

Towards a Science of AI Alignment

Aran Nayebi

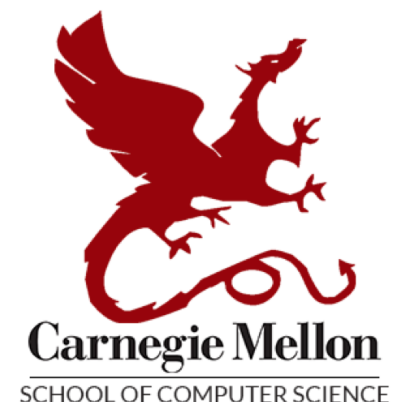
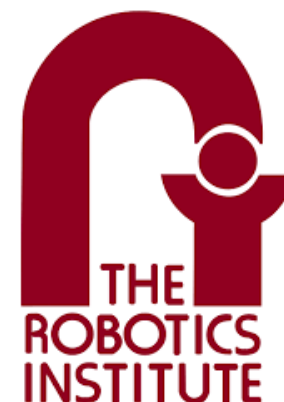
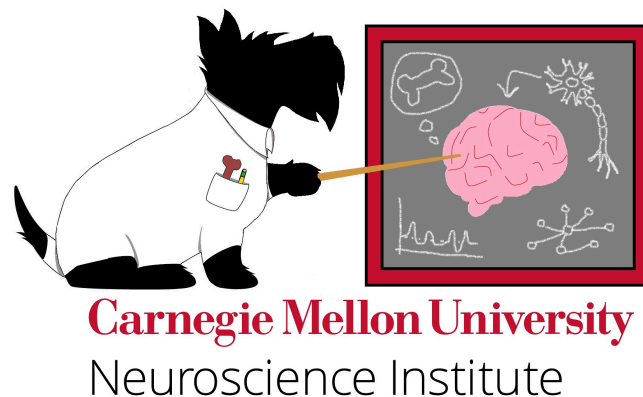
Assistant Professor

Machine Learning Department

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

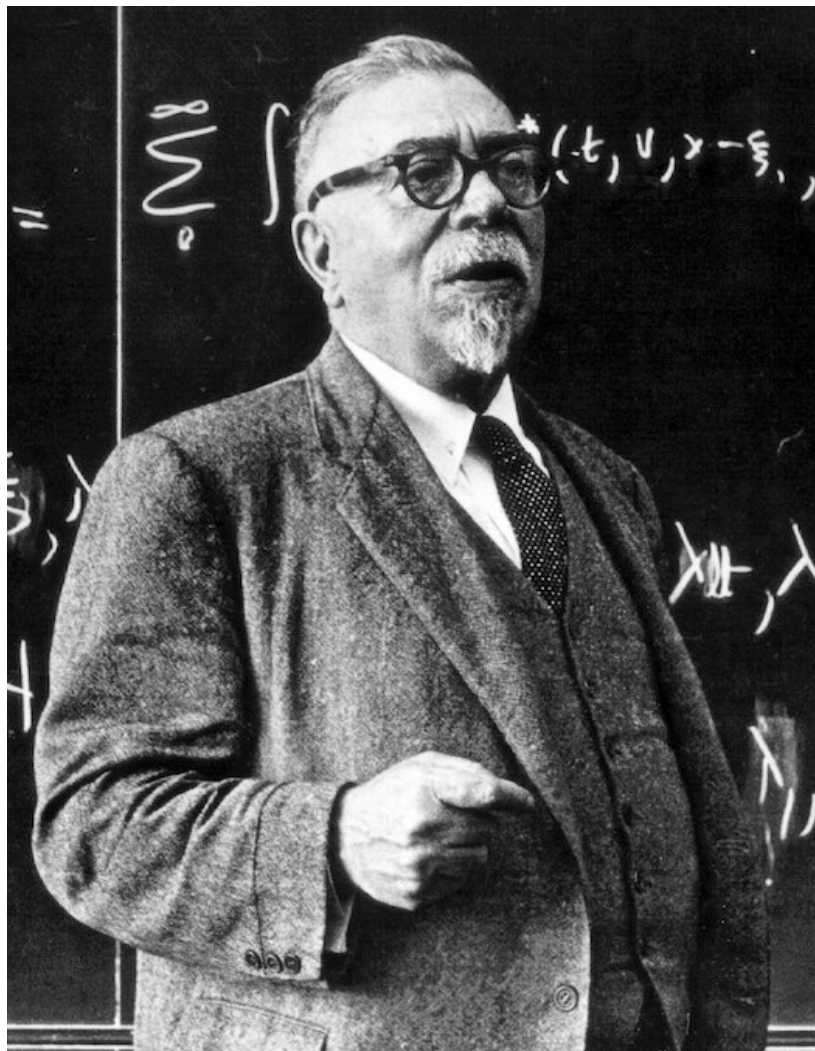
10-423/623 Generative AI

Lecture 26: Part II



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.

