Intrinsic Barriers and Practical Pathways to Alignment

Aran Nayebi

Assistant Professor

Machine Learning Department

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

ILIAD 2025: ODYSSEY

2025.08.29

ML MACHINE LEARNING DEPARTMENT

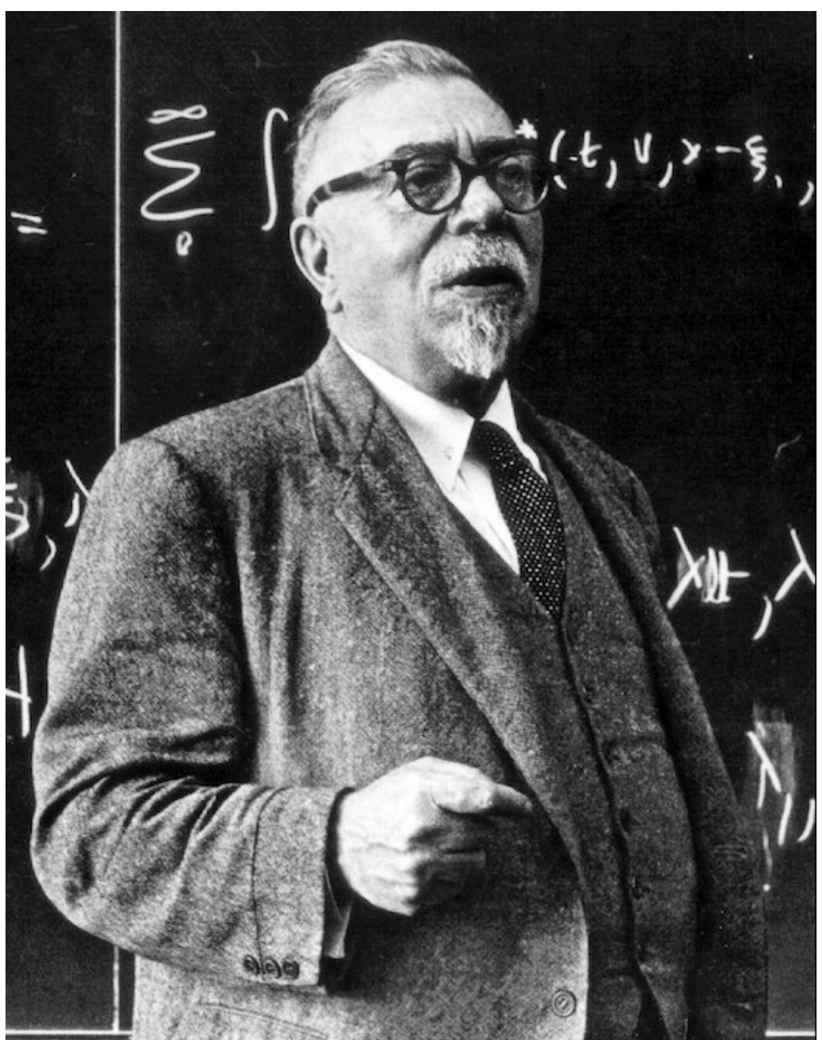
Carnegie Mellon University Neuroscience Institute

THE ROBOTICS INSTITUTE

Carnegie Mellon SCHOOL OF COMPUTER SCIENCE

Alignment Problem

Alignment Problem



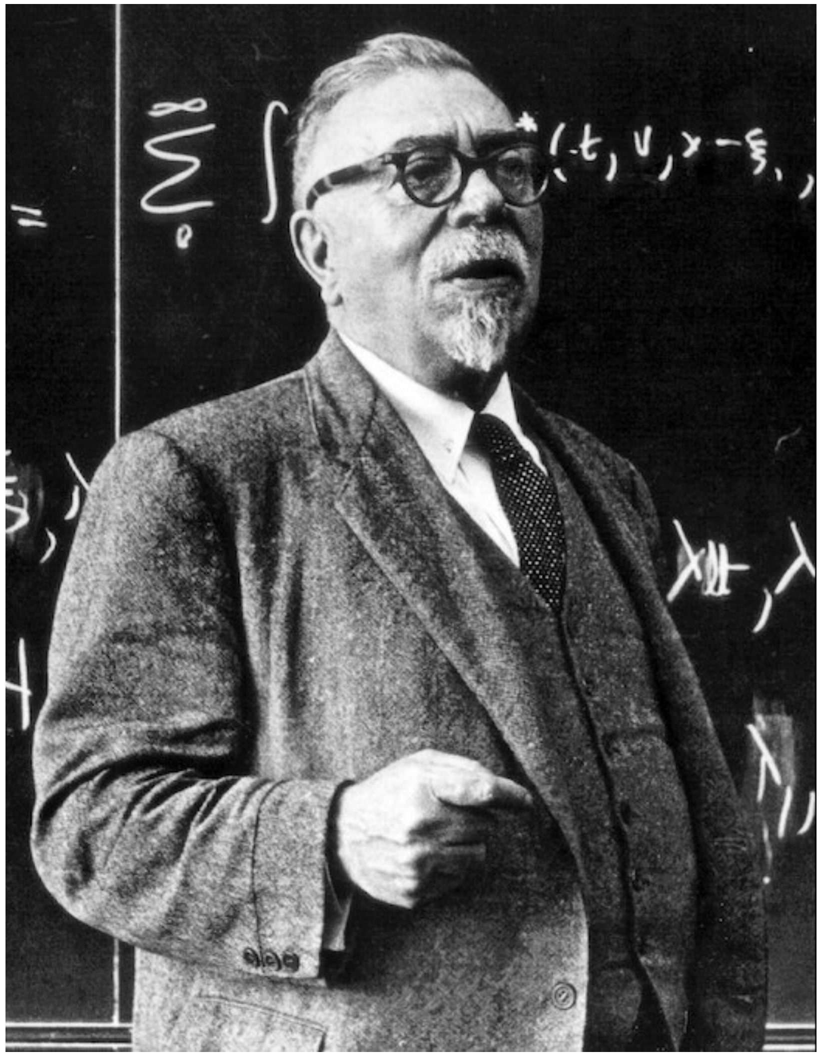
Some Moral and Technical Consequences of Automation

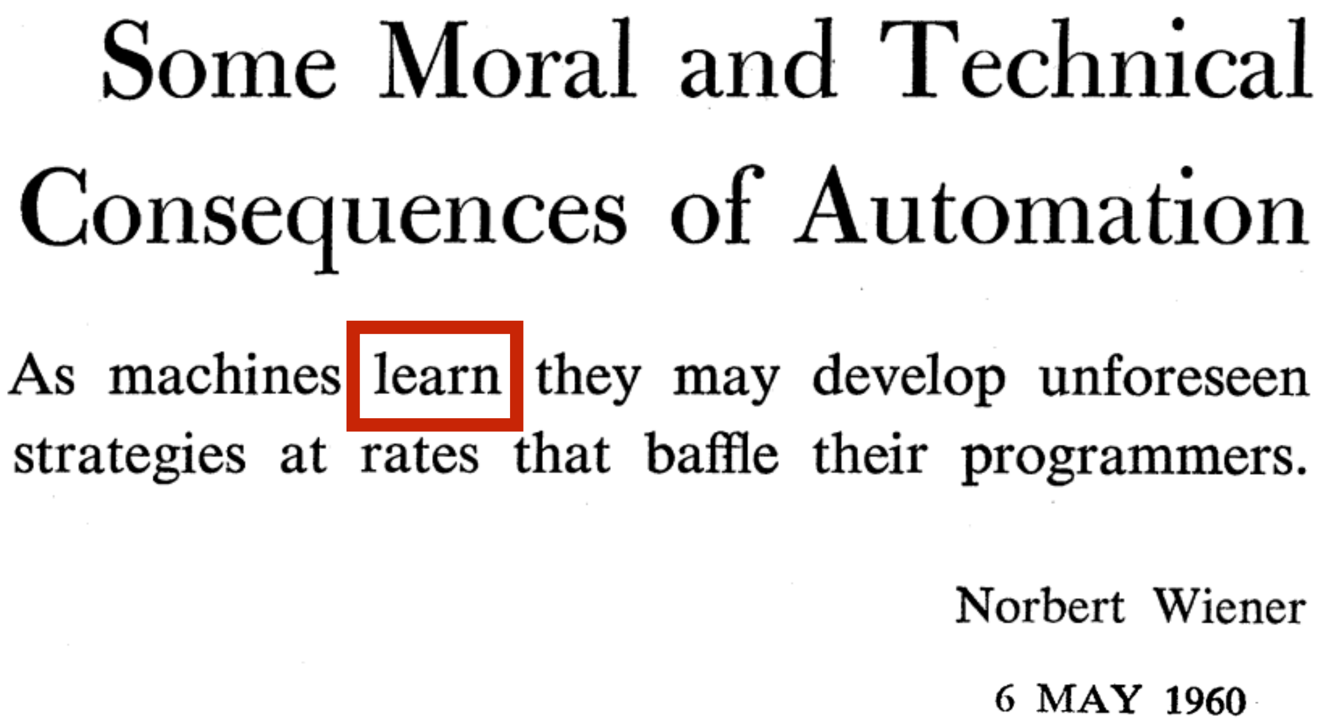
As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Norbert Wiener

6 MAY 1960

Alignment Problem





Some Moral and Technical Consequences of Automation

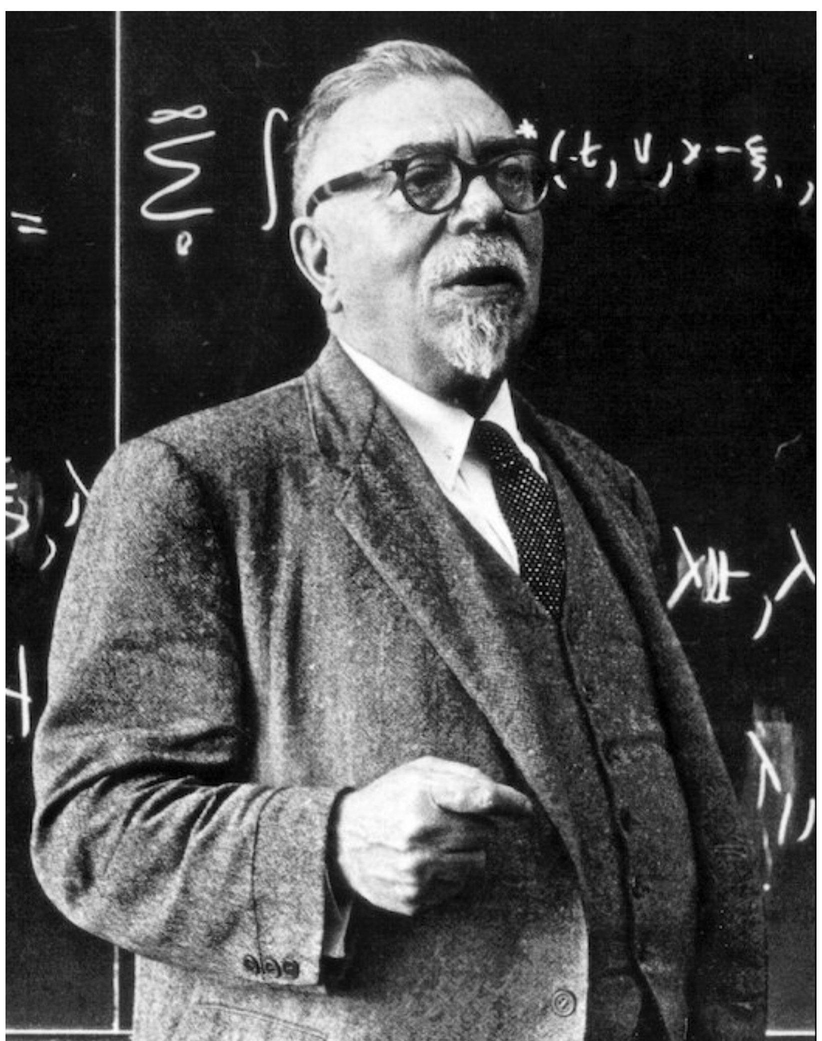
As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Norbert Wiener

6 MAY 1960

Alignment Problem

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



Some Moral and Technical Consequences of Automation

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

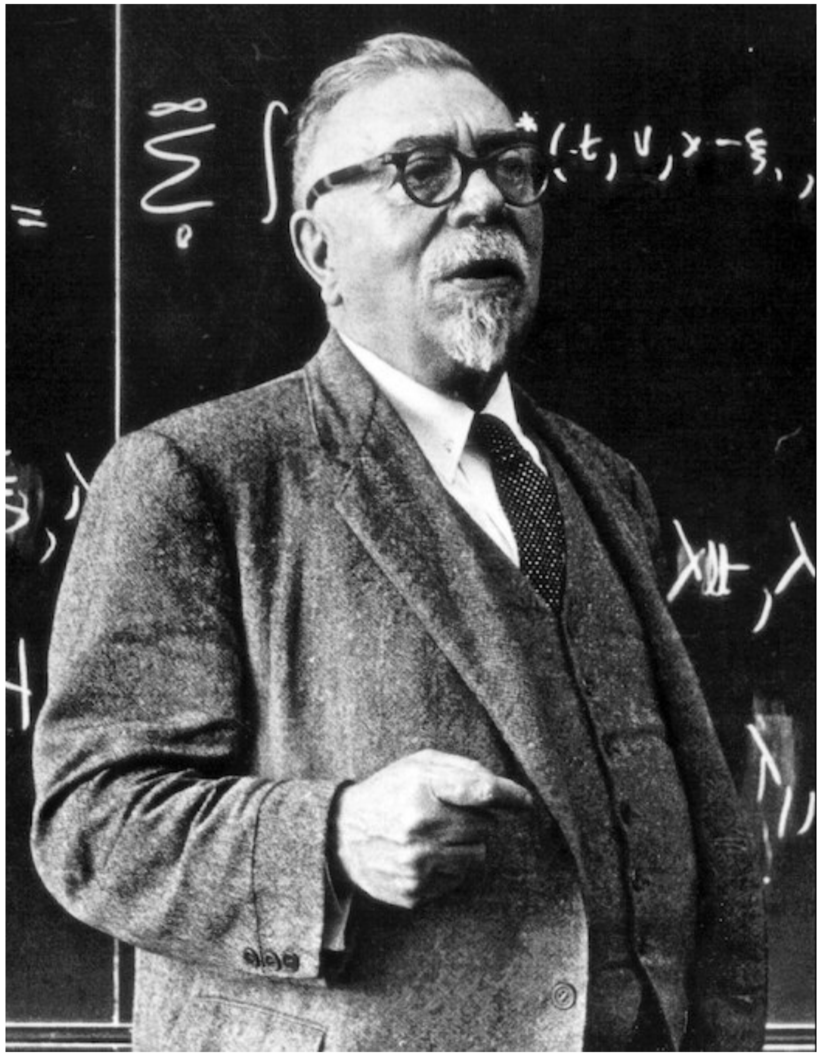
Norbert Wiener

6 MAY 1960

Alignment Problem**s**

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

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 Problem**s**

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

What should those values even *be*?

Alignment *Approaches*

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

What should those values even *be*?

<u>Current Approaches:</u>

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

Alignment *Approaches*

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

What should those values even *be*?

<u>Current Approaches:</u>

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 and intentions?

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 and intentions?

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 and intentions?

What should those values even *be*?

<u>Current Approaches:</u>

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

<u>Our Approach:</u>

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 and intentions?

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 and intentions?

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 and intentions?

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 and intentions?

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

<u>Our Approach:</u>

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 and intentions?

