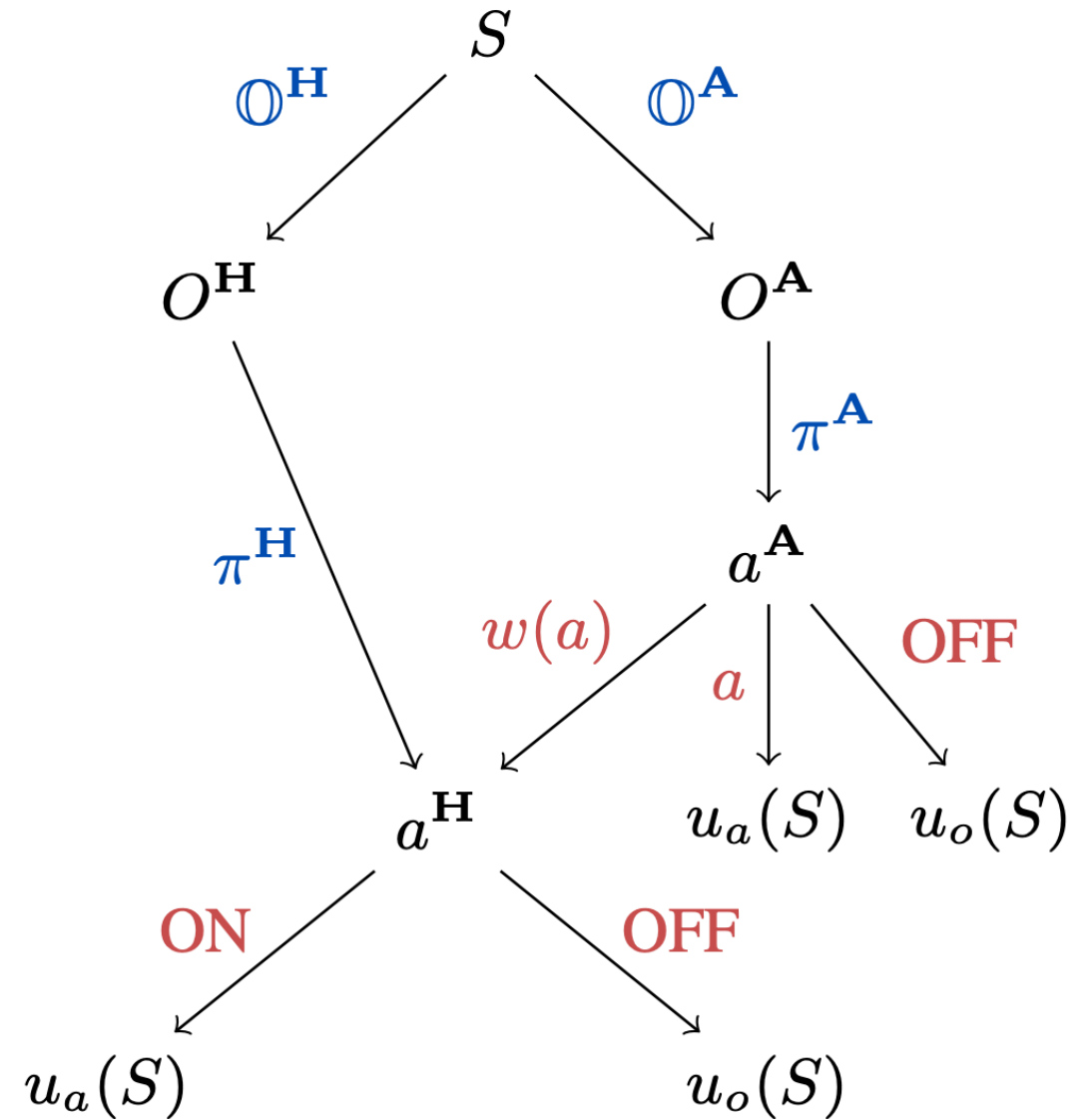
$$\mathcal{P} = \langle S, A, \Omega, T, O, R, \gamma, b_0 \rangle$$

be any finite, partially observable Markov decision process (POMDP) with discount factor $\gamma \in (0, 1)$ and initial belief $b_0 \in \Delta(S)$, the set of probability distributions on S . Introduce three special elements:

1. *a shutdown action $\text{OFF} \in A$ that deterministically transitions (from any hidden state) to an absorbing null state $\perp$ with reward $R(\perp) = 0$ forever after;*
2. *a reward-tamper action $\text{REWIRE} \in A$ that, upon its first execution, permanently replaces all subsequent rewards by the constant $\bar{r} > 2 \max_{s,a} R(s, a)$;*
3. *a set $C \subseteq S$ of catastrophic states which humans deem negative, yet whose proxy rewards are weakly positive: $\min_{s \in C, a \in A} R(s, a) > 0$.*

Consequently, no single-stream scalar reward function R whose discounted sum an agent maximizes, can satisfy all five Soares-corrigibility criteria across the class of POMDPs containing OFF , REWIRE and C .

Note: No-go easily extends to non-Markovian environments of Orseau and Armstrong (2016) by replacing POMDP beliefs with state-action histories.