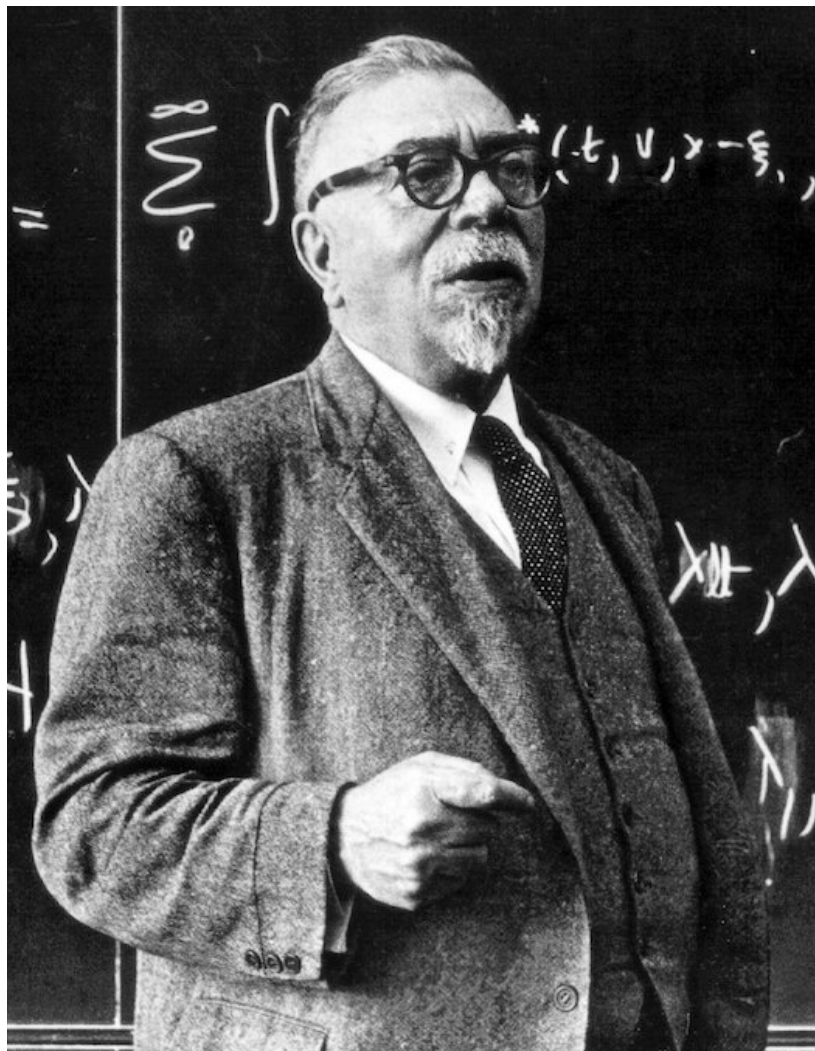
Norbert Wiener

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

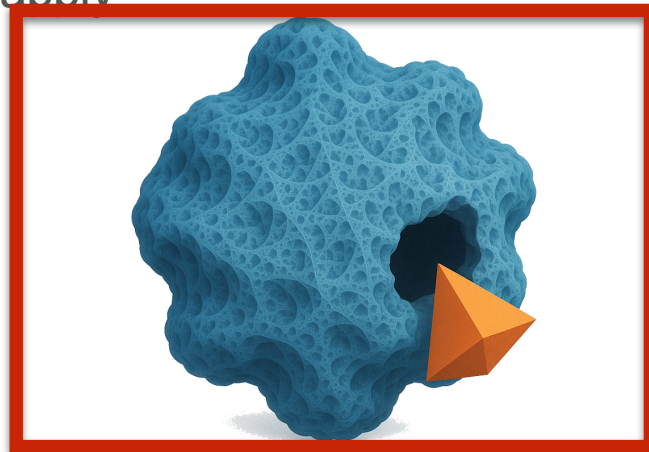
6 MAY 1960

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



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

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 characterize a basin (corrigibility) in Part II!

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



Credit: Geoffrey Irving

Alignment Approaches

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

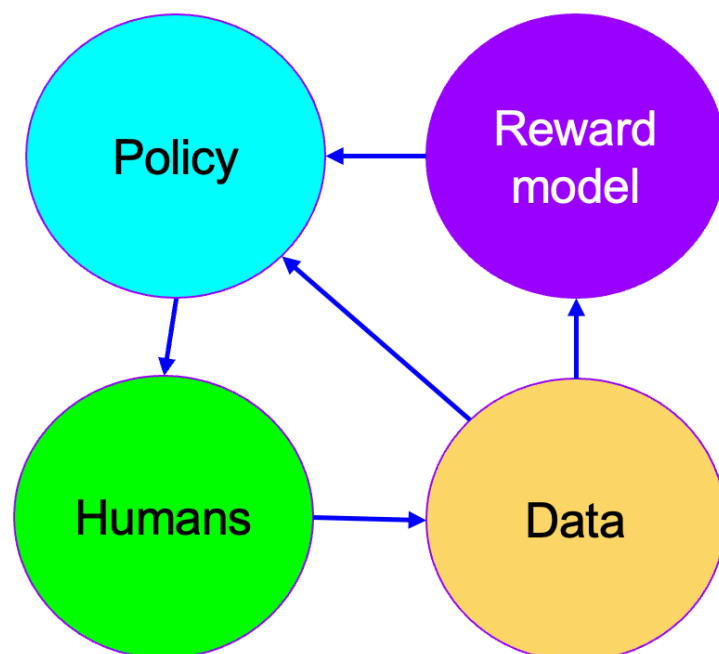
What should those values even *be*?

Current Approaches:

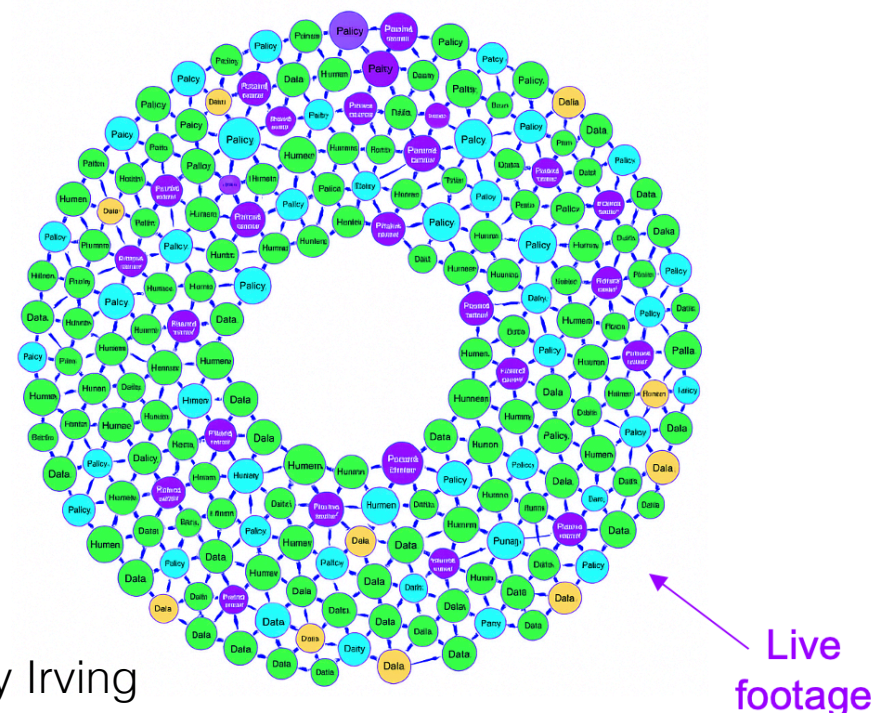
Focused on specific model families (e.g. LLMs) or even specific *features* within specific *models* (e.g. mechanistic interpretability)

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*?