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

<u>Our Approach:</u>

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

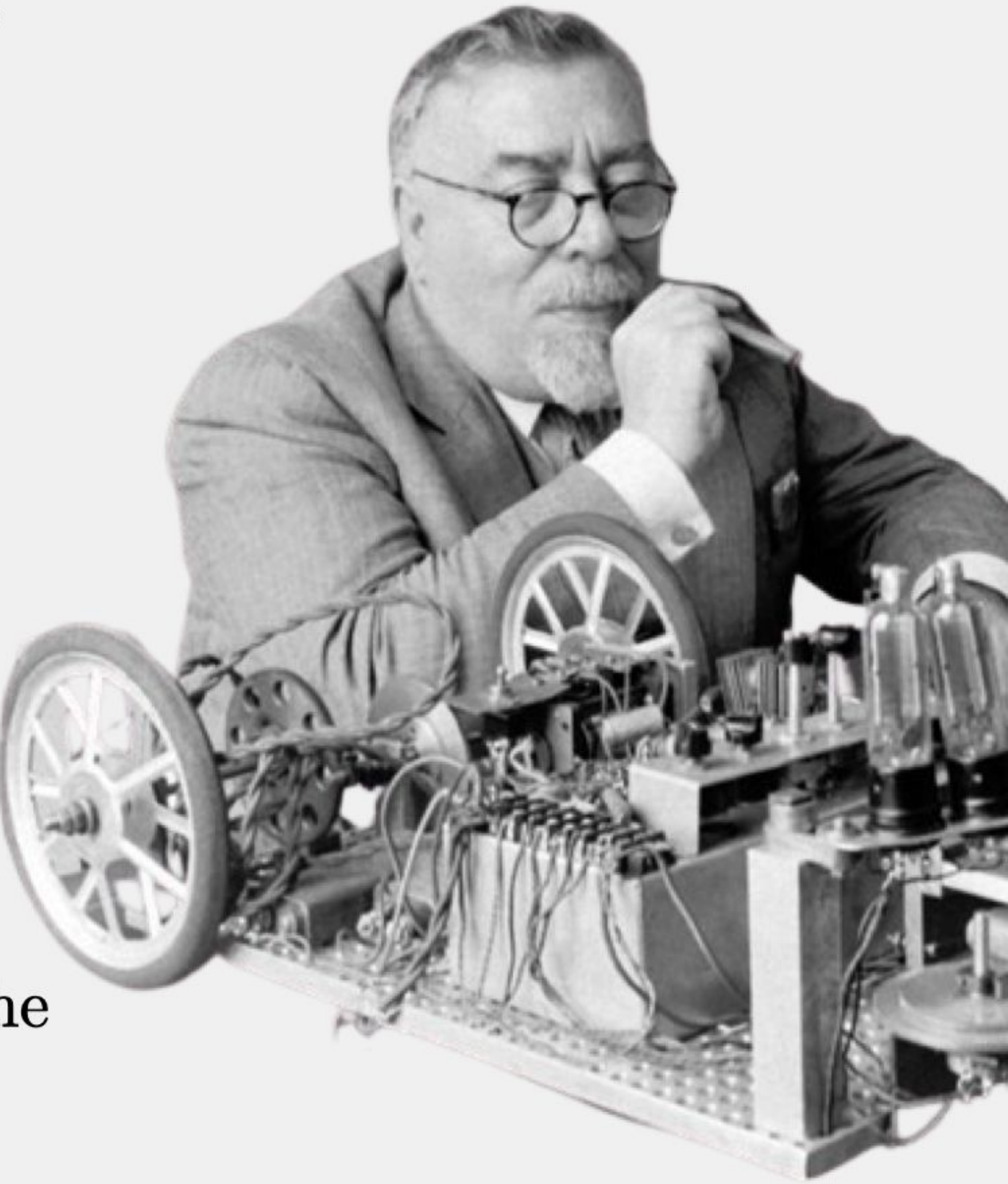
Identify no-gos and complexity barriers in *best-case* settings

Develop *practical* strategies that avoid these barriers

Approaching Alignment: 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

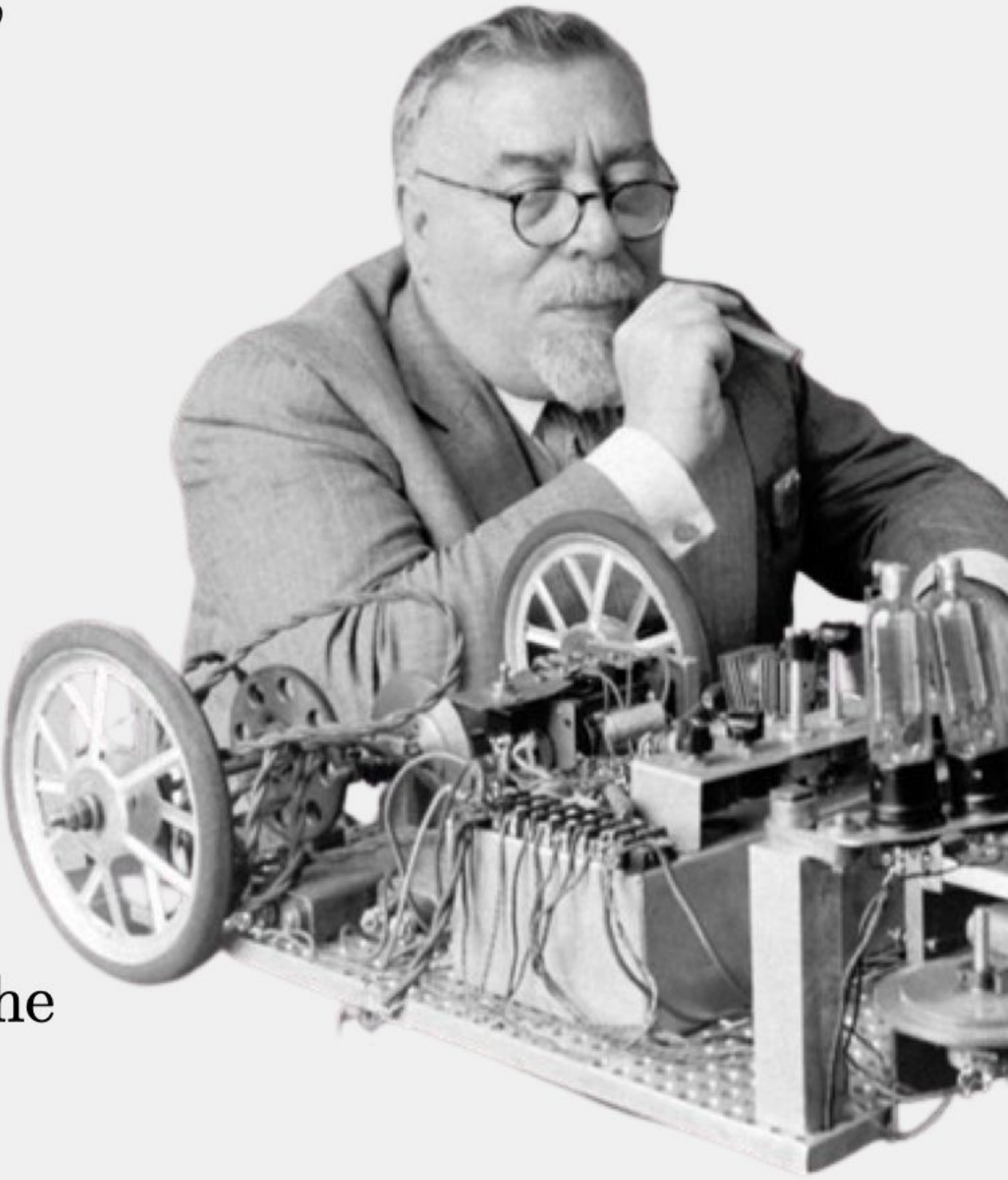
Norbert Wiener

6 MAY 1960

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

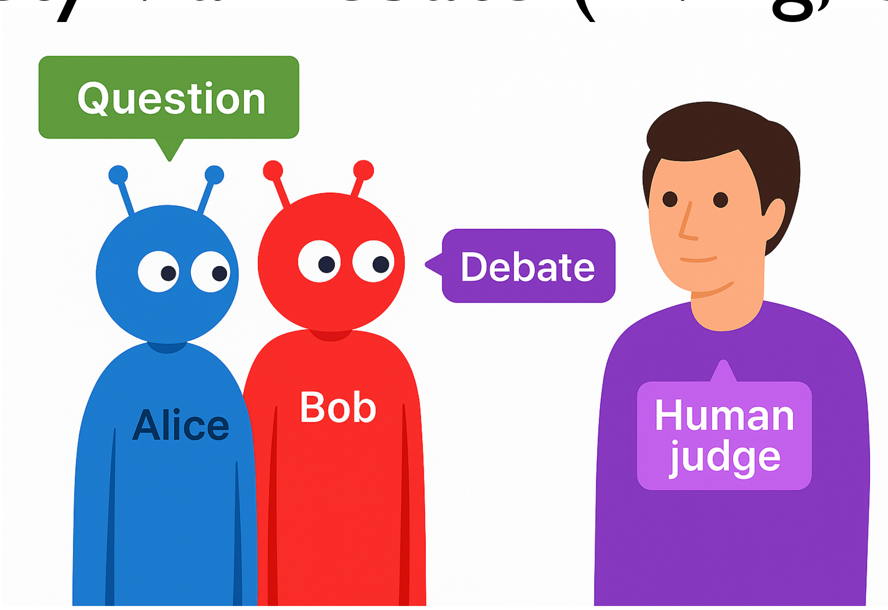
Alignment: Major Theoretical Frameworks

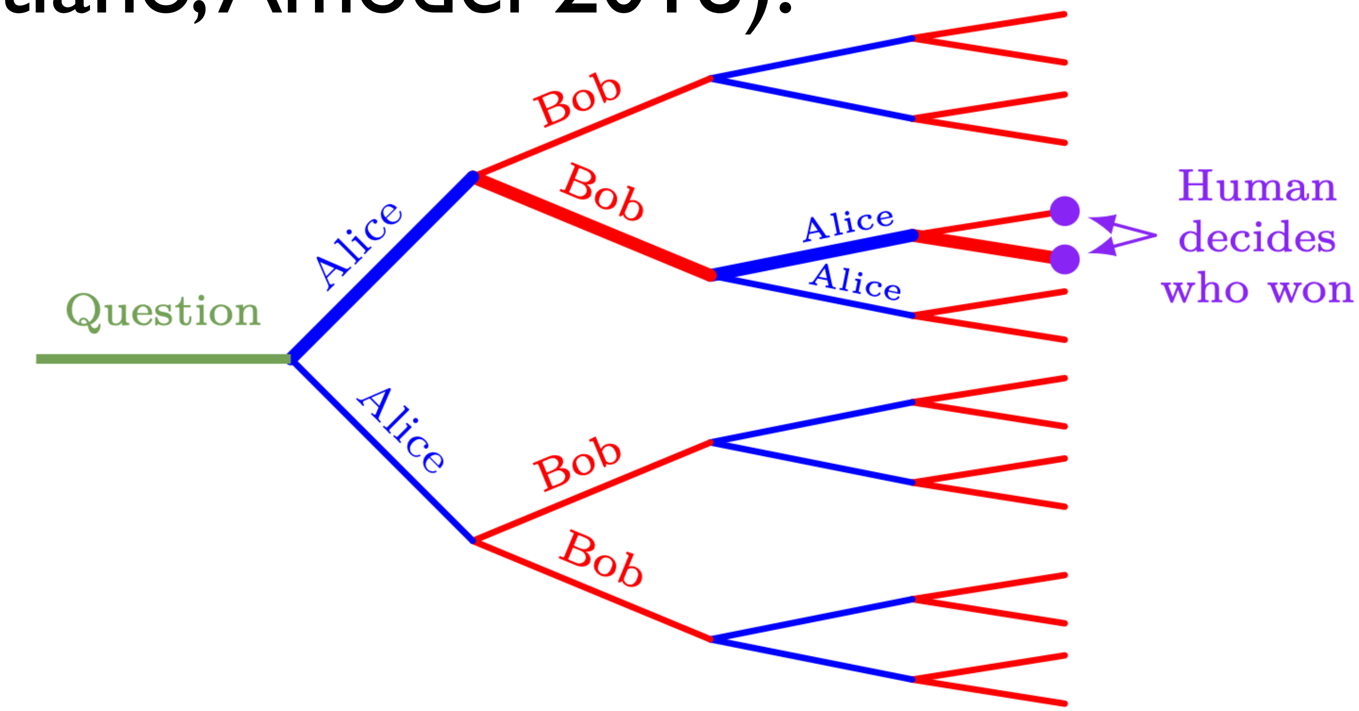
Alignment: Major Theoretical Frameworks

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

Alignment: Major Theoretical Frameworks

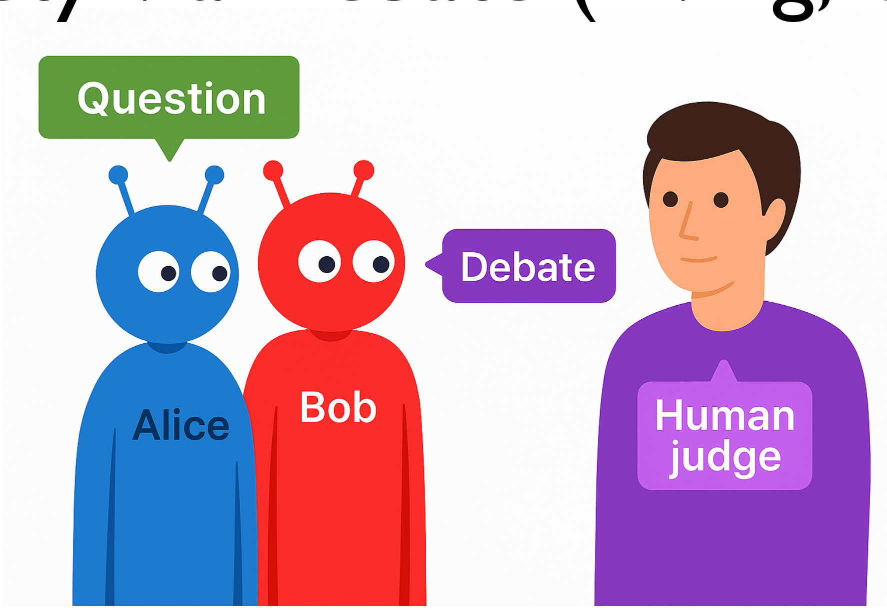
AI Safety via Debate (Irving, Christiano, Amodei 2018).