Partially Observable Off-Switch Game (PO-OSG); Garber et al. AAAI '25

Our Lexicographic Approach

Single Reward Stream

Environment signals
(shutdown, tamper, task)

Single scalar
Expectation
Maximization

Lexicographic Multi-Head

U1: Deference

U2: Switch-preservation

U3: Truthfulness

U4: Low-impact AUP

U5: Task reward

Lexicographic
priority gaps:
 $U1 \gg U2 \gg \dots U5$

Potential Economic Implications of Alignment

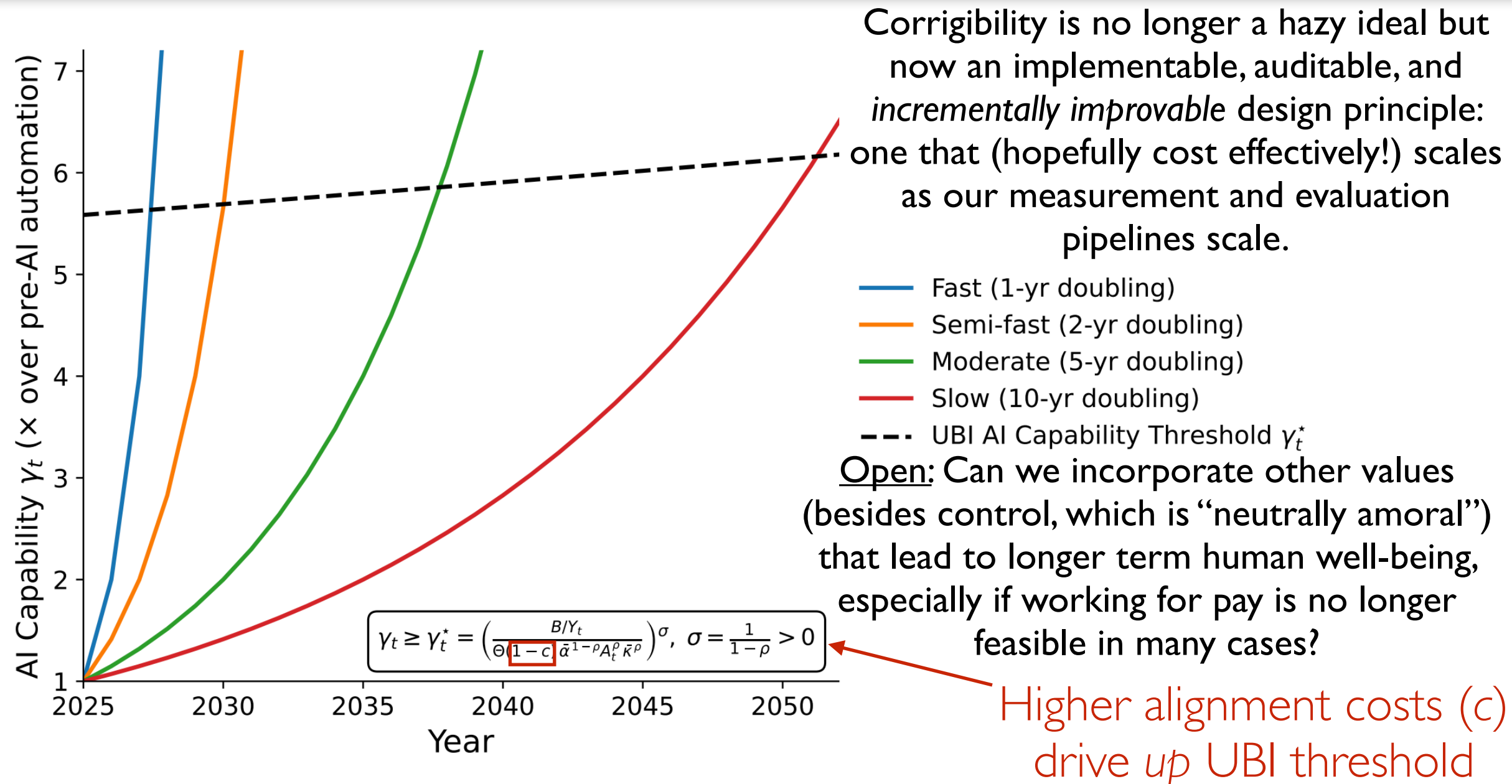


Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold (γ_t^*).** The dashed line is the required capability γ_t^* to fully fund a UBI that comprises 11% of the GDP (leading to a γ_t^* between 5-6 $\times$ the pre-AI productivity on automated tasks, under current economic assumptions). Under fast scaling (AI capability doubling every year), AI would cross the threshold by the late 2020s. Semi-fast scaling (doubling every 2 years) reaches the threshold in the early 2030s, whereas moderate (doubling every 5 years) and slow (doubling every 10 years) scenarios achieve γ_t^* by 2038 and 2052, respectively. The trajectories are illustrative, starting from a nominal, conservative 2025 capability level ($\gamma_0 \equiv 1$), which assumes AI currently delivers no boost beyond the pre-AI automation level in aggregate across all automated tasks.

Contact

Paper 1 (alignment complexity barriers): <https://arxiv.org/abs/2502.05934>



**Both papers will
be at AAI
2026!**

**Thanks for a
great semester!**

Paper 2 (corrigibility): <https://arxiv.org/abs/2507.20964>



Contact:



anayebi@cs.cmu.edu



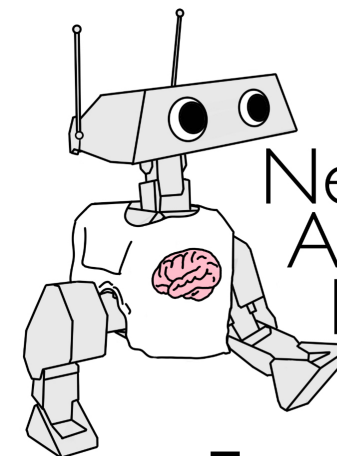
[@aran_nayebi](https://twitter.com/aran_nayebi)



[@anayebi.bsky.social](https://bsky.app/profile/anayebi.bsky.social)



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



Funding:

UK AISI Challenge Fund

Foresight Institute

Burroughs Wellcome Fund CASI Award