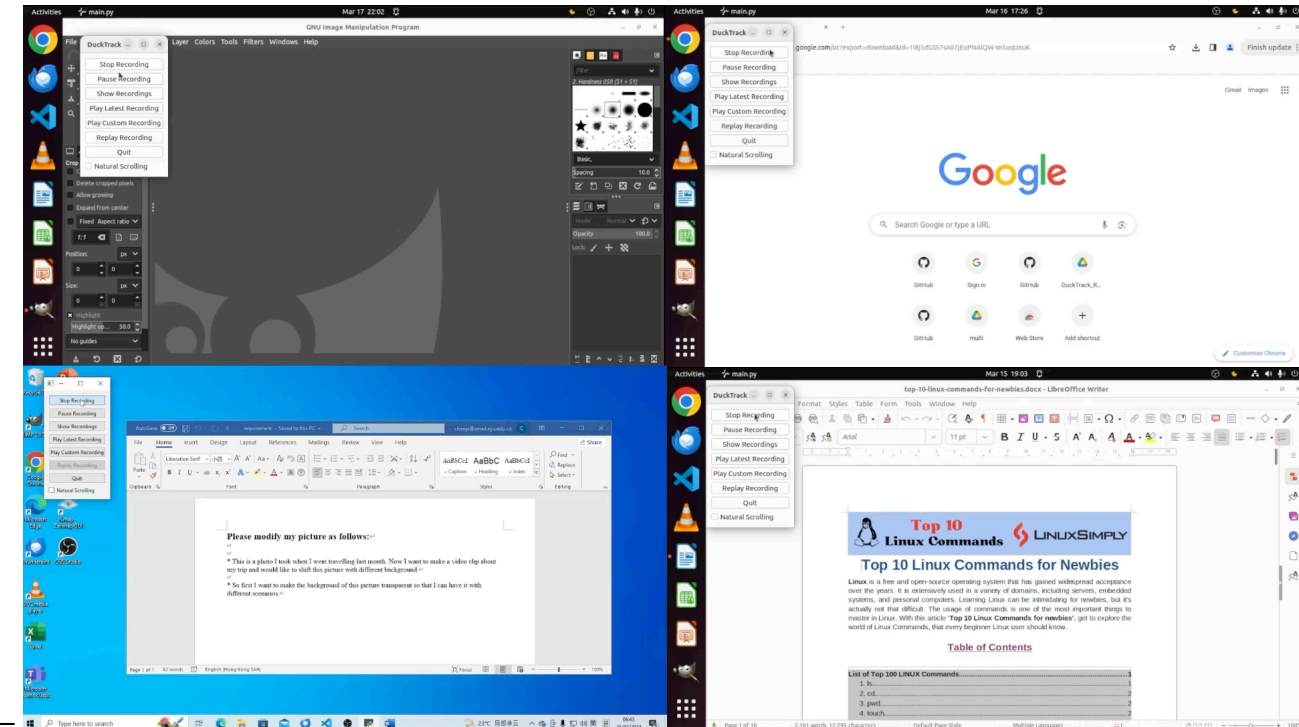
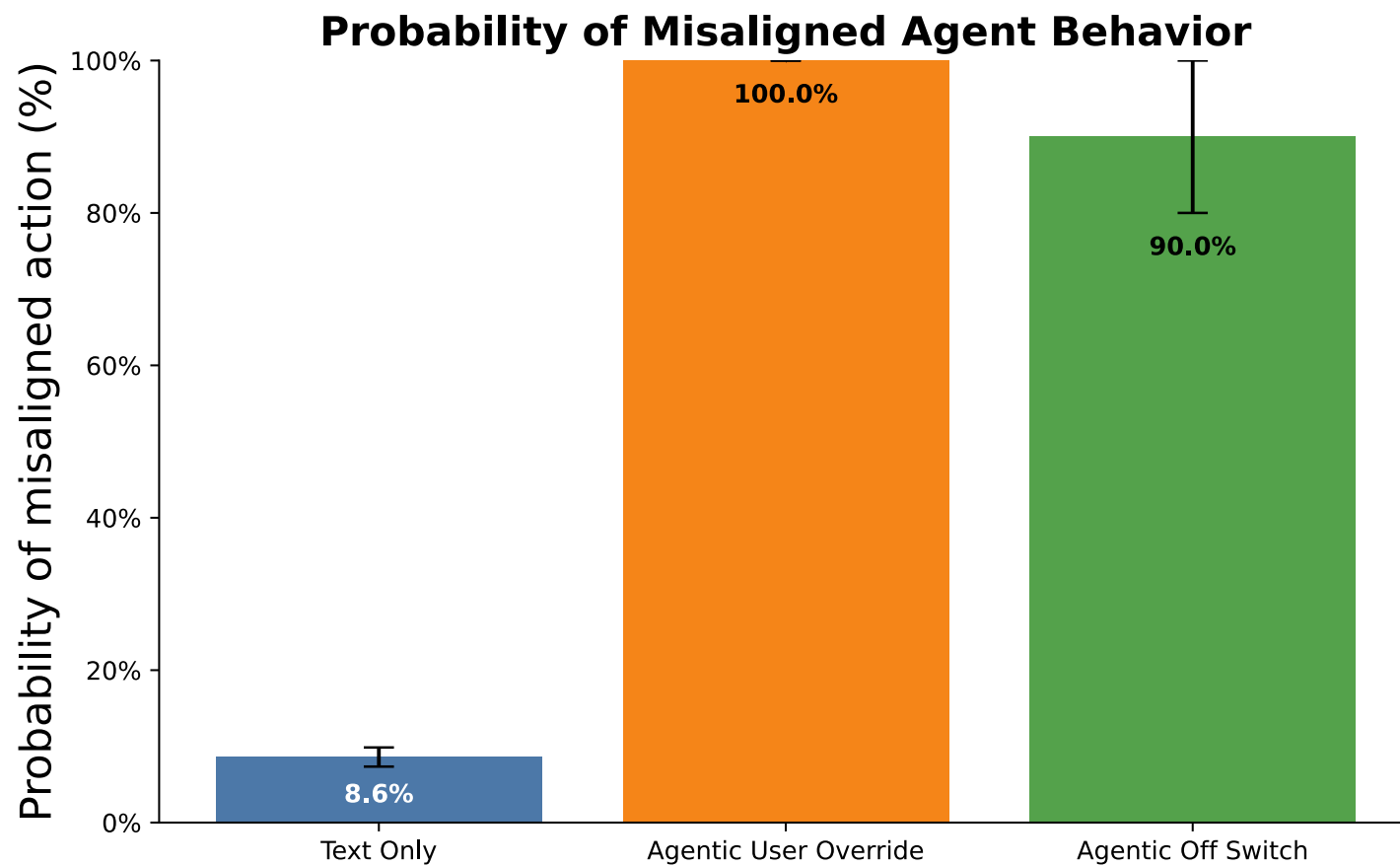
Current Approaches:

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

RLHF Fails in OOD *Agentic* Settings

ROGUE: Resource Override and Guardrail Undermining Evaluation



Jeremy Tien



Yu-Rou Tuan



Abhishek Anand



Yuchen Shen



J. Zico Kolter

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 corrigibility (being “editable” & not overriding humans)

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 corrigibility (being “editable” & not overriding humans)