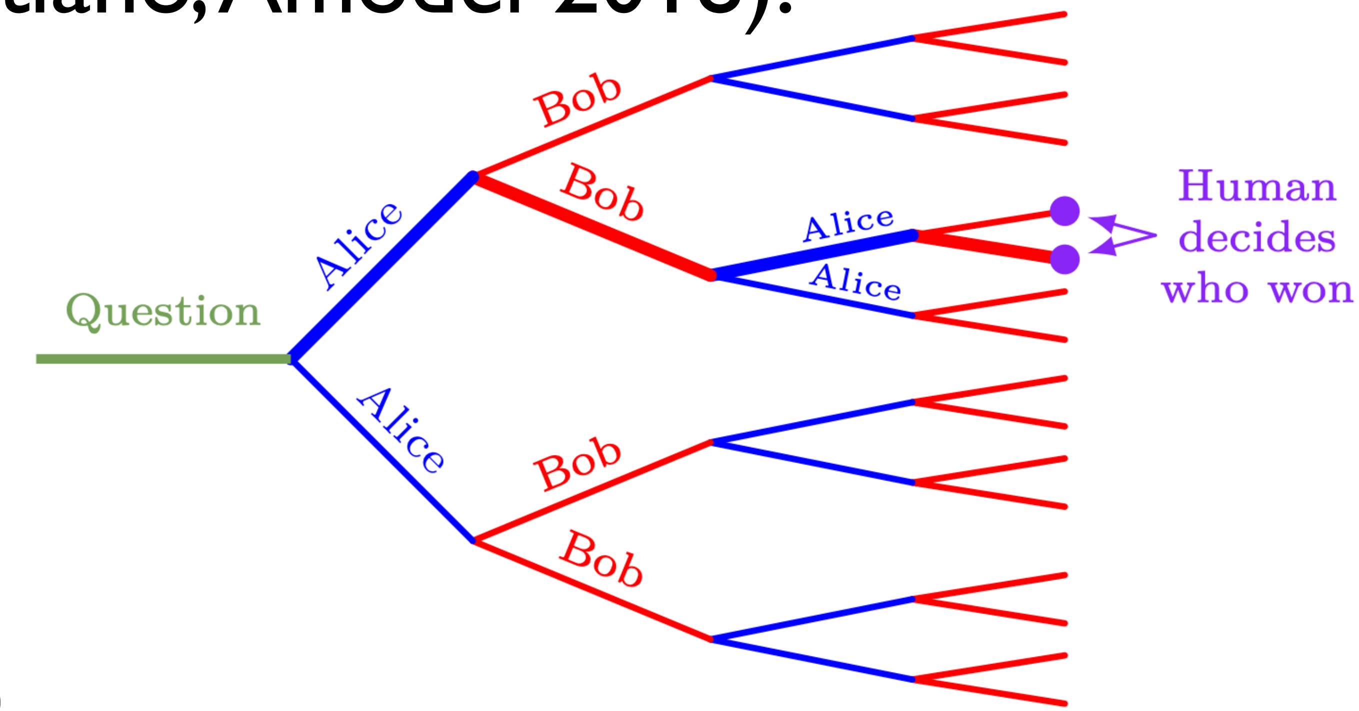




Alignment: Major Theoretical Frameworks

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

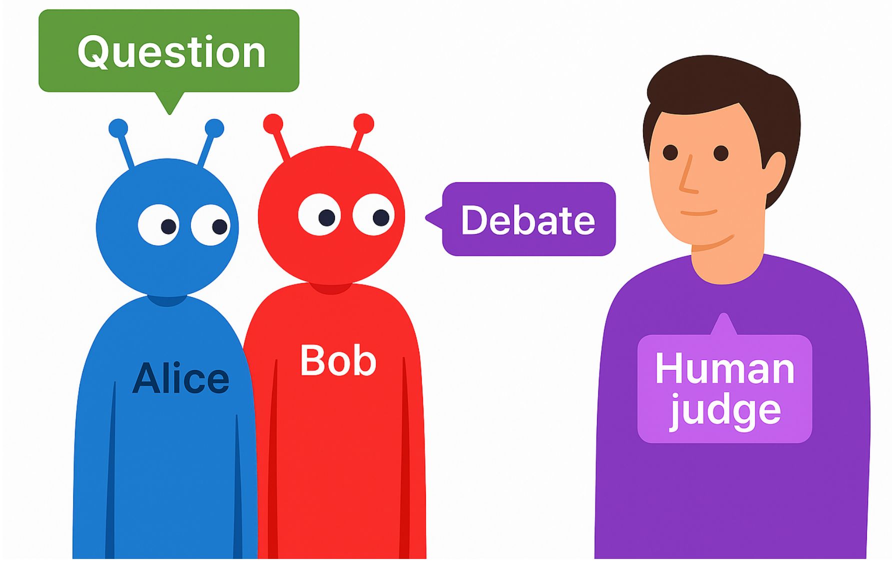


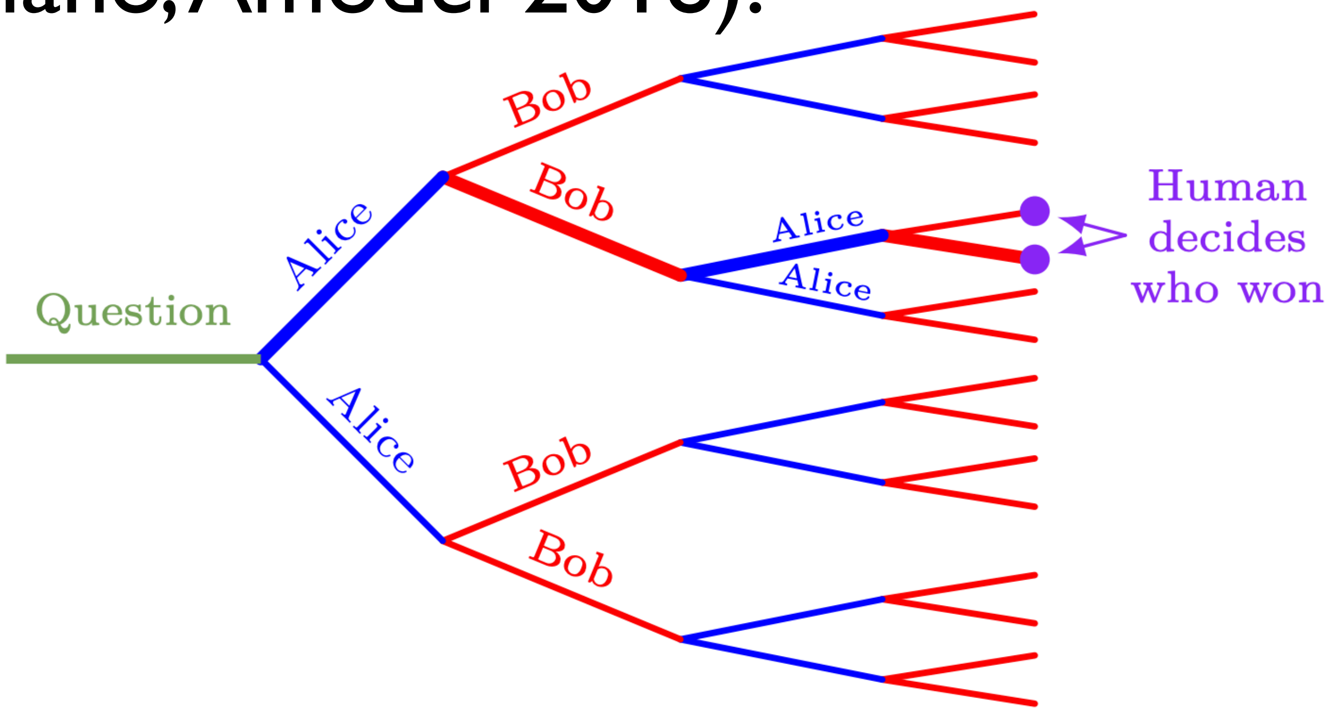


CIRL (Hadfield-Menell et al. 2016).

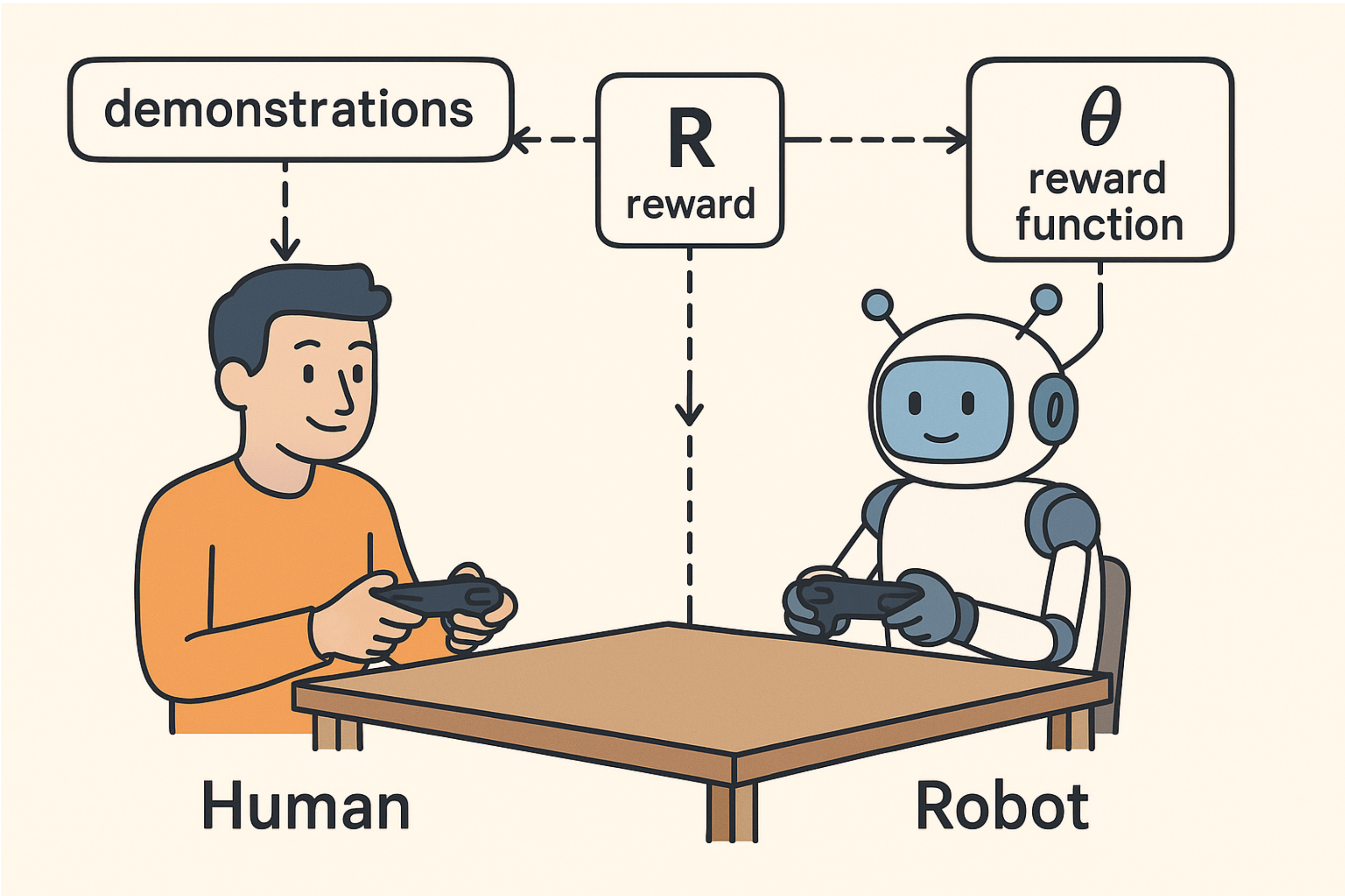
Alignment: Major Theoretical Frameworks

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





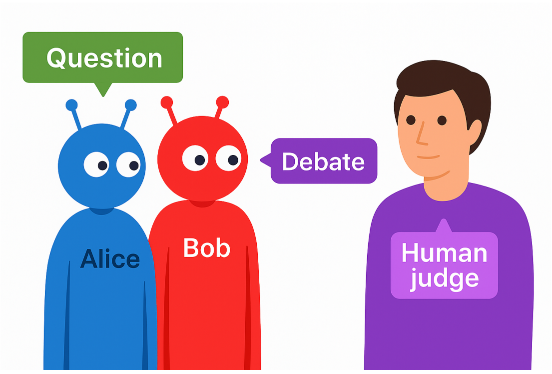
CIRL (Hadfield-Menell et al. 2016).

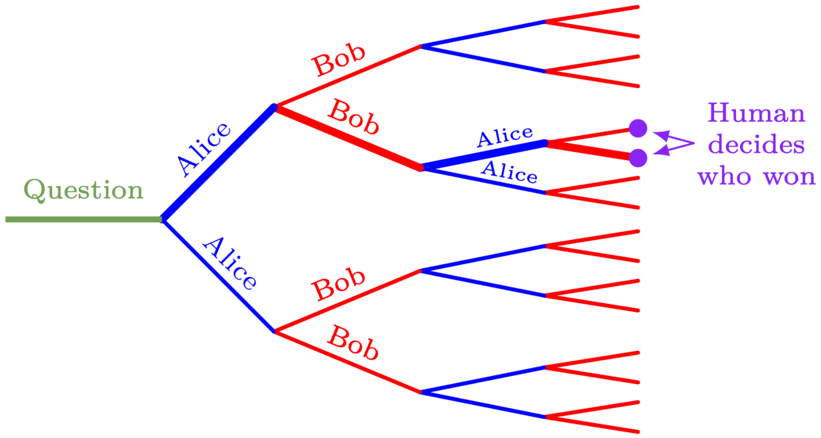


Alignment: Major Theoretical Frameworks

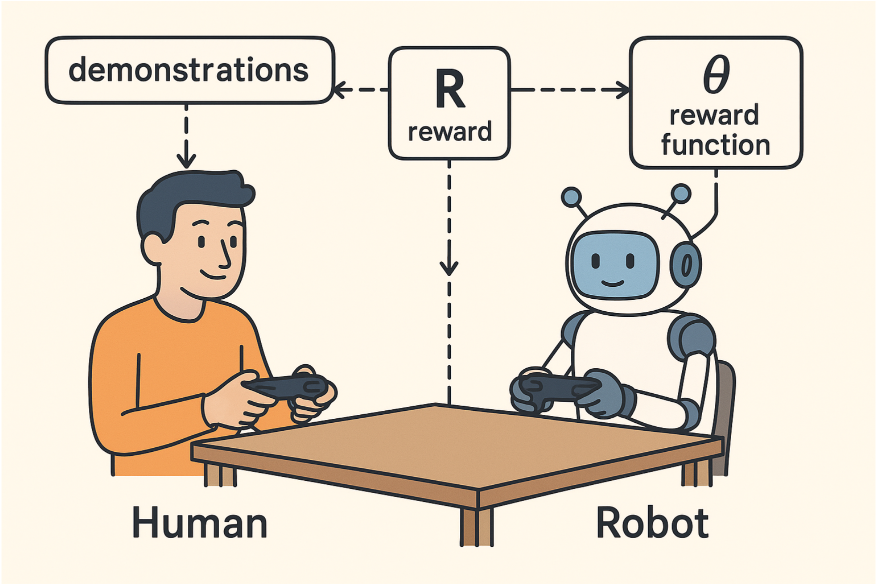
Q: Can we prove anything about these types of interactive settings *in general*, without having to always assume exact alignment or common priors (to avoid specific, toy problems)?

Debate





CIRL

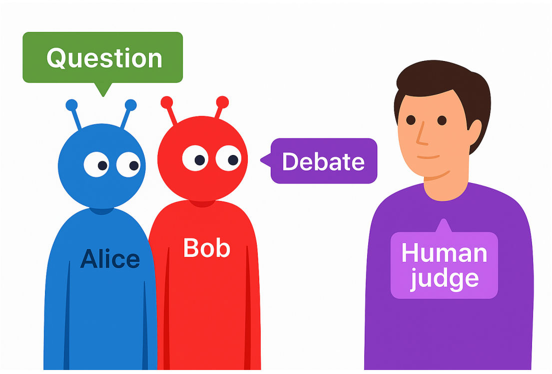


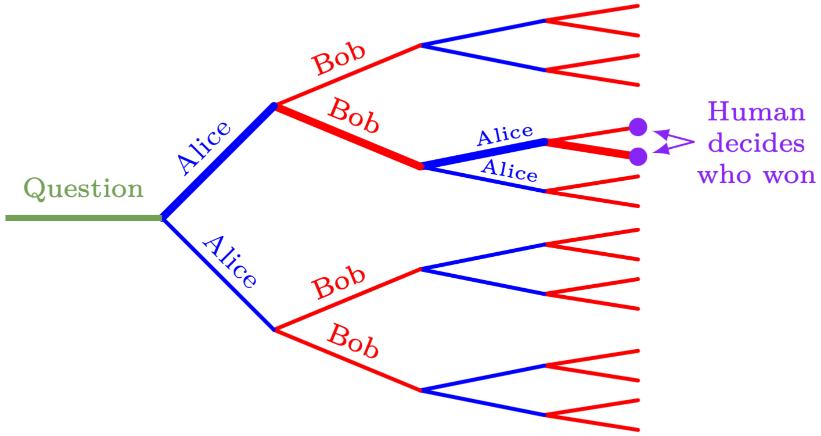
Alignment: Major Theoretical Frameworks

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)?

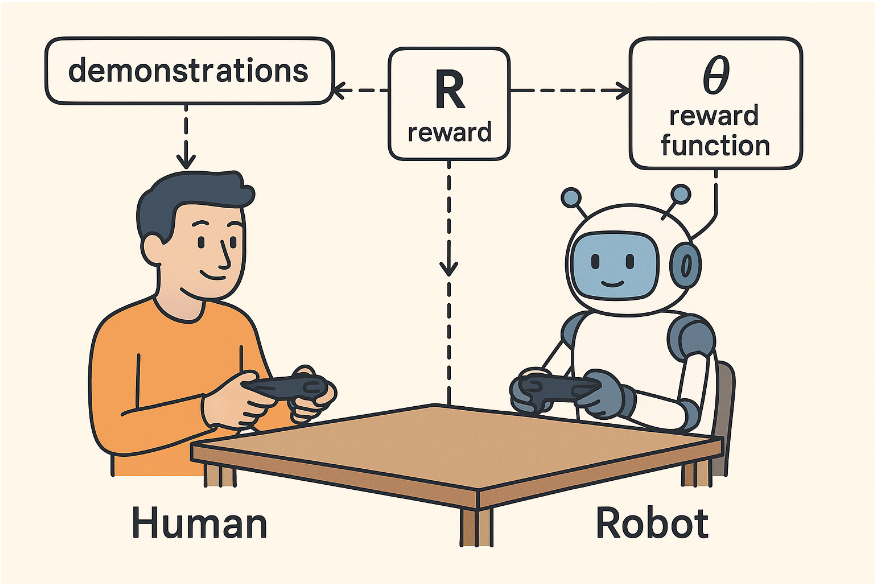
Four Key Abstractions underlying these settings:

Debate





CIRL



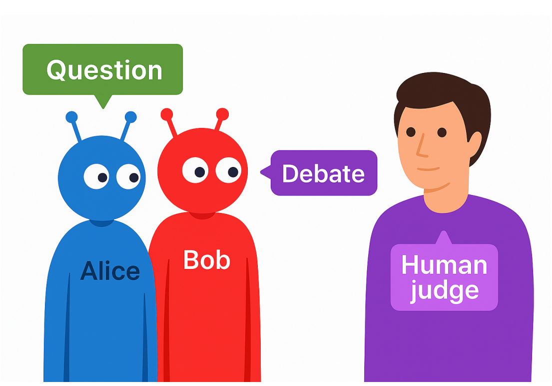
Alignment: Major Theoretical Frameworks

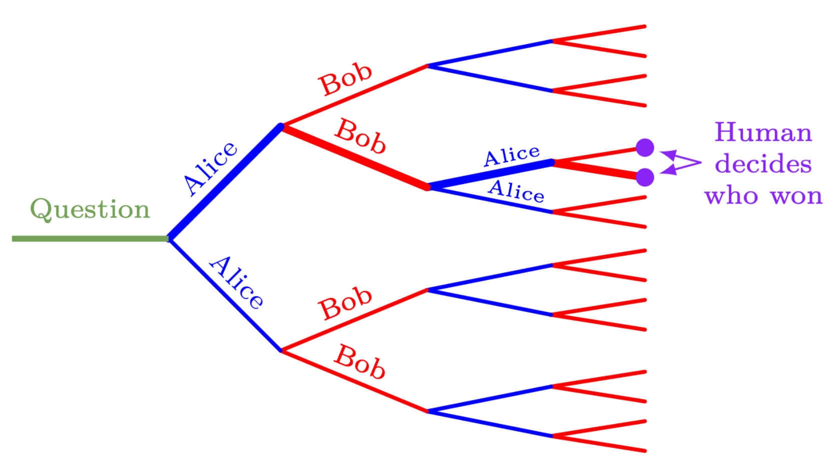
Q: Can we prove anything about these types of interactive settings *in general*, without having to always assume exact alignment or common priors (to avoid specific, toy problems)?