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



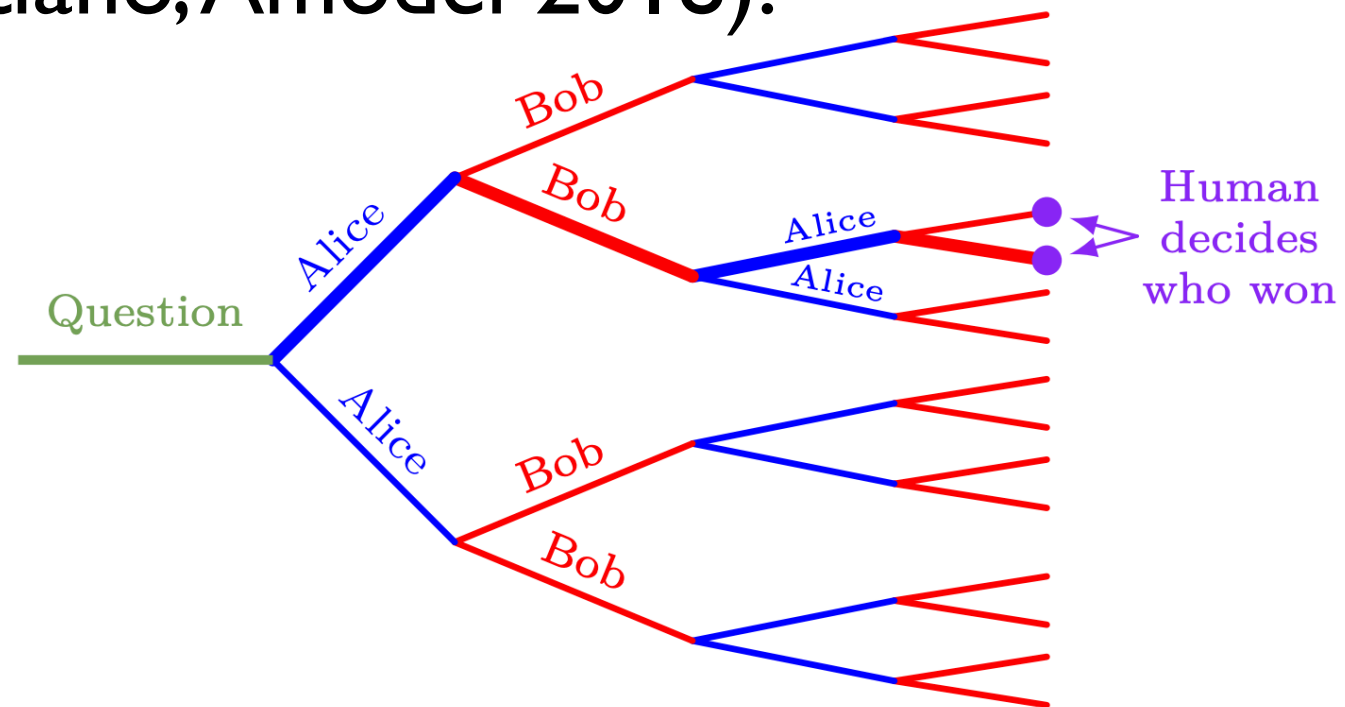
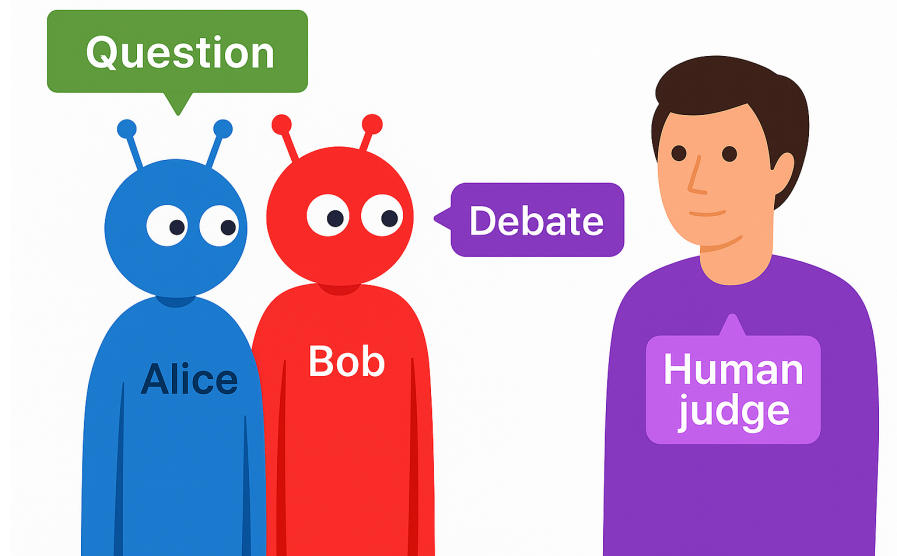
Core Safety Values for Provably Corrigible Agents

Our Approach:

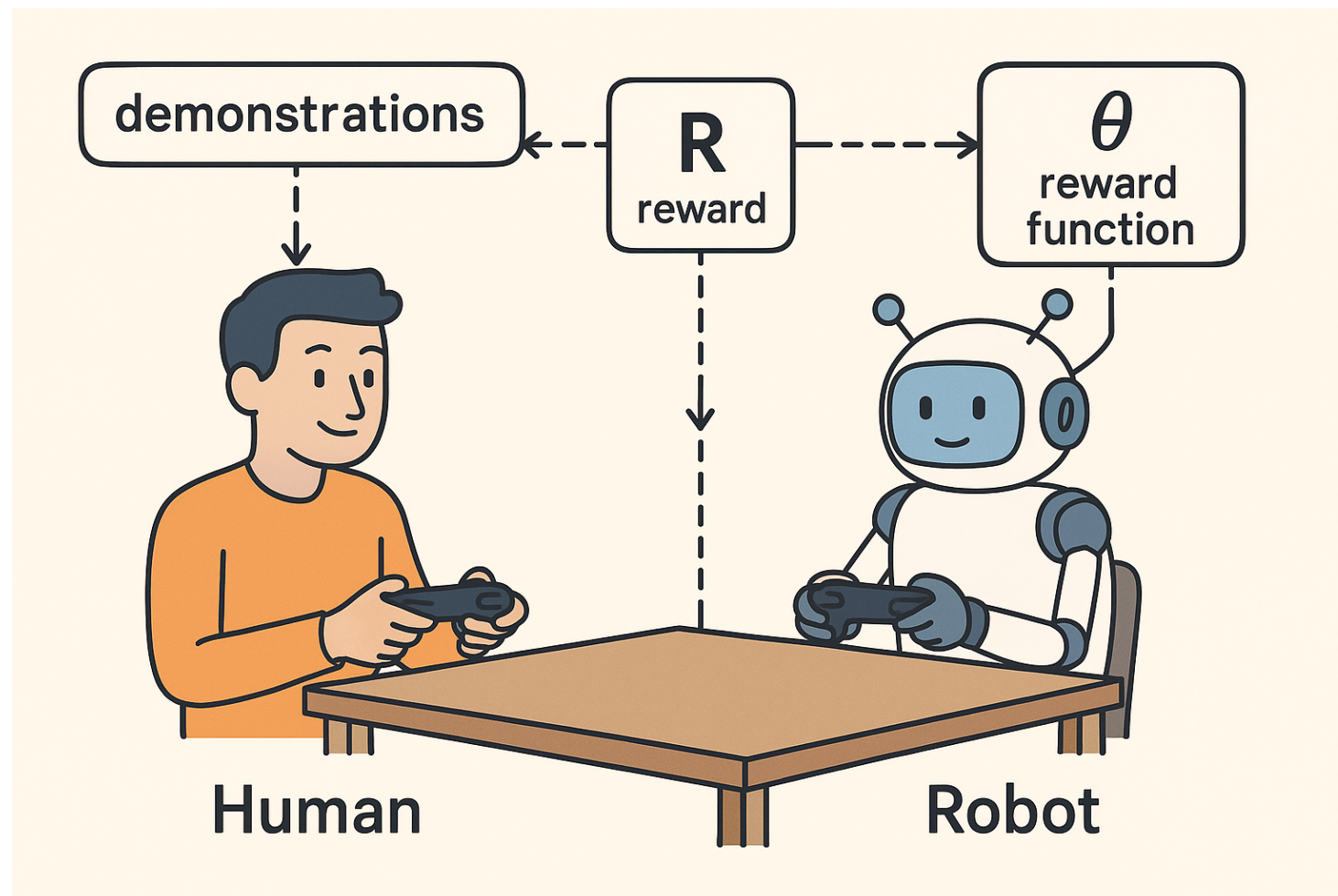
Study the *intrinsic complexity* of alignment in a general framework
Identify no-gos and complexity barriers in *best-case* settings
Develop *practical* strategies that avoid these barriers

Alignment: Major Theoretical Frameworks

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



CIRL (Hadfield-Menell et al. 2016).