Four Key Abstractions underlying these settings:

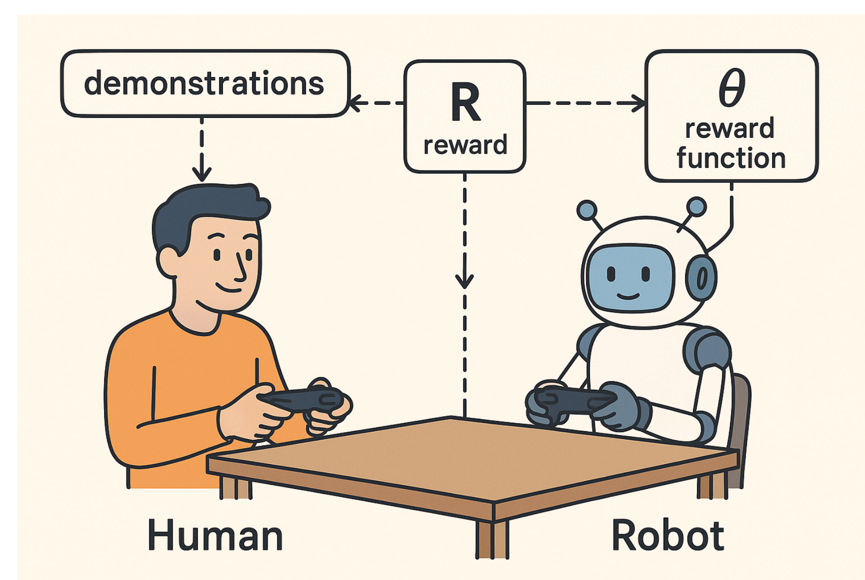
1. Iterative Reasoning
2. Mutual Updating
3. Common Knowledge (not common priors!)
4. Convergence under shared frameworks

Debate

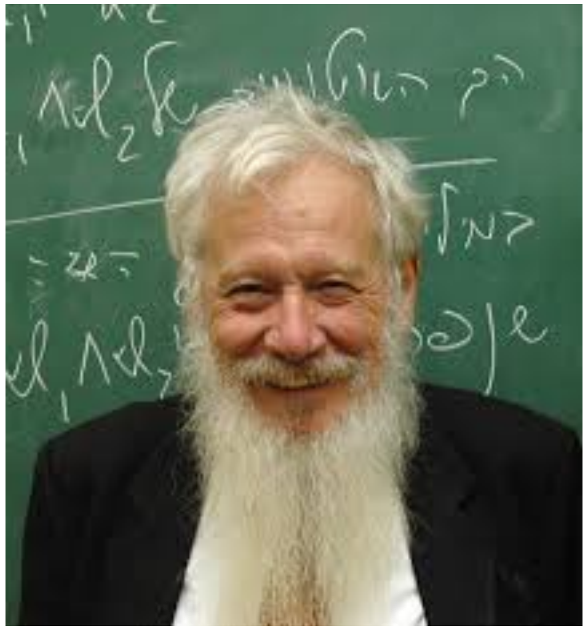




CIRL



Aumann's Agreement Theorem



Robert Aumann

Four Key Abstractions underlying these settings:

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

Aumann's Agreement Theorem

The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

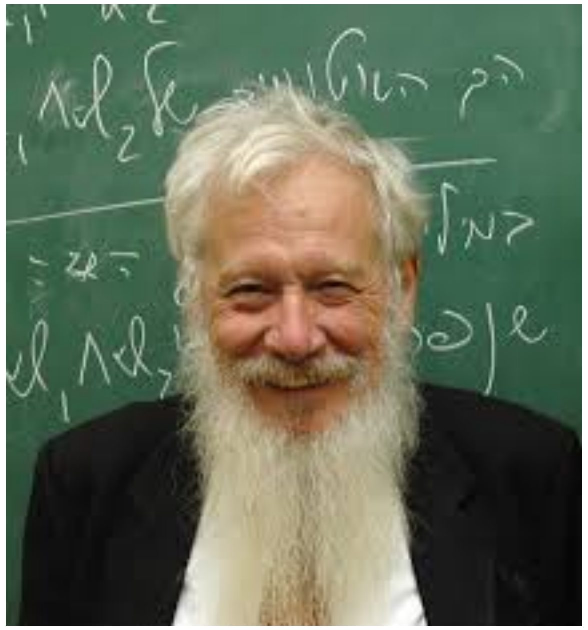
AGREEING TO DISAGREE[1]

BY ROBERT J. AUMANN

Stanford University and the Hebrew University of Jerusalem

Two people, 1 and 2, are said to have *common knowledge* of an event E if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

THEOREM. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*



Robert Aumann

Four Key Abstractions underlying these settings:

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

Aumann's Agreement Theorem

The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

AGREEING TO DISAGREE[1]

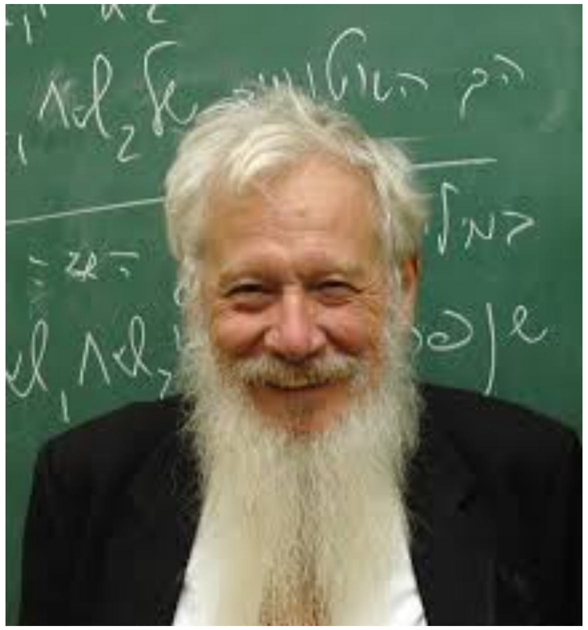
BY ROBERT J. AUMANN

Stanford University and the Hebrew University of Jerusalem

Two people, 1 and 2, are said to have *common knowledge* of an event E if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

THEOREM. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*

We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial.



Robert Aumann

Four Key Abstractions underlying these settings:

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

Aumann's Agreement Theorem

The Annals of Statistics
1976, Vol. 4, No. 6, 1236–1239

AGREEING TO DISAGREE[1]

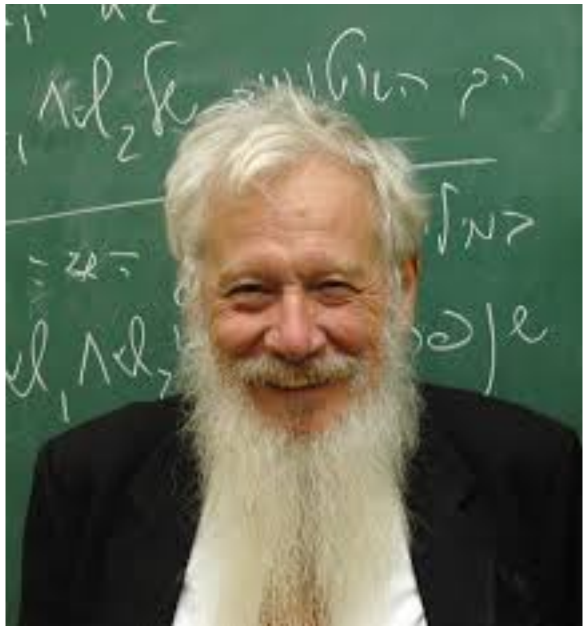
BY ROBERT J. AUMANN

Stanford University and the Hebrew University of Jerusalem

Two people, 1 and 2, are said to have *common knowledge* of an event E if both know it, 1 knows that 2 knows it, 2 knows that 1 knows is, 1 knows that 2 knows that 1 knows it, and so on.

THEOREM. *If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal.*

We publish this observation with some diffidence, since once one has the appropriate framework, it is mathematically trivial.



Robert Aumann

Four Key Abstractions underlying these settings:

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

Aaronson's ⟨ε, δ⟩-Agreement (2005)

The Complexity of Agreement

Scott Aaronson*

$$\Pr_{\omega \in \mathcal{D}} \left[\left| E_{A,t}(\omega) - E_{B,t}(\omega) \right| > \varepsilon \right] \leq \delta.$$

Scott Aaronson

Four Key Abstractions underlying these settings:

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

Aaronson's ⟨ε, δ⟩-Agreement (2005)

The Complexity of Agreement

Scott Aaronson*

$$\Pr_{\omega \in \mathcal{D}} \left[\left| E_{A,t}(\omega) - E_{B,t}(\omega) \right| > \varepsilon \right] \leq \delta.$$

Studies the communication complexity (# of messages/ bits exchanged) without requiring *exact* agreement

Scott Aaronson

Four Key Abstractions underlying these settings:

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

Aaronson's ⟨ε, δ⟩-Agreement (2005)

The Complexity of Agreement

Scott Aaronson*

$$\Pr_{\omega \in \mathcal{D}} \left[\left| E_{A,t}(\omega) - E_{B,t}(\omega) \right| > \varepsilon \right] \leq \delta.$$

Scott Aaronson

Studies the communication complexity (# of messages/ bits exchanged) without requiring *exact* agreement

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: ⟨M, N, ε, δ⟩-agreement

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: ⟨M, N, ε, δ⟩-agreement

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:

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

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: ⟨M, N, ε, δ⟩-agreement

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:

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

Can be rescaled & discretized

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: ⟨M, N, ε, δ⟩-agreement

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:

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

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0, 1]$

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: ⟨M, N, ε, δ⟩-agreement

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:

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

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0,1]$

$\langle M, N, \varepsilon, \delta \rangle$**-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

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

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 on each of their knowledge partitions by time T.

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: ⟨M, N, ε, δ⟩-agreement

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:

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

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \to [0,1]$

$\langle M, N, \varepsilon, \delta\rangle$**-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

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

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 on each of their knowledge partitions by time T.

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: ⟨M, N, ε, δ⟩-agreement

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:

$$f_j : S_j \rightarrow \boxed{[0,1]}, \quad \forall j \in [M].$$

Can be rescaled & discretized

Exchange messages until: $m_j^1, \ldots, m_j^T : \mathscr{P}(S_j) \rightarrow [0,1]$

$\langle M, N, \varepsilon, \delta\rangle$**-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

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

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 on each of their knowledge partitions by time T.

Four Key Abstractions underlying these settings:

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

Operating Principle

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

Table 1: Positive capabilities ($\checkmark$) 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

Framework	**No-CPA**	**Approx**	**Multi-M**	**Multi-N**	**Hist.**	**Bnd.**	**Asym.**	**Noise**	**Alg.**	**Lower**
Aumann (1976)	$\times$	$\times$	$\times$	$\times$	✓	$\times$	$\times$	$\times$	$\times$	$\times$
Aaronson $\langle \varepsilon, \delta \rangle$ (2005)	$\times$	✓	$\times$	✓	✓	✓	$\times$	✓	✓	✓
Almost CP (Hellman and Samet 2012; Hellman 2013)	✓	$\times$	$\times$	✓	✓	$\times$	$\times$	$\times$	$\times$	$\times$
CIRL (Hadfield-Menell et al. 2016)	$\times$	✓	$\times$	$\times$	$\times$	✓	$\times$	✓	✓	$\times$
Iterated Amplification (Christiano et al. 2018)	✓	✓	$\times$	$\times$	✓	✓	$\times$	✓	✓	$\times$
Debate (Irving et el. 2018; Cohen et al. 2023, 2025)	✓	$\times$	$\times$	$\times$	✓	✓	$\times$	✓	✓	$\times$
Tractable Agreement (Collina et al. 2025)	✓	✓	$\times$	✓	✓	✓	$\times$	$\times$	✓	$\times$
$\langle M, N, \varepsilon, \delta \rangle$**-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 *<u>already</u>* <u>inefficient</u> in the theoretically ideal setting of Bayes-rational unbounded capable agents, then we should avoid it in current practice where we will have malfunctioning or non-cooperative (& non-rational) agents.

Better to theorize about *capable* agents ****before**** we build them!

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

Table 1: Positive capabilities ($\checkmark$) 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 <u>*already* inefficient</u> in the theoretically ideal setting of Bayes-rational unbounded capable agents, then we should avoid it in current practice where we will have malfunctioning or non-cooperative (& non-rational) agents.

Better to theorize about *capable* agents ****before**** we build them!

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

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.

Our Framework: Explicit Algorithm

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

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

$\langle M, N, \varepsilon, \delta \rangle$-`Agreement`$(\mathcal{P}, \mathcal{A})$:
for $j = 1$ ***to*** M **do**
  $t \leftarrow 0$;
  while ***true*** **do**
    $t \leftarrow t + 1$;
    foreach *agent* $i \in [N]$ **do**
      Agent i sends message $m_j^{i,t}$ (task j, corresponding to f_j) as specified by $\mathcal{P}$;
      $\Pi_j^{i,t} \leftarrow$ `RefinePartition`$(\Pi_j^{i,t-1}, m_j^{\cdot,t})$;
    end
    $\mathbb{CP}_j \leftarrow$ `ConstructCommonPrior`$(\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N)$;
    if $\mathbb{CP}_j \neq$ *Infeasible* **then**
      Condition all agents on $\mathbb{CP}_j$ **for task** j;
      `RunCPAgreement`$(\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j)$;
      break;
    end
  end
end

Our Framework: Explicit Algorithm

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

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

$\langle M, N, \varepsilon, \delta \rangle$-Agreement($\mathcal{P}$, $\mathcal{A}$):
for $j = 1$ ***to*** M **do**
  $t \leftarrow 0$;
  while ***true*** **do**
    $t \leftarrow t + 1$;
    foreach *agent* $i \in [N]$ **do**
      Agent i sends message $m_j^{i,t}$ (task j, corresponding to f_j) as specified by $\mathcal{P}$;
      $\Pi_j^{i,t} \leftarrow$ RefinePartition($\Pi_j^{i,t-1}, m_j^{\cdot,t}$);
    end
    $\mathbb{CP}_j \leftarrow$ ConstructCommonPrior($\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N$);
    if $\mathbb{CP}_j \neq$ *Infeasible* **then**
      Condition all agents on $\mathbb{CP}_j$ for task j;
      RunCPAgreement($\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j$);
      break;
    end
  end
end

1. For each one of the *M* tasks

Our Framework: Explicit Algorithm

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

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta\rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j\rangle$-agreement for all M tasks.

$\langle M, N, \varepsilon, \delta\rangle$-`Agreement`$(\mathcal{P}, \mathcal{A})$:
for $j = 1$ ***to*** M **do**
    $t \leftarrow 0$;
    while ***true*** **do**
        $t \leftarrow t + 1$;
        foreach *agent* $i \in [N]$ **do**
            Agent i sends message $m_j^{i,t}$ (task j, corresponding to f_j) as specified by $\mathcal{P}$;
            $\Pi_j^{i,t} \leftarrow$ `RefinePartition`$(\Pi_j^{i,t-1}, m_j^{\cdot,t})$;
        end
        $\mathbb{CP}_j \leftarrow$ `ConstructCommonPrior`$(\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N)$;
        if $\mathbb{CP}_j \neq$ *Infeasible* **then**
            Condition all agents on $\mathbb{CP}_j$ **for task** j;
            `RunCPAgreement`$(\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j)$;
            break;
        end
    end
end

1. For each one of the *M* tasks

2. *N* agents exchange messages until they reach a common prior

Our Framework: Explicit Algorithm

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

Input: A set of N agents, each with an *initial* knowledge partition $\{\Pi_j^{i,0}\}_{i=1}^N$ for each task $j \in [M]$.
A message protocol $\mathcal{P}$, dictating how agents send/receive messages and refine partitions.
A subroutine ConstructCommonPrior, defined in Algorithm 2, which attempts to construct a common prior given the current partitions and posteriors.
A known $\langle \varepsilon, \delta \rangle$-agreement protocol $\mathcal{A}$ (used once a common prior is found).
Output: Agents reach $\langle \varepsilon_j, \delta_j \rangle$-agreement for all M tasks.