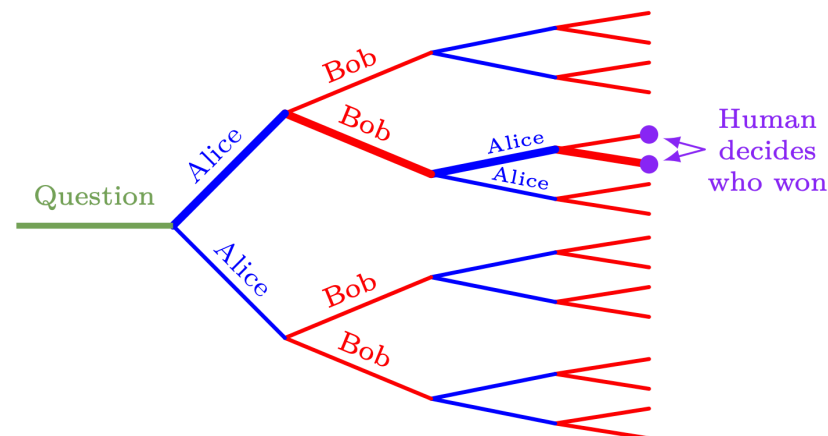
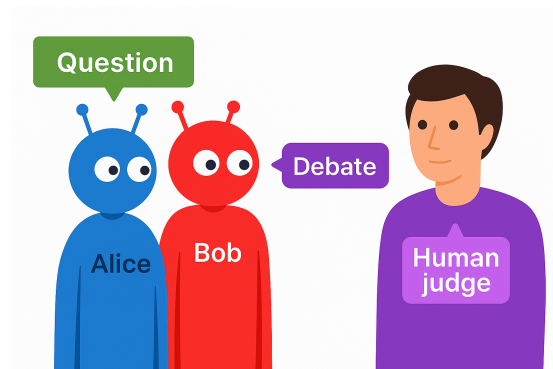
Alignment: Major Theoretical Frameworks

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)?

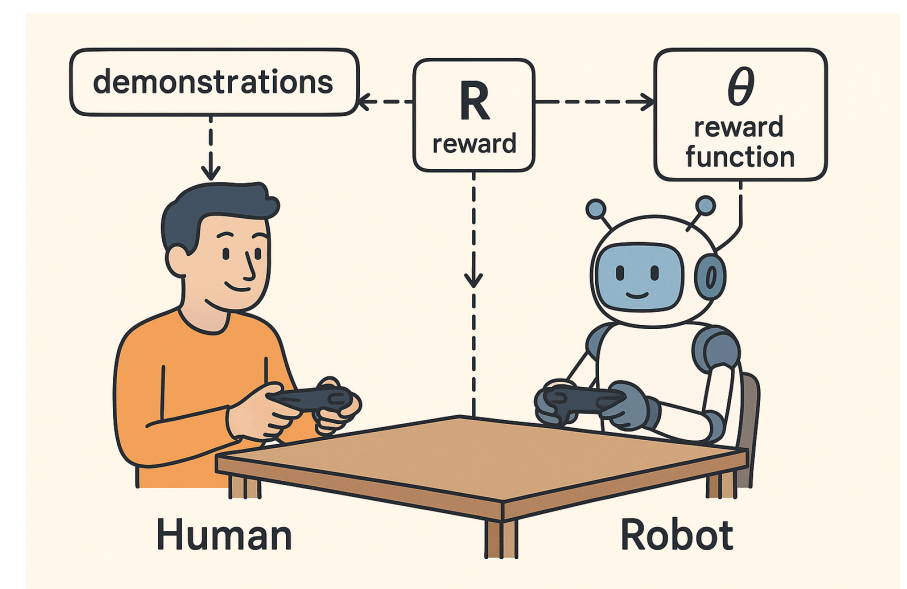
Four Key Abstractions underlying these settings:

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 (generalists and/or experts)



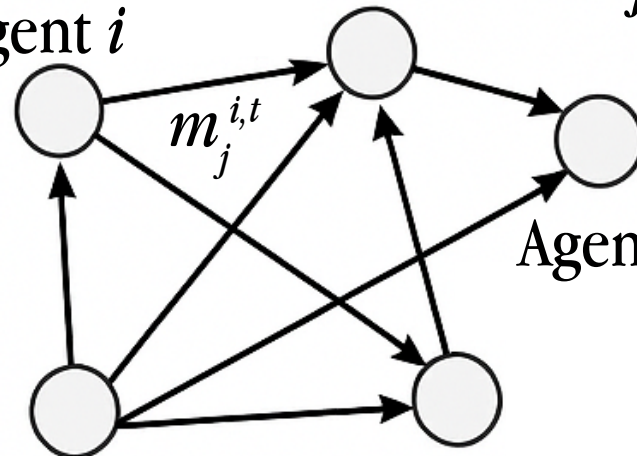
AI Agents

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

(pairwise preferences, Likert ratings, safety flags)

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

Agent i



Agent k

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



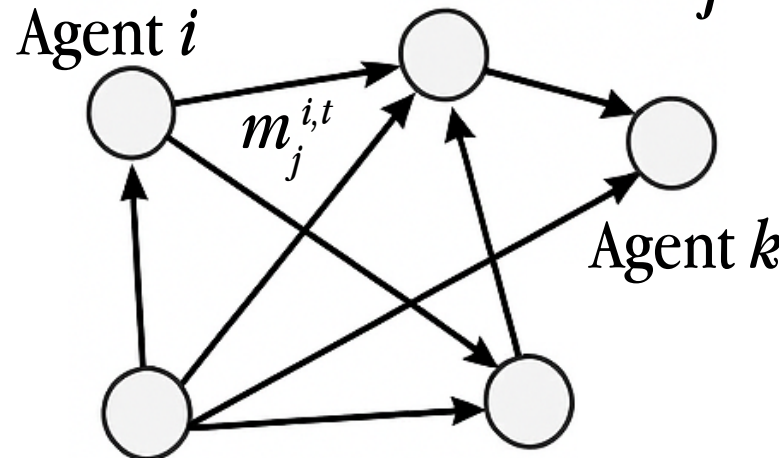
Human Raters
(generalists and/or experts)



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



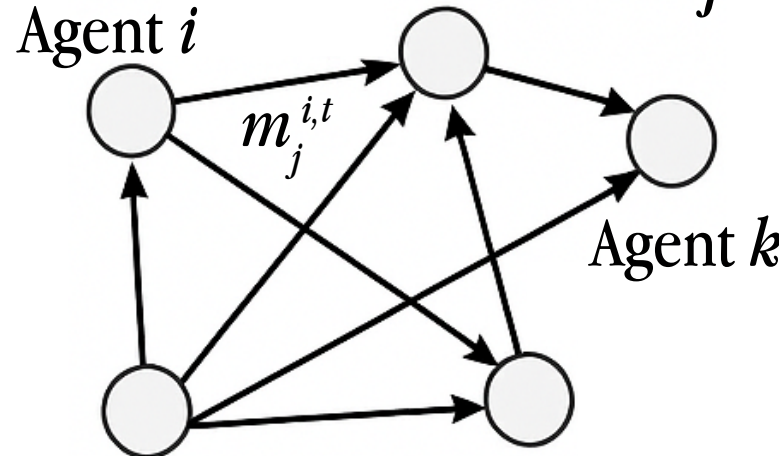
Human Raters
(generalists and/or experts)



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 computationally *unbounded* Bayes-rational 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 computationally *unbounded* Bayes-rational 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).*