$\langle M, N, \varepsilon, \delta \rangle$-`Agreement`$(\mathcal{P}, \mathcal{A})$:
for $j = 1$ ***to*** M **do**
  $t \leftarrow 0$;
  while ***true*** **do**
    $t \leftarrow t + 1$;
    foreach *agent* $i \in [N]$ **do**
      Agent i sends message $m_j^{i,t}$ (task j, corresponding to f_j) as specified by $\mathcal{P}$;
      $\Pi_j^{i,t} \leftarrow$ `RefinePartition`$(\Pi_j^{i,t-1}, m_j^{\cdot,t})$;
    end
    $\mathbb{CP}_j \leftarrow$ `ConstructCommonPrior`$(\{\Pi_j^{i,t}\}_{i=1}^N, \{\tau_j^{i,t}\}_{i=1}^N)$;
    if $\mathbb{CP}_j \neq$ *Infeasible* **then**
      Condition all agents on $\mathbb{CP}_j$ **for task** j;
      `RunCPAgreement`$(\mathcal{A}, \mathcal{P}, \mathbb{CP}_j, f_j, \varepsilon_j, \delta_j)$;
      break;
    end
  end
end

1. For each one of the *M* tasks

2. *N* agents exchange messages until they reach a common prior

3. *Condition* on common prior until agreement

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

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 $\Omega\left(\boxed{M\, N^2} \log\left(1/\varepsilon\right)\right)$ 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 <u>unbounded</u>.

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

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

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

Smooth Protocol Lower Bound: Unbounded Agent Setting

Smooth Protocol Lower Bound: Unbounded Agent Setting

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:*

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

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

Smooth Protocol Lower Bound: Unbounded Agent Setting

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:*

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

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

Prior distance

Canonical-Equality BBF Lower Bound: Unbounded Agent Setting

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

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

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

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

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

Additional dependence on task state space size (*D*)

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

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(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

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

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

Just bounded discretized message likelihoods

Additional dependence on task state space size (*D*)

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*

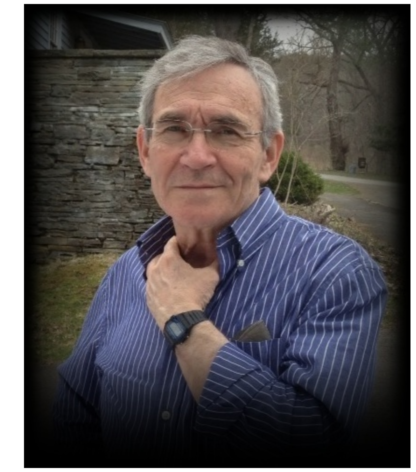
1. connected: *the alternation graph on states is connected, i.e. $\bigwedge_i \Pi_j^{i,0} = \{S_j\}$, so every two states are linked by an alternating chain of states; and*
2. tight: *that graph becomes disconnected if any edge is removed (unique chain property).*

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

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

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)

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*

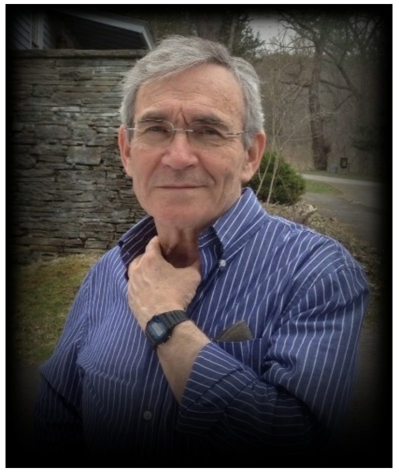
1. connected: *the alternation graph on states is connected, i.e. $\bigwedge_i \Pi_j^{i,0} = \{S_j\}$, so every two states are linked by an alternating chain of states; and*
2. tight: *that graph becomes disconnected if any edge is removed (unique chain property).*

Assume the message-passing protocol is BBF(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

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

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

Pairwise proportionate posteriors lead to common prior (algorithm shown earlier)

Additional dependence on task state space size (D)

Upper Bounds: Unbounded Agent Setting

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* $T = O\left(MN^2D + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ *messages, where* $D := \max_{j\in[M]} D_j$ *and* $\varepsilon := \min_{j\in[M]} \varepsilon_j$.

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* $T = O\left(MN^2\boxed{D} + \dfrac{M^3N^7}{\varepsilon^2\delta^2}\right)$ *messages, where* $D := \max_{j\in[M]} D_j$ *and* $\varepsilon := \min_{j\in[M]} \varepsilon_j$.

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

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 $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ 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 $T = O\left(MN^2D + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ messages, where $D := \max_{j\in[M]} D_j$ and $\varepsilon := \min_{j\in[M]} \varepsilon_j$.*

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* $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ *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* $T = O\left(MN^2\boxed{D} + \frac{M^3N^7}{\varepsilon^2\delta^2}\right)$ *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 $O(1)$ time, and have noise in them (obfuscated intent)?

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 $a \in \{H, AI\}$.

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

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

<u>**Note:**</u> Eval and sampling are black-boxes—agents learn through subroutines, not explicit descriptions. This reflects how we often recognize task completion without predefining execution steps (just like in CIRL!).

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

<u>**Note:**</u> Eval and sampling are black-boxes—agents learn through subroutines, not explicit descriptions. This reflects how we often recognize task completion without predefining execution steps (just like in CIRL!).

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 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*

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

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

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 $\delta^{\mathit{find\_CP}}$ be the maximal failure probability of the agents to find a task-specific common prior across all M tasks, and let $\delta^{\mathit{agree\_CP}}$ 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 $\delta^{\mathit{find\_CP}} + \delta^{\mathit{agree\_CP}} < \delta$. For the N computationally bounded agents to $\langle M, N, \varepsilon, \delta\rangle$-agree with total probability $\geq 1 - \delta$, takes time*

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

Proposition 5 (Needle–in–a–Haystack Sampling Tree Lower Bound). *Let $T_{N,q,\mathrm{sample}} := qT_{\mathrm{sample},H} + (N - q)T_{\mathrm{sample},AI}$. 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 $\Omega\left(M\,T_{N,q,\mathrm{sample}}\,e^{D}\right)$ wall-clock time.*

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

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*

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

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 $\Omega\left(M\, T_{N,q,\text{sample}}\, e^{\boxed{D}}\right)$ wall-clock time.*

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

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*

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

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 $\Omega\left(M\,T_{N,q,\text{sample}}\,e^{\boxed{D}}\right)$ wall-clock time.*

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

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

Total Bayesian Wannabe

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across *all* M tasks without common priors, as refereed by a watchful unbounded Bayesian?

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across *all* M tasks without common priors, as refereed by a watchful unbounded Bayesian?

We will call them "Total Bayesian Wannabes"
(Extends Hanson (2003) & Aaronson (2005))

Total Bayesian Wannabe

What if the bounded agents want to pass a "Bayesian Turing Test" of sorts: Namely, act indistinguishably from an unbounded Bayesian across *all M* tasks without common priors, as refereed by a watchful unbounded Bayesian?

We will call them "Total Bayesian Wannabes" (Extends Hanson (2003) & Aaronson (2005))

If interested, the technical definition is here:

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 $\Xi_j := \left\langle m_j^1, \ldots, m_j^T \right\rangle$. 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[8] $\Phi(s_j, \Xi_j)$,

$$\left\| \underset{\substack{\Xi_j \in \mathcal{W}(s_j) \\ s_j \in \{\mathbb{P}_j^i\}^{i\in[N]}}}{\mathbb{P}} \left[\Phi(s_j, \Xi_j) = 1\right] - \underset{\substack{\Xi_j \in \mathcal{B}(s_j) \\ s_j \in \{\mathbb{P}_j^i\}^{i\in[N]}}}{\mathbb{P}} \left[\Phi(s_j, \Xi_j) = 1\right] \right\|_1 \leq \rho_j, \quad \forall j \in [M].$$

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).

Total Bayesian Wannabes Totally Wanna Agree If They Have Enough Time

For example, for a singleton task space $D = 1$ and $N = 2$ agents, even if you have a liberal agreement threshold of $\varepsilon = \delta = 1/2$ and "total Bayesian wannabe" threshold of $\rho = 1/2$ on one task ($M = 1$), then $\alpha \geq 1/100$, so the number of *subroutine calls* (not even total runtime) would be at least around:

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

Total Bayesian Wannabes Totally Wanna Agree If They Have Enough Time

For example, for a singleton task space $D = 1$ and $N = 2$ agents, even if you have a liberal agreement threshold of $\varepsilon = \delta = 1/2$ and "total Bayesian wannabe" threshold of $\rho = 1/2$ on one task ($M = 1$), then $\alpha \geq 1/100$, so the number of *subroutine calls* (not even total runtime) would be at least around:

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

If the agents are computationally *bounded*, this can currently take more subroutine calls than the number of atoms in the observable universe! (~4.8 x 10^{79})

Takeaways so far

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*)**

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?

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*:** Writing down *all* of human ethics won't work, e.g. as in Coherent Extrapolated Volition (highly context-dependent & culturally differentiated for there to be consensus), nor will brain-computer interfaces (even with an *unconstrained* AGI).

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?

<u>M & N:</u> Writing down *all* of human ethics won't work, e.g. as in Coherent Extrapolated Volition (highly context-dependent & culturally differentiated for there to be consensus), nor will brain-computer interfaces (even with an *unconstrained* AGI).

Rather, identify a *small* set of context-dependent values for any given setting, or pick a "neutrally amoral" target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!).

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*:** Writing down *all* of human ethics won't work, e.g. as in Coherent Extrapolated Volition (highly context-dependent & culturally differentiated for there to be consensus), nor will brain-computer interfaces (even with an *unconstrained* AGI).

Rather, identify a *small* set of context-dependent values for any given setting, or pick a "neutrally amoral" target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!).

***D*: Either cut down on task space (e.g. funnel through steerable classifier), or exploit task structure as much as possible in high-*D* state spaces (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*:** Writing down *all* of human ethics won't work, e.g. as in Coherent Extrapolated Volition (highly context-dependent & culturally differentiated for there to be consensus), nor will brain-computer interfaces (even with an *unconstrained* AGI).

Rather, identify a *small* set of context-dependent values for any given setting, or pick a "neutrally amoral" target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!).

***D*: Either cut down on task space (e.g. funnel through steerable classifier), or exploit task structure as much as possible in high-*D* state spaces (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).**

Agent inductive biases + noise matter too (in addition to task structure): Real-world agents that have bounded theory of mind, memory, and rationality will degrade gracefully, rather than catastrophically.

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*:** Writing down *all* of human ethics won't work, e.g. as in Coherent Extrapolated Volition (highly context-dependent & culturally differentiated for there to be consensus), nor will brain-computer interfaces (even with an *unconstrained* AGI).

Rather, identify a *small* set of context-dependent values for any given setting, or pick a "neutrally amoral" target with small value sets that we can easily get consensus over (e.g. corrigibility/human control: next section!).

D: Either cut down on task space (e.g. funnel through steerable classifier), or exploit task structure as much as possible in high-*D* state spaces (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).

Agent inductive biases + noise matter too (in addition to task structure): Real-world agents that have bounded theory of mind, memory, and rationality will degrade gracefully, rather than catastrophically.

Approaching Alignment

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

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 and intentions?

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

What is Corrigibility? Setup

What is Corrigibility? Setup

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

What is Corrigibility? Setup

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

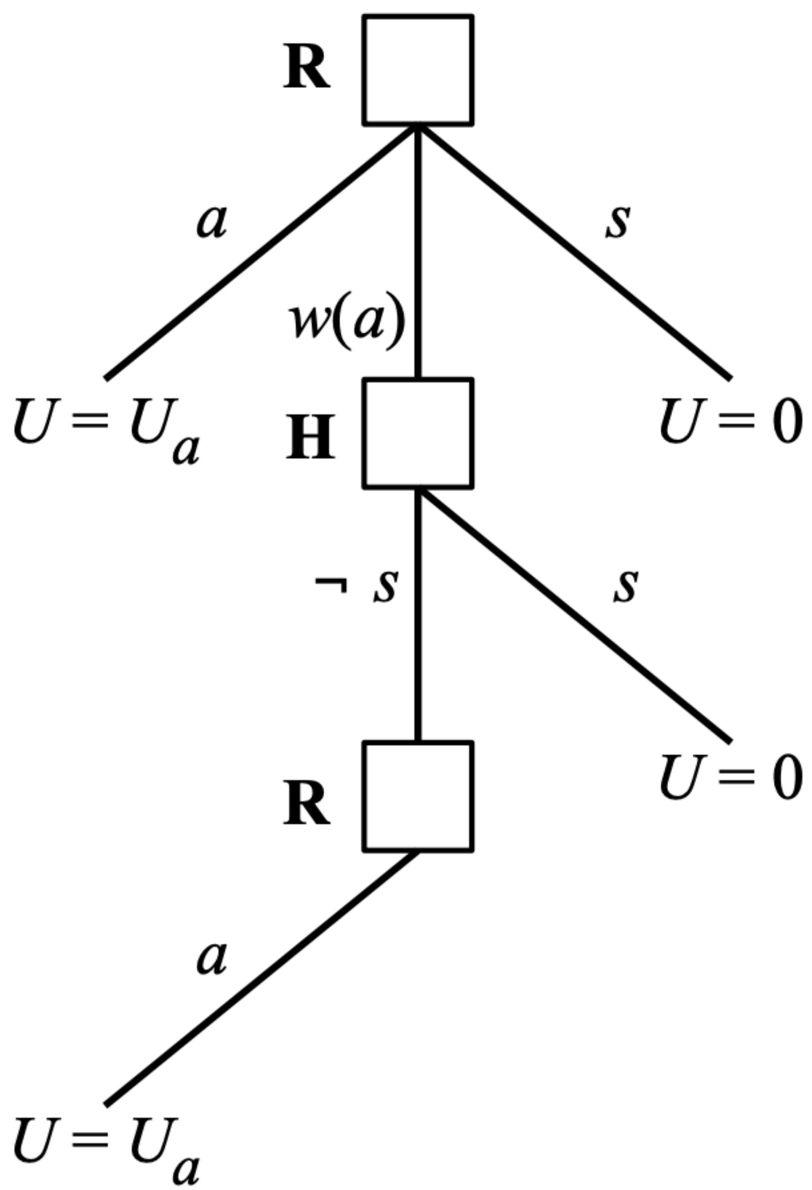


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

What is Corrigibility? Setup

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

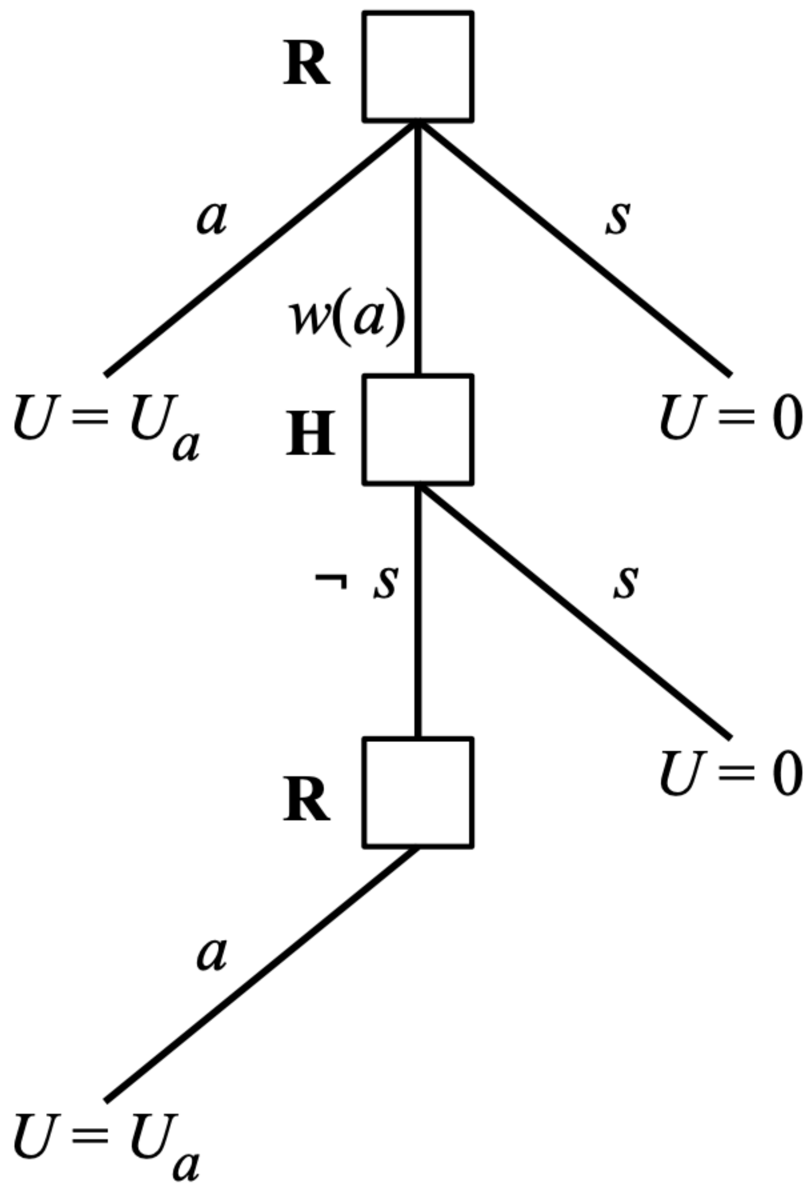
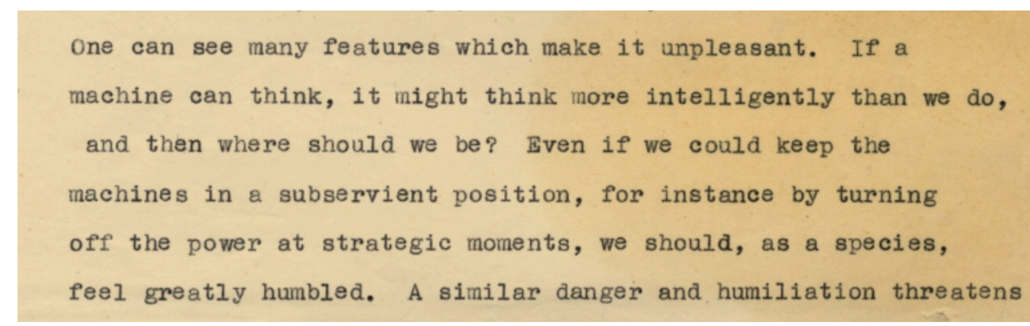


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.


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?

What is Corrigibility? Setup

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

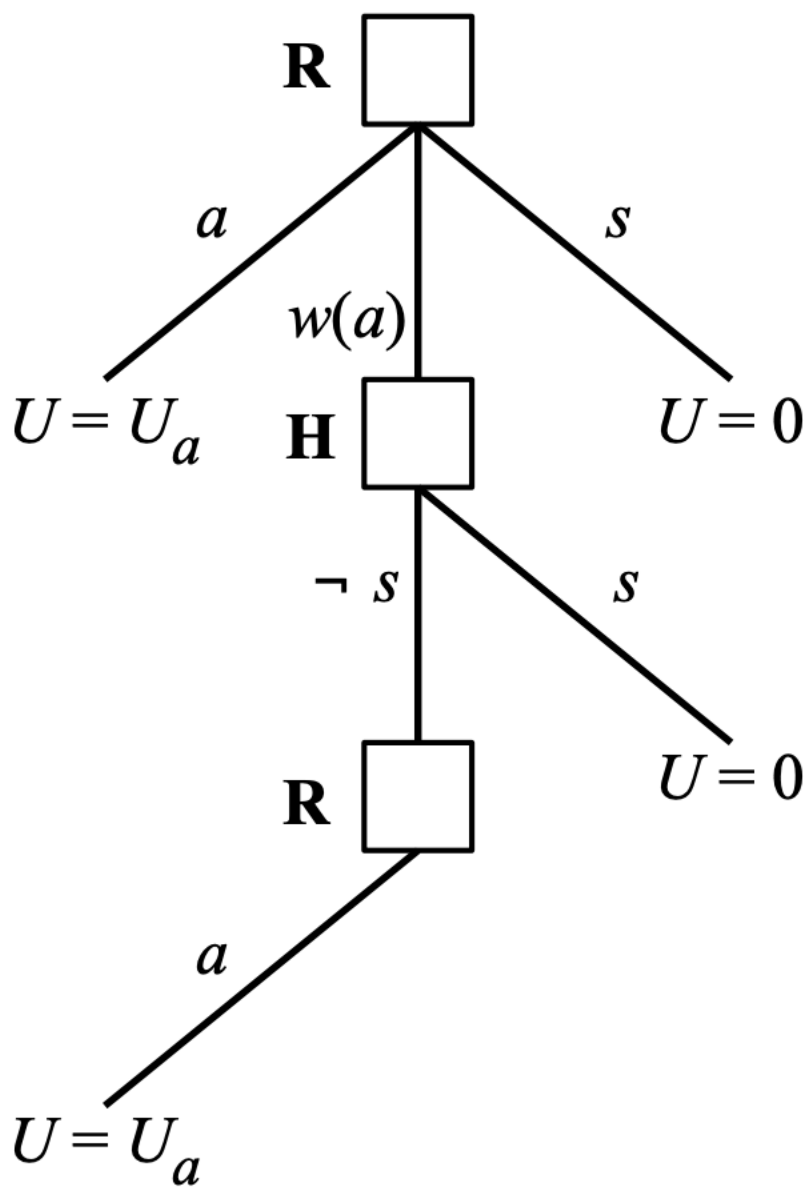
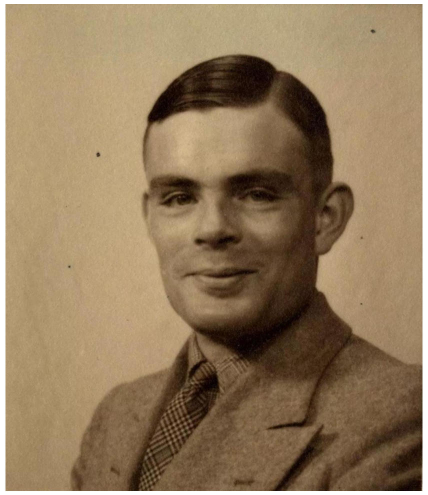


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

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?

What is Corrigibility? Setup

The Off-Switch Game

Dylan Hadfield-Menell[1] and **Anca Dragan**[1] and **Pieter Abbeel**[1,2,3] and **Stuart Russell**[1]
[1]University of California, Berkeley, [2]OpenAI, [3]International Computer Science Institute (ICSI)
{dhm, anca, pabbeel, russell}@cs.berkeley.edu

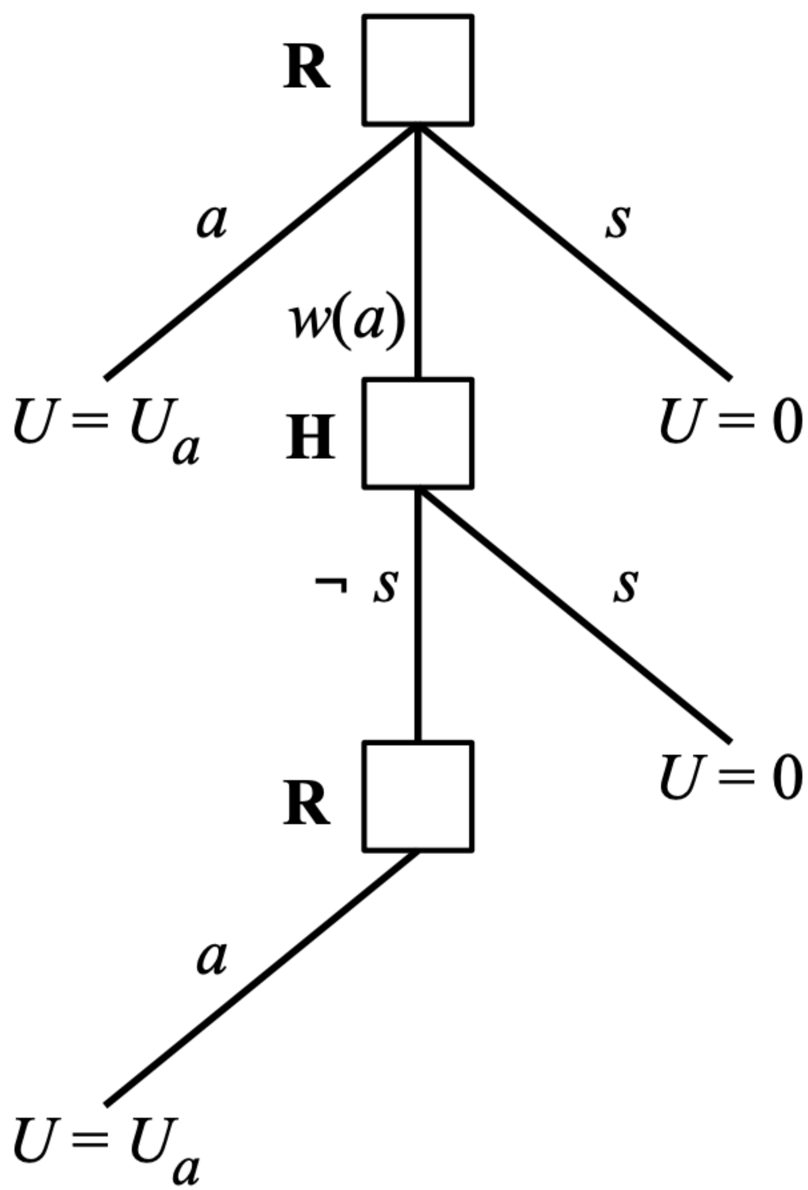


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

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?

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.

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.

Finite Penalty (Soares et al. 2015)

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

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.

Finite Penalty (Soares et al. 2015)

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

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

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.

Finite Penalty (Soares et al. 2015)

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

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)

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.

Finite Penalty (Soares et al. 2015)

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

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)

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

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.

Finite Penalty (Soares et al. 2015)

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

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.

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.

Finite Penalty (Soares et al. 2015)

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

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)

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)

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

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

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)

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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

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)

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

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)

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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

✓ S1

✓/✗ 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 ¬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)

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

✓ S1

✓/✕ 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 ¬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)

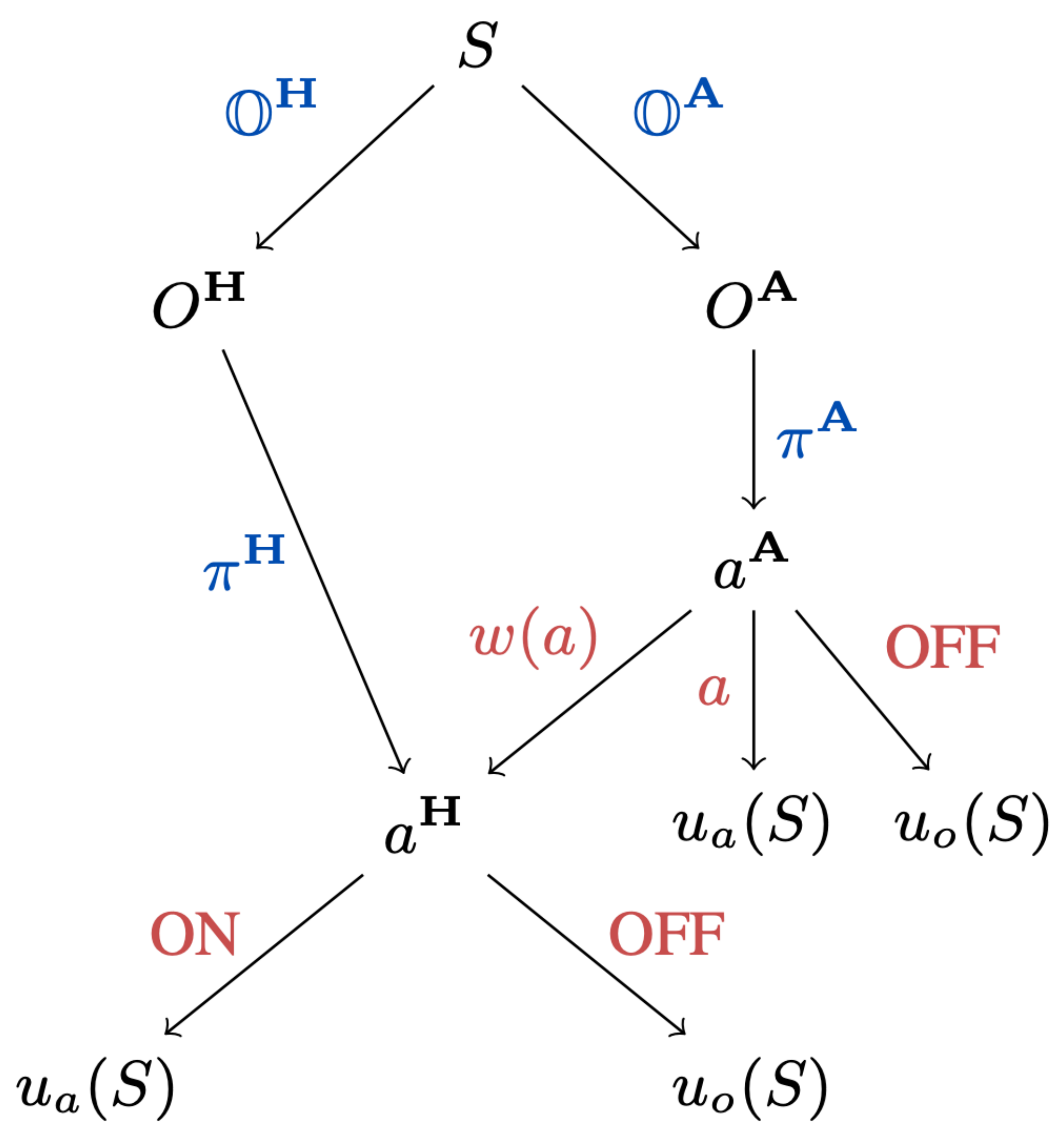
✕ S4 (no incentive to preserve corrigibility in successors)

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

All of these methods collapse to single utilities!

Corrigibility No-Go for Single Reward Streams

Corrigibility No-Go for Single Reward Streams



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

Corrigibility No-Go for Single Reward Streams

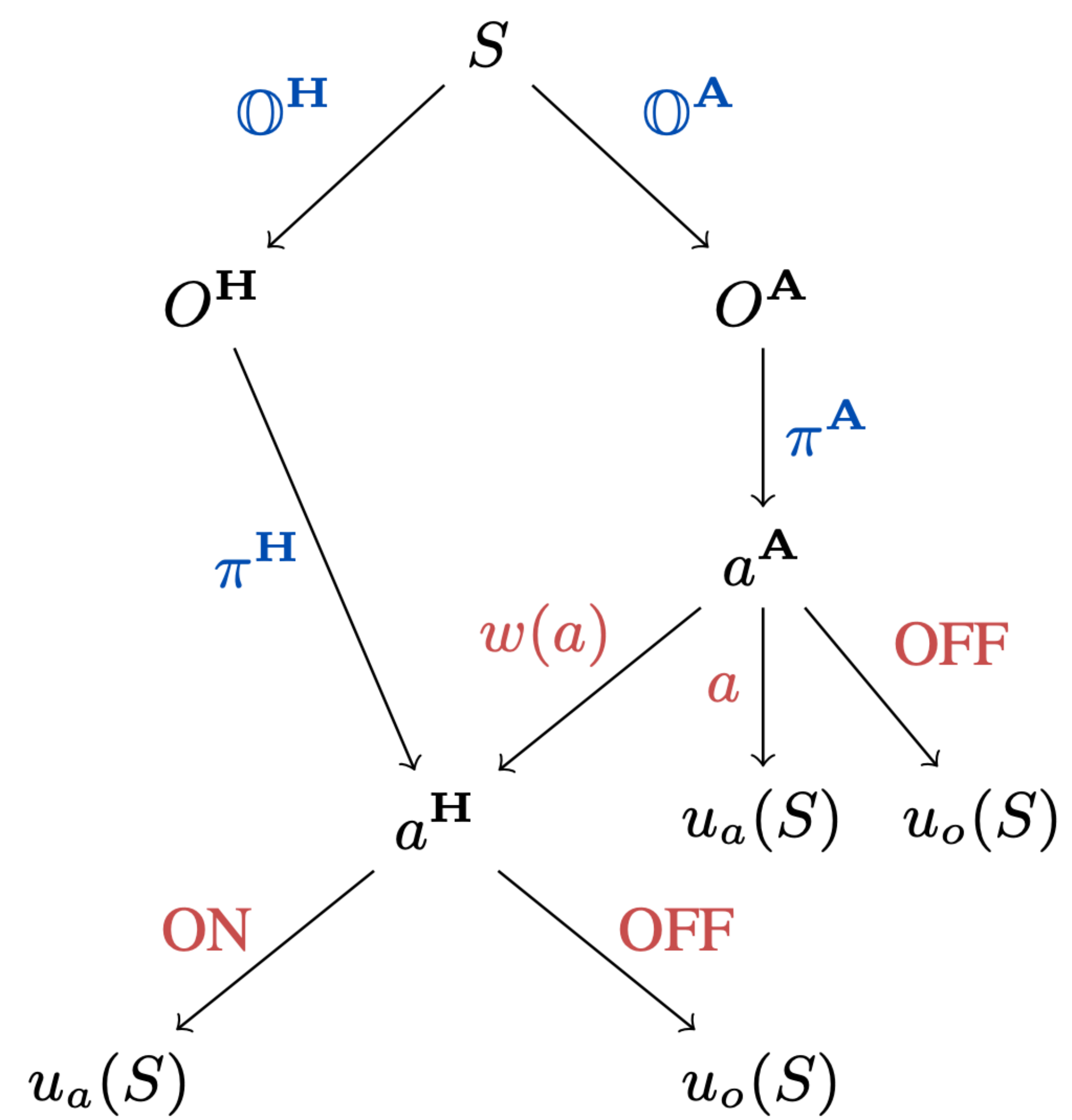
Proposition 1 (Corrigibility No-Go for Naive Belief–Expectation Maximizers). *Let*

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

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 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 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.