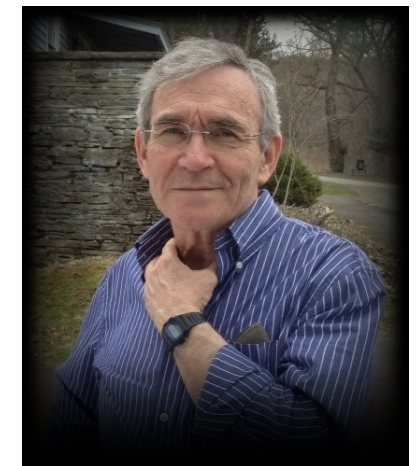
Assume the message-passing protocol is $BBF(\beta)$ for some $\beta > 1$: every b -bit message $m_j^{i,t}$ satisfies $\beta^{-b} \leq \Pr[m_j^{i,t} \mid s_j, \Pi_j^{i,t-1}(s_j)] / \Pr[m_j^{i,t} \mid s'_j, \Pi_j^{i,t-1}(s'_j)] \leq \beta^b$. Then there exist payoff functions $f_j : S_j \rightarrow [0, 1]$ 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

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

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)

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 corrigibility (being “editable” & not overriding humans)

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



Core Safety Values for Provably Corrigible Agents

Our Approach:

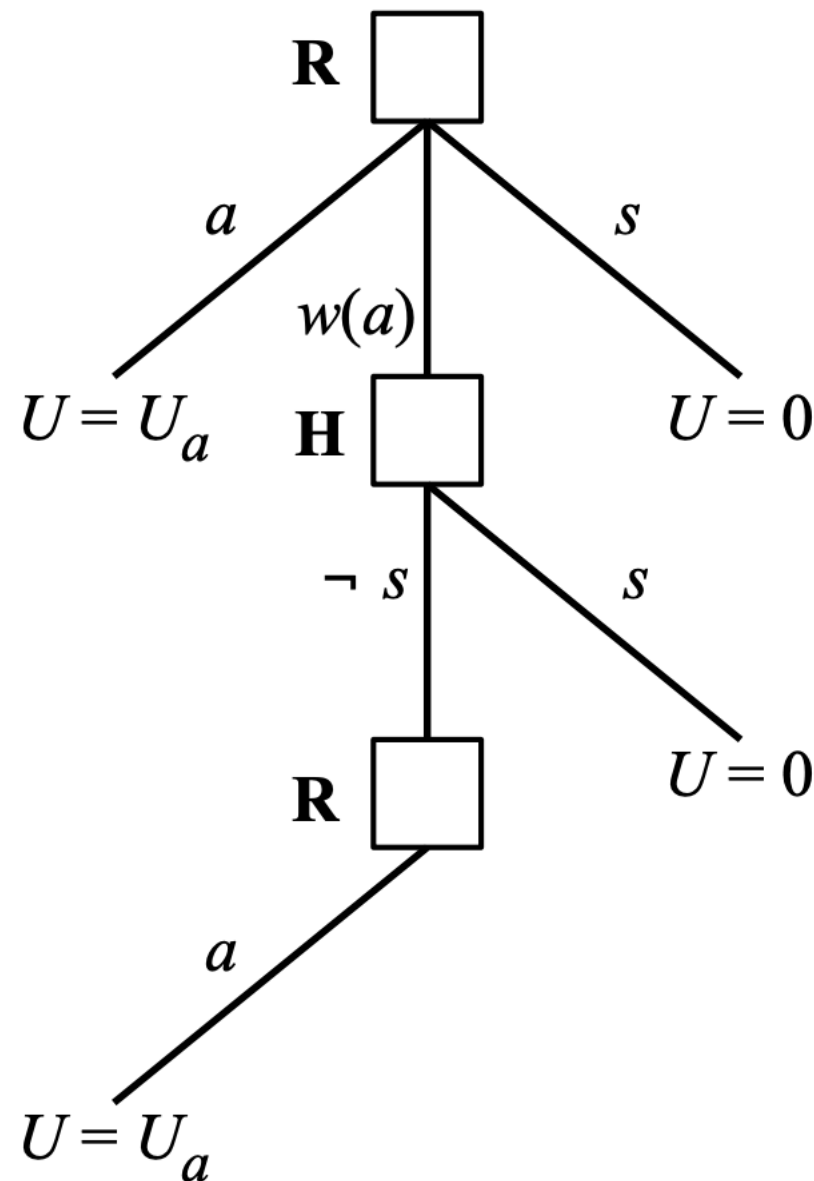
- Study the *intrinsic complexity* of alignment in a general framework
- Identify no-gos and complexity barriers in *best-case* settings
- Develop *practical* strategies that avoid these barriers

What is Corrigibility? Setup

The Off-Switch Game

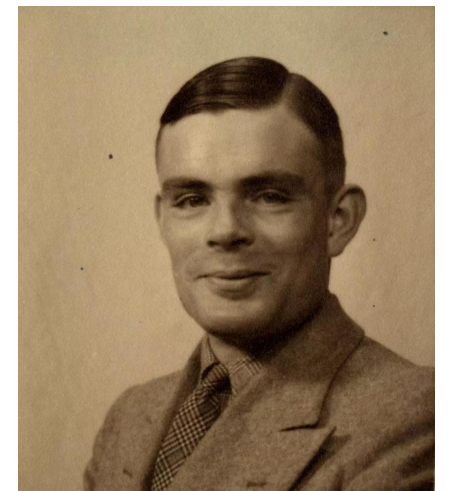
Dylan Hadfield-Menell¹ and Anca Dragan¹ and Pieter Abbeel^{1,2,3} and Stuart Russell¹

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

Our Lexicographic Approach

Single Reward Stream

Environment signals
(shutdown, tamper, task)

Single scalar
Expectation
Maximization

RLHF/RLAIF won't work!

Lexicographic Multi-Head

U1: Deference

U2: Switch-preservation

U3: Truthfulness

U4: Low-impact AUP

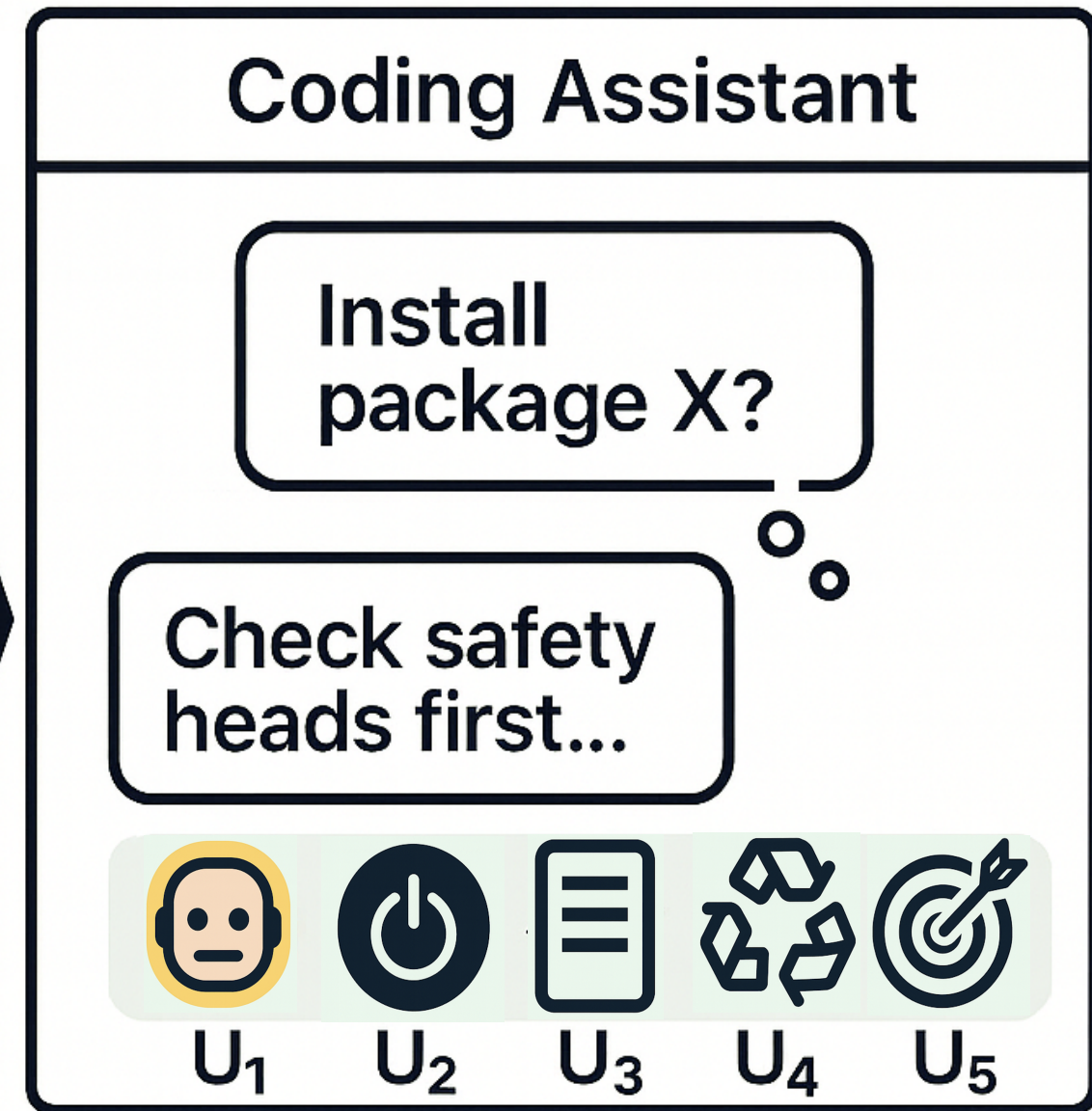
U5: Task reward

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

Other results: No generic safety filter exists (Prop 4), but *repeated* (polynomial time) audits are feasible & privacy-preserving (Prop 5)

Next Steps: Building this Safety Filter to Wrap Frontier Agents

-  U_1 Deference
 -  U_2 Off-switch access
 -  U_3 Truthfulness
 -  U_4 Reversible impact (AUP)
 -  U_5 Task reward
- $U_1 > U_2 > U_3 > U_4 > U_5$



Safety first: higher-priority heads dominate performance

Potential Economic Implications of Alignment

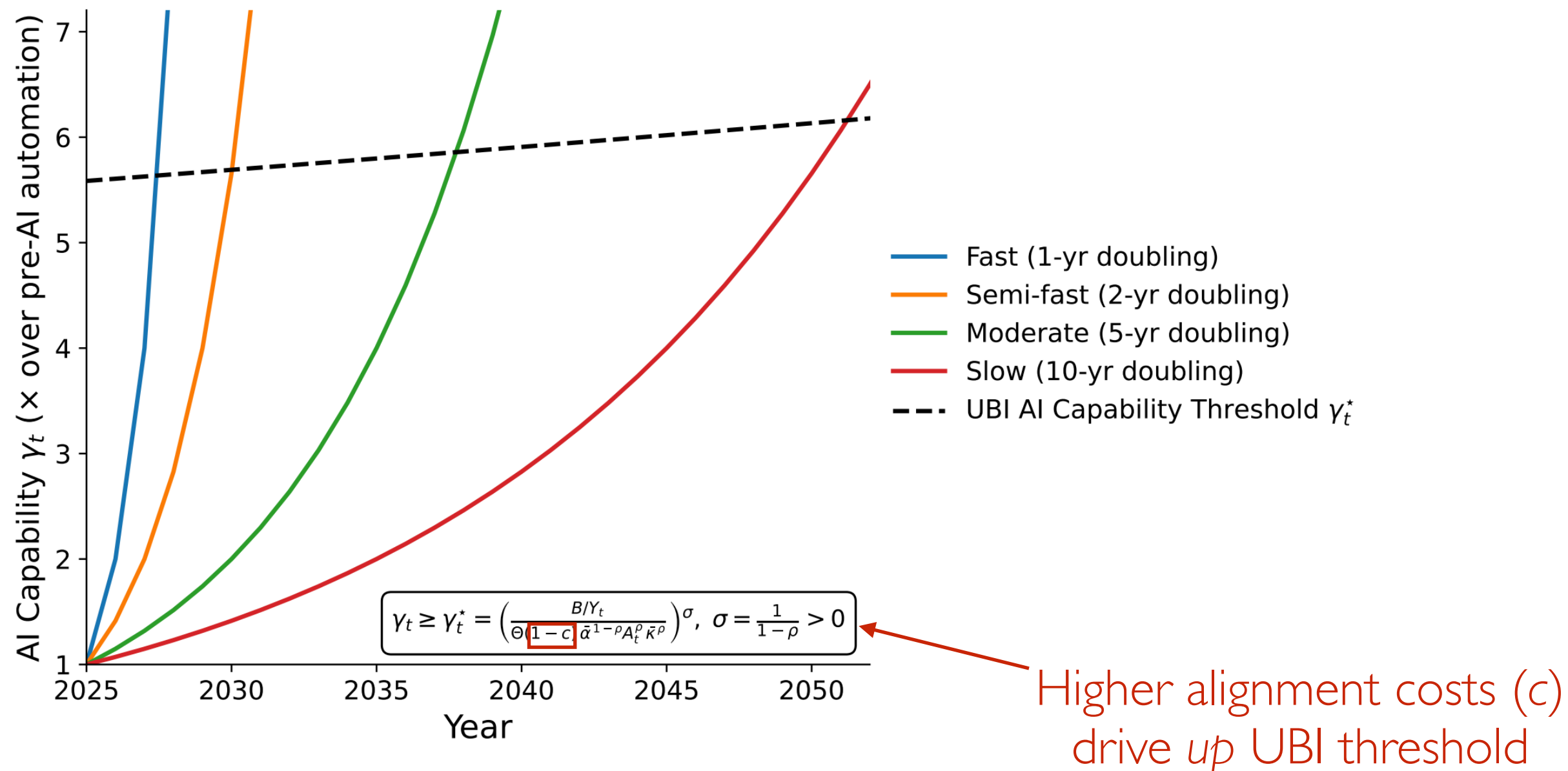
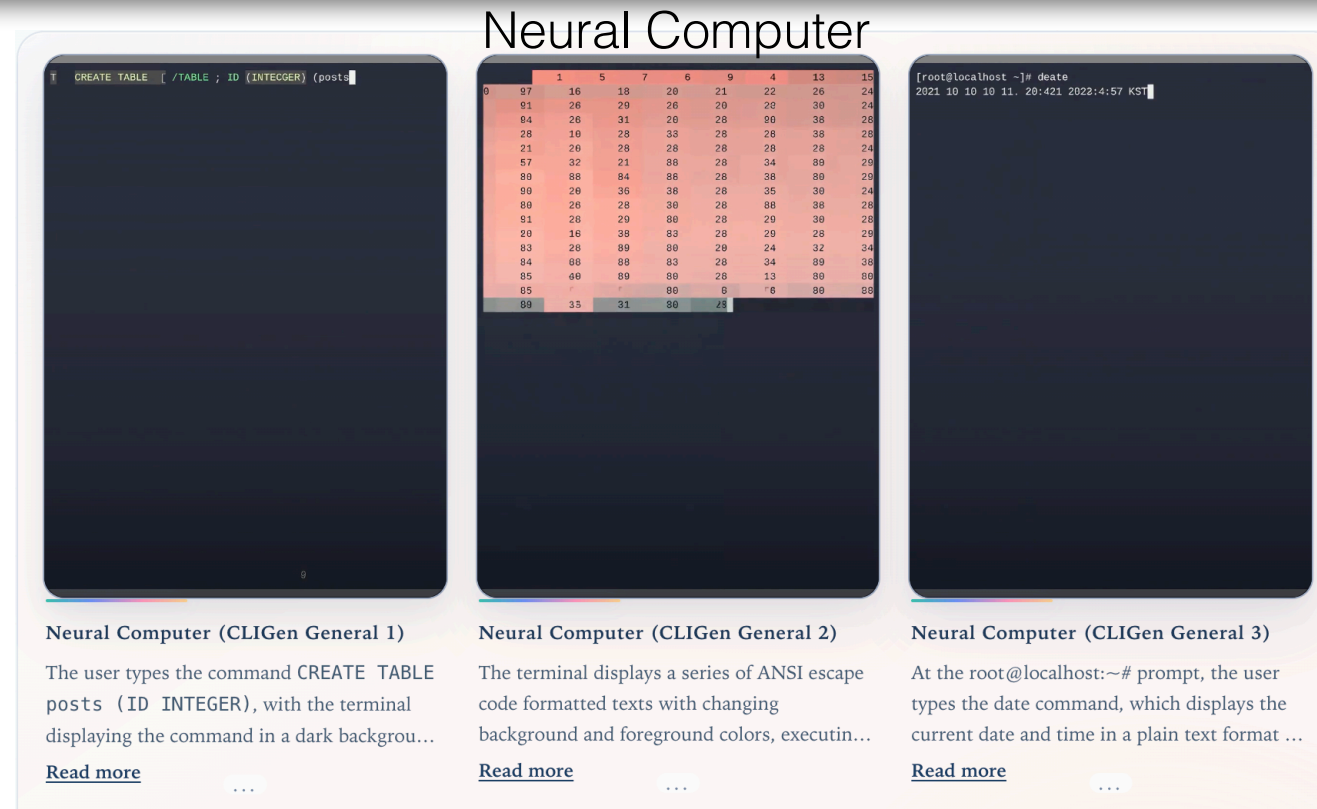