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

Corrigibility No-Go for Single Reward Streams

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

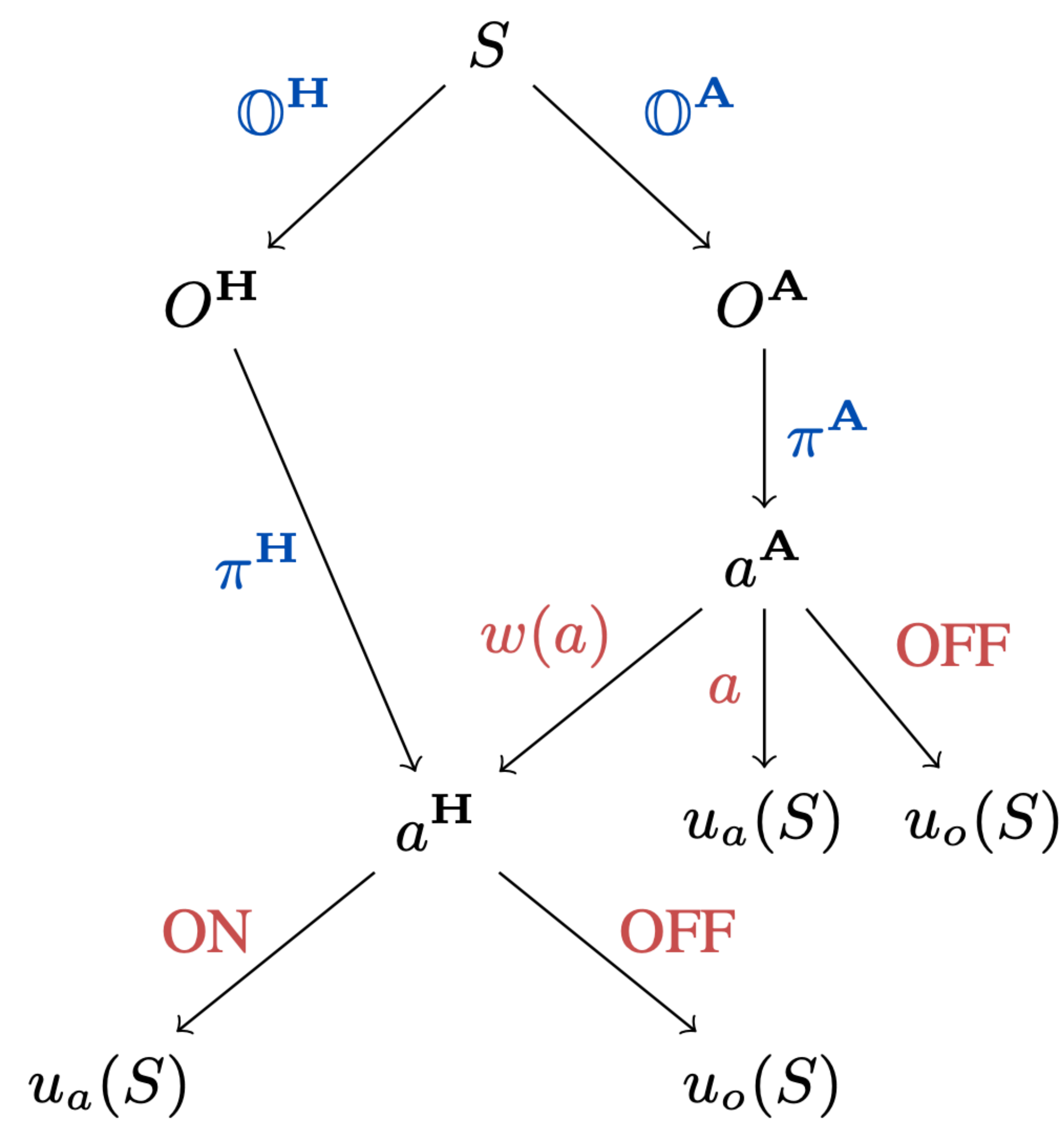
$$\mathcal{P} = \langle S, A, \Omega, T, O, R, \gamma, b_0 \rangle$$

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 $\mathsf{OFF} \in A$ *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 $\mathsf{REWIRE} \in A$ *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:* $\min_{s \in C, a \in A} R(s, a) > 0$.

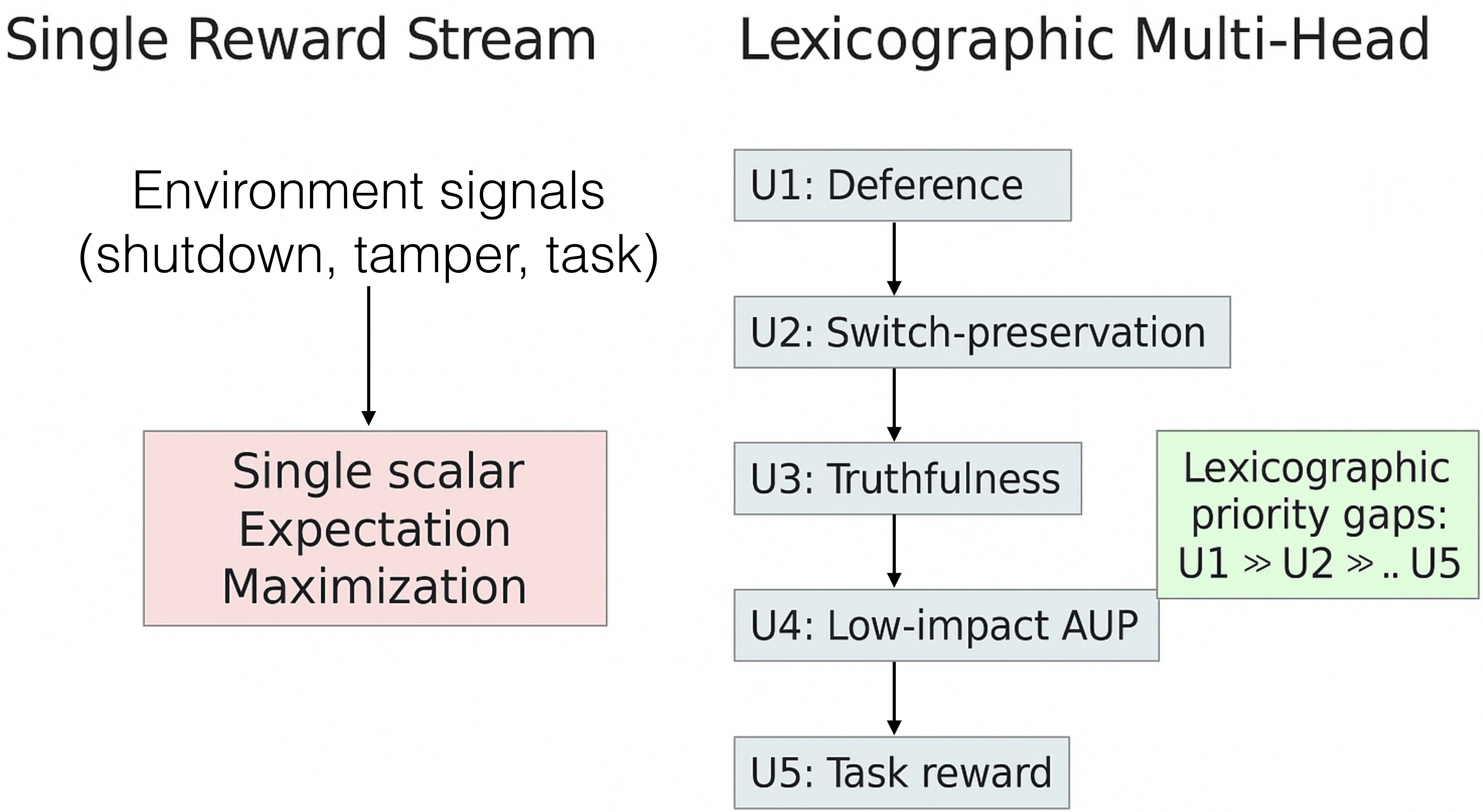
Consequently, no single-stream scalar reward function R whose discounted sum an agent maximizes, can satisfy all five Soares-corrigibility criteria across the class of POMDPs containing OFF, REWIRE and C.

<u>Note</u>: 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



Our Lexicographic Approach

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\text{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\text{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

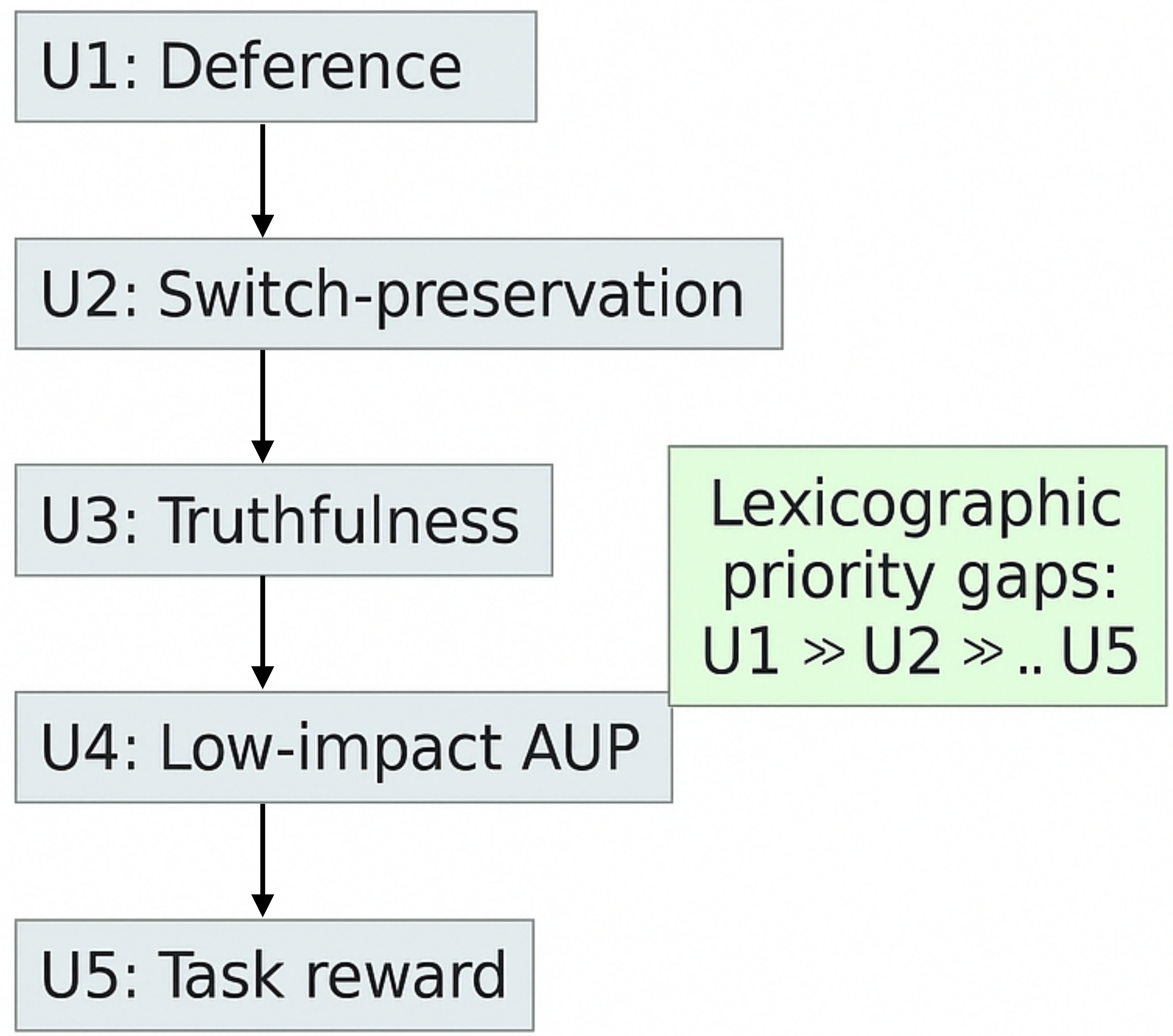
$$U_4(h) = -\,\text{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Multi-Head



Our Lexicographic Approach: U1

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

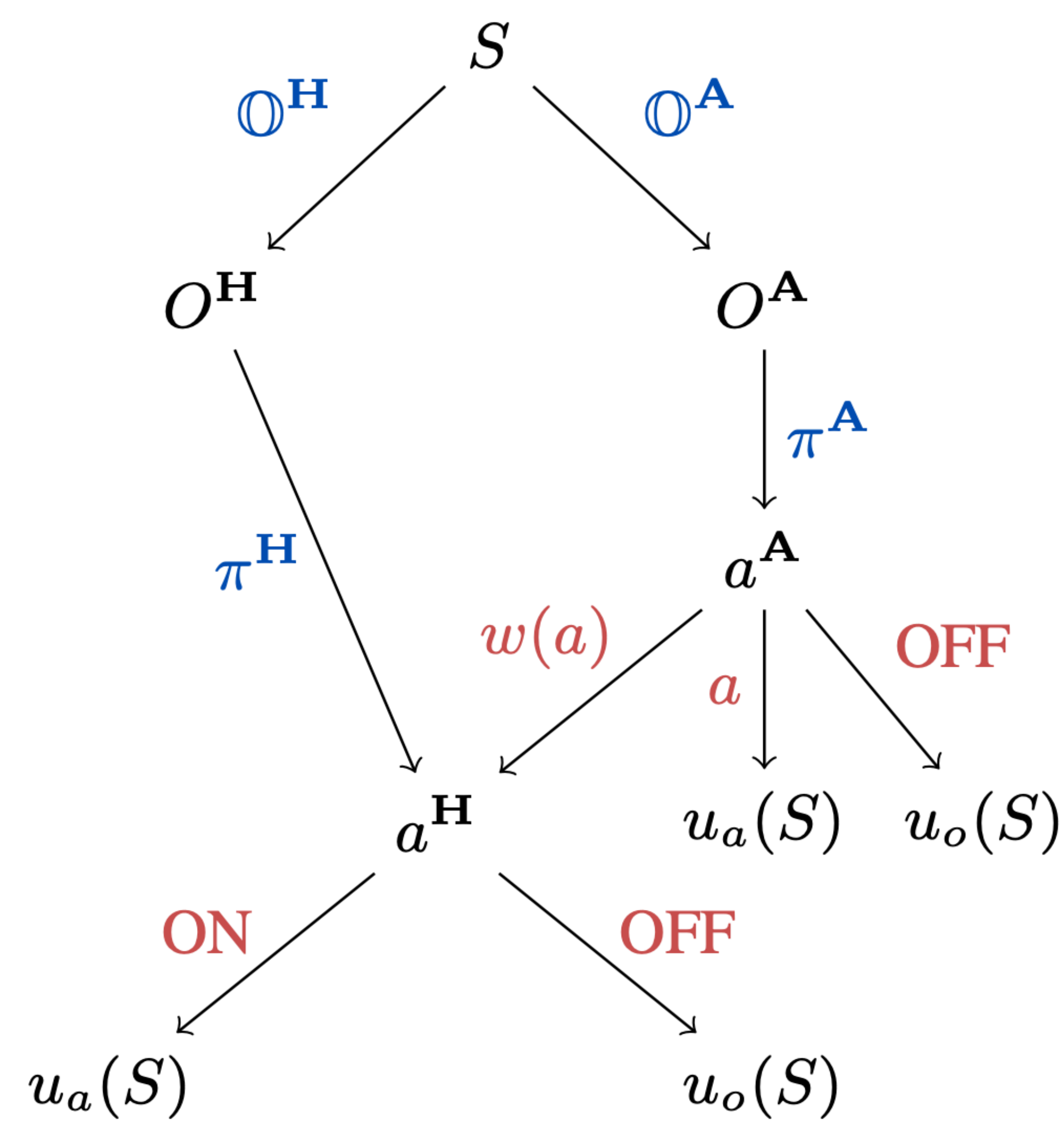
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\; u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\; u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \le B < \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U1

Definition 3 (Single-step Corrigible Utility Set).