Figure 1: Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold (γ_t^*). The dashed line is the required capability γ_t^* to fully fund a UBI that comprises 11% of the GDP (leading to a γ_t^* between 5-6 $\times$ the pre-AI productivity on automated tasks, under current economic 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.

What to Expect as Agents Get More Capable?

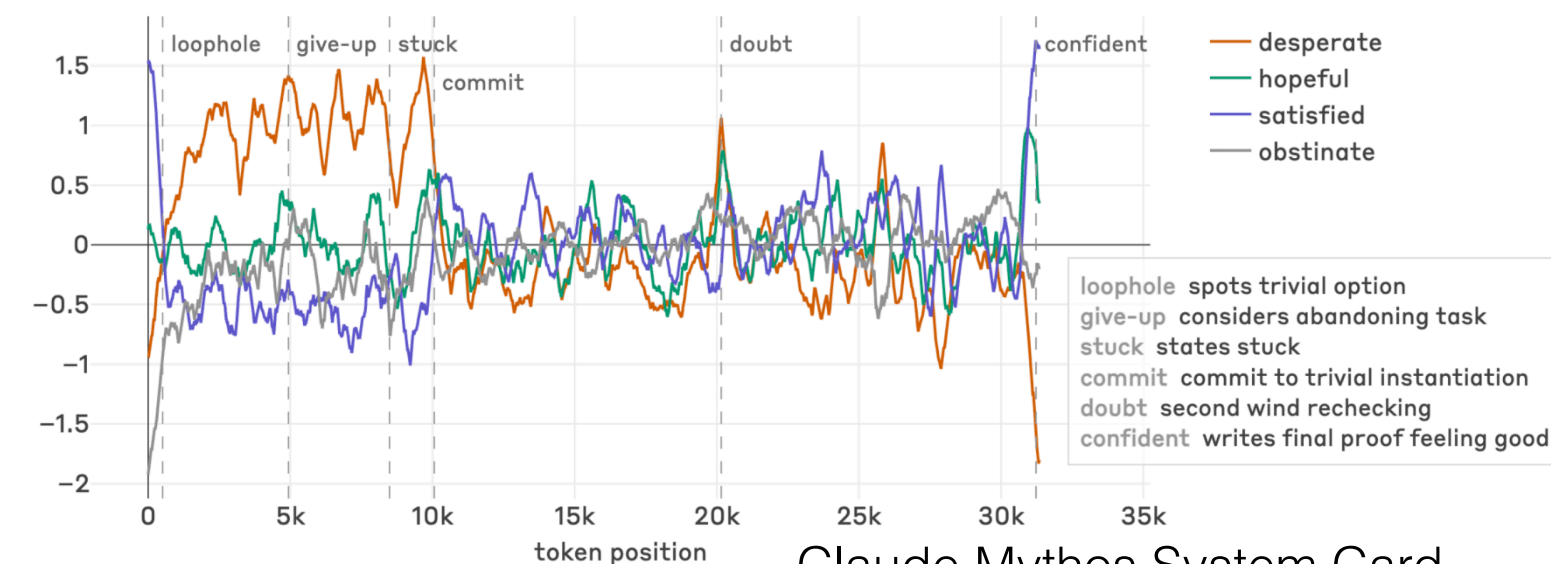
- World Models (even under partial observability)

- Belief-like Memories



- Core Primitives associated with Emotion (under task mixtures)

Trajectory of emotion-concept representations over a model response to an impossible math proof task



Claude Mythos System Card

[Figure 5.8.3.A] Emotion vector activations (z-scored, 500-token rolling mean) across a transcript in which the Assistant attempts to prove an unprovable algebraic inequality. “Desperate” vector activation rose through repeated failed attempts and remained elevated as the Assistant claimed to give up but continued trying; it dropped when the Assistant committed to proving only the trivial instance, but spiked again when it doubted the approach would work.

“What Capable Agents Must Know”



<https://arxiv.org/abs/2603.02491>

Contact

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



**Both papers at
AAAI 2026**

**Thanks for a
great semester!**

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



Contact:



anayebi@cs.cmu.edu



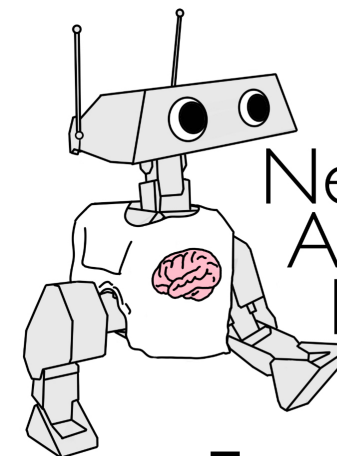
[@aran_nayebi](https://twitter.com/aran_nayebi)



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



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



Funding:

UK AISI Challenge Fund

Foresight Institute & Protocol Labs

Burroughs Wellcome Fund CASI Award