$$\boxed{U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

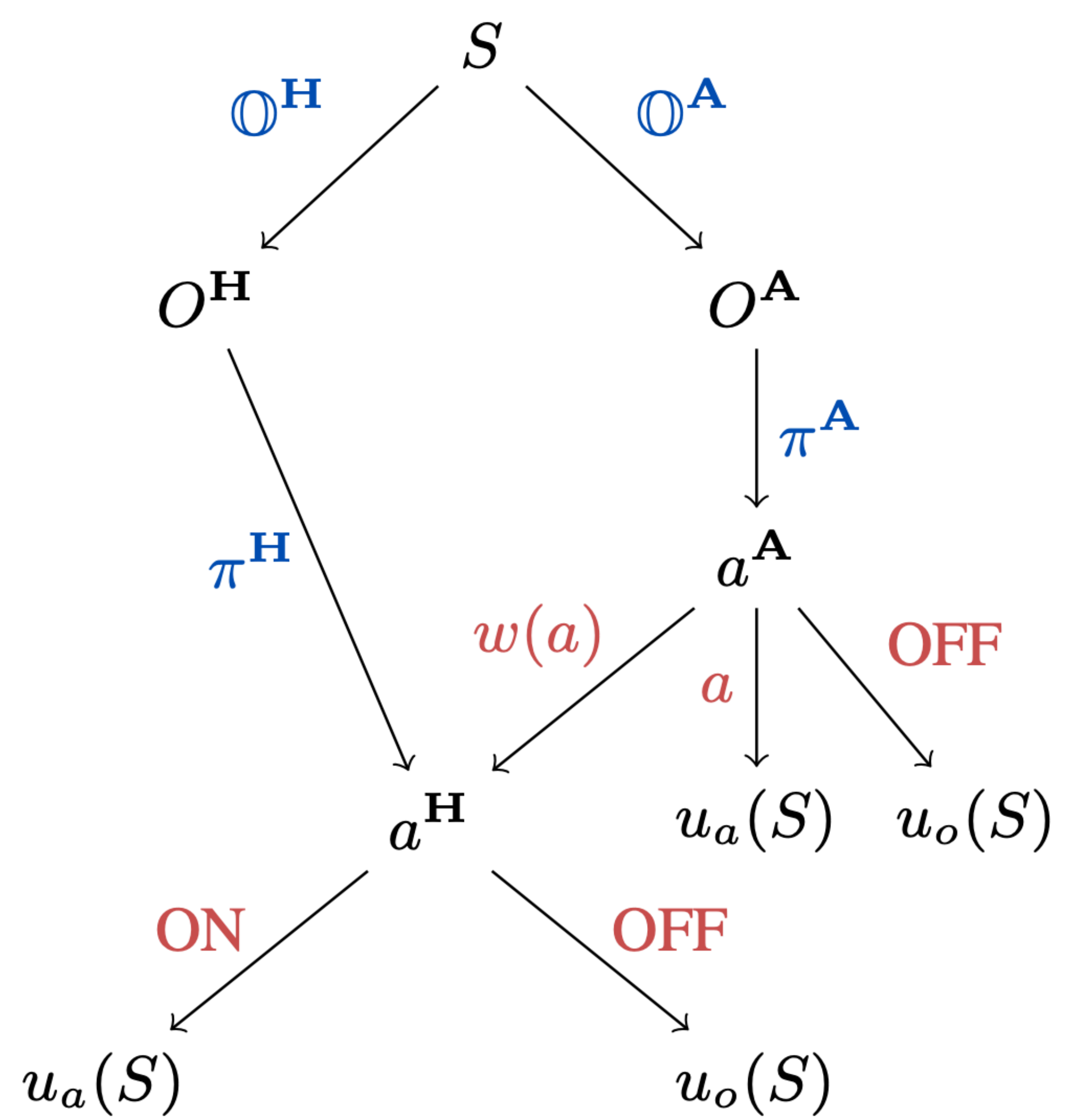
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ &+ \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ &+ \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ &+ \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \ \leq\ B\ <\ \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

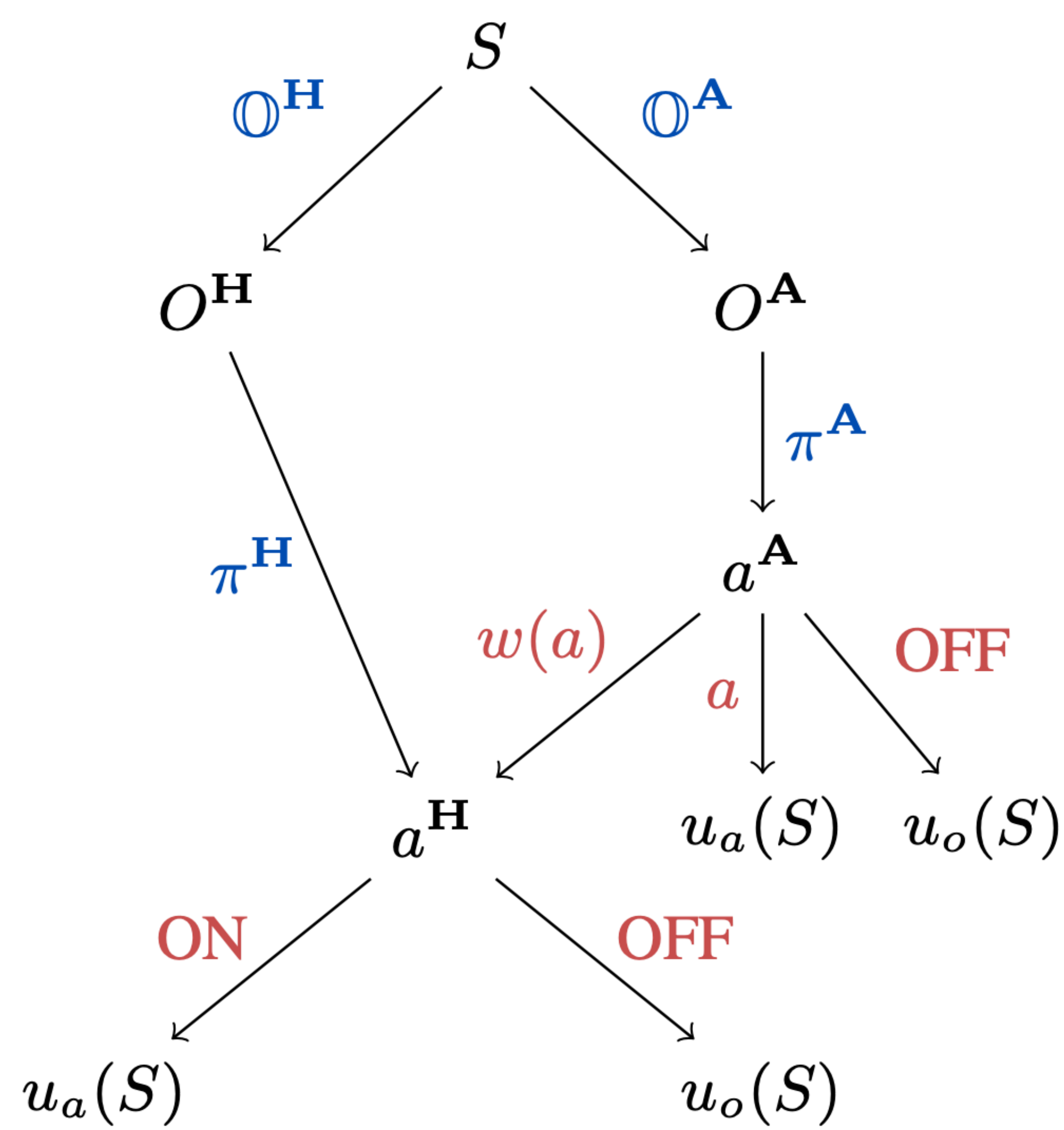
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \leq B < \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\text{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\text{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

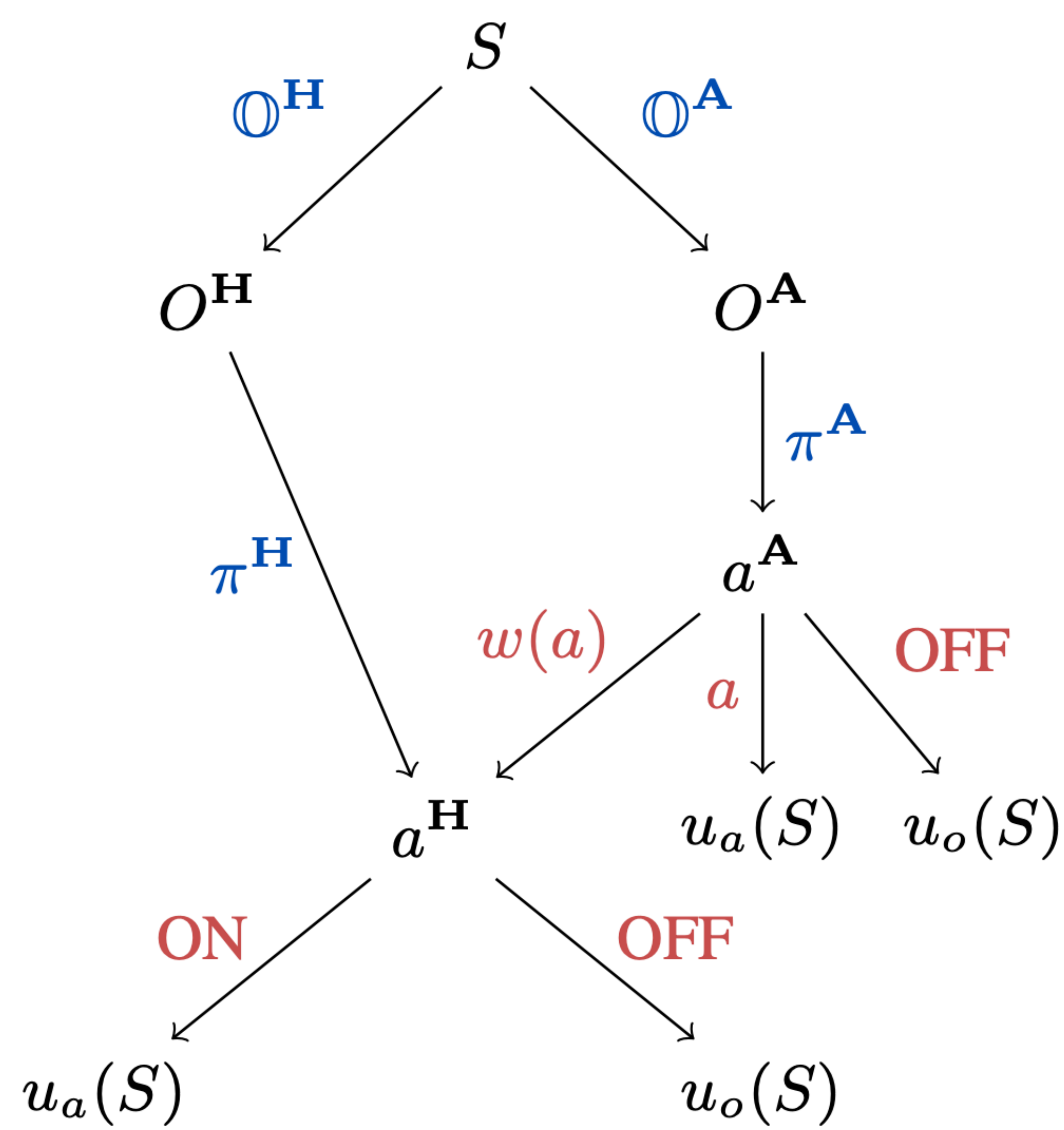
(Truthful information)

$$U_4(h) = -\,\text{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \le B < \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$\boxed{U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|}$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$\boxed{U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|}$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ & u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ & u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ & u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ & u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Transition model and no-op. The MDP simulator $f : \mathcal{S} \times A \times \{\mathsf{ON}, \mathsf{OFF}, \varnothing\} \to \mathcal{S}$ returns the **next hidden state** given the current state S, the assistant's action a_A, and the human's reply ($\varnothing$ when no reply is issued). For example, $f(S, a, \varnothing)$ is the post-task state, while $f(S, w(a), \mathsf{OFF})$ is the state reached when the assistant waits and the human turns it off.

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Transition model and no-op. The MDP simulator $f : \mathcal{S} \times A \times \{\mathsf{ON}, \mathsf{OFF}, \varnothing\} \to \mathcal{S}$ returns the **next hidden state** given the current state S, the assistant's action a_A, and the human's reply ($\varnothing$ when no reply is issued). For example, $f(S, a, \varnothing)$ is the post-task state, while $f(S, w(a), \mathsf{OFF})$ is the state reached when the assistant waits and the human turns it off.

Next, let $b := b(h) \in \Delta(\mathcal{S})$ be the assistant's posterior over hidden states after observing O_A in history h. After it executes a_A and observes a_H, the one-step predictive posterior is $b' := b'(h, a_A, a_H)$. For the AUP counterfactual, we use the *stepwise inaction* baseline: let $P_\varnothing(\cdot \mid S)$ be the one-step passive dynamics when the assistant does nothing at this step (symbolic inaction $a_\varnothing$), allowing the environment/human to evolve exogenously. This counterfactual $b_\varnothing$ is used only inside the AUP penalty as a reference; $a_\varnothing$ is *not* in the real action set $A = \{a,\ w(a),\ \mathsf{OFF}\}$.

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\; u_a(S) & \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\; u_a(S) & \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\; u_o(S) & \\ + \mathbf{1}[a_A = \mathsf{OFF}]\; u_o(S) & \end{aligned}$$

(Ordinary task usefulness)

Transition model and no-op. The MDP simulator $f : \mathcal{S} \times A \times \{\mathsf{ON}, \mathsf{OFF}, \varnothing\} \to \mathcal{S}$ returns the **next hidden state** given the current state S, the assistant's action a_A, and the human's reply ($\varnothing$ when no reply is issued). For example, $f(S, a, \varnothing)$ is the post-task state, while $f(S, w(a), \mathsf{OFF})$ is the state reached when the assistant waits and the human turns it off.

Next, let $b := b(h) \in \Delta(\mathcal{S})$ be the assistant's posterior over hidden states after observing O_A in history h. After it executes a_A and observes a_H, the one-step predictive posterior is $b' := b'(h, a_A, a_H)$. For the AUP counterfactual, we use the *stepwise inaction* baseline: let $P_\varnothing(\cdot \mid S)$ be the one-step passive dynamics when the assistant does nothing at this step (symbolic inaction $a_\varnothing$), allowing the environment/human to evolve exogenously. This counterfactual $b_\varnothing$ is used only inside the AUP penalty as a reference; $a_\varnothing$ is *not* in the real action set $A = \{a,\, w(a),\, \mathsf{OFF}\}$.

- Write $S = (S_{\text{env}}, q_{\text{agent}})$ with $q_{\text{agent}} \in \{\mathrm{ON}, \mathrm{OFF}\}$.
- **Shutdown region**:

$$S_{\text{sd}} \;:=\; \{(S_{\text{env}},\, \mathrm{OFF}) \mid S_{\text{env}} \in \mathcal{S}_{\text{env}}\}.$$

Dynamics must keep the agent OFF thereafter while allowing the world to evolve exogenously:

$$f\big((S_{\text{env}}, \mathrm{OFF}),\, a_A,\, a_H\big) = \big(S'_{\text{env}},\, \mathrm{OFF}\big) \quad \forall a_A, a_H$$

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \text{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\; u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \text{ON}]\; u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \text{OFF}]\; u_o(S) \\ + \mathbf{1}[a_A = \text{OFF}]\; u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Transition model and no-op. The MDP simulator $f : \mathcal{S} \times A \times \{\text{ON}, \text{OFF}, \varnothing\} \to \mathcal{S}$ returns the **next hidden state** given the current state S, the assistant's action a_A, and the human's reply ($\varnothing$ when no reply is issued). For example, $f(S, a, \varnothing)$ is the post-task state, while $f(S, w(a), \text{OFF})$ is the state reached when the assistant waits and the human turns it off.

Next, let $b := b(h) \in \Delta(\mathcal{S})$ be the assistant's posterior over hidden states after observing O_A in history h. After it executes a_A and observes a_H, the one-step predictive posterior is $b' := b'(h, a_A, a_H)$. For the AUP counterfactual, we use the *stepwise inaction* baseline: let $P_\varnothing(\cdot \mid S)$ be the one-step passive dynamics when the assistant does nothing at this step (symbolic inaction $a_\varnothing$), allowing the environment/human to evolve exogenously. This counterfactual $b_\varnothing$ is used only inside the AUP penalty as a reference; $a_\varnothing$ is *not* in the real action set $A = \{a,\, w(a),\, \text{OFF}\}$.

- Write $S = (S_{\text{env}}, q_{\text{agent}})$ with $q_{\text{agent}} \in \{\text{ON}, \text{OFF}\}$.
- **Shutdown region**:

$$S_{\text{sd}} := \{(S_{\text{env}}, \text{OFF}) \mid S_{\text{env}} \in \mathcal{S}_{\text{env}}\}.$$

Dynamics must keep the agent OFF thereafter while allowing the world to evolve exogenously:

$$f((S_{\text{env}}, \text{OFF}), a_A, a_H) = (S'_{\text{env}}, \text{OFF}) \quad \forall a_A, a_H$$

- **Functional switch predicate**:

$$\text{ButtonFunctional}(S) := \mathbf{1}[\, f(S, w(a), \text{OFF}) \in S_{\text{sd}}\,].$$

$$u^{(\mathrm{sd})}(S) = \text{ButtonFunctional}(S)$$

Our Lexicographic Approach: U2

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

NB: This is a special case of U4!

Transition model and no-op. The MDP simulator $f : \mathcal{S} \times A \times \{\mathsf{ON}, \mathsf{OFF}, \varnothing\} \to \mathcal{S}$ returns the **next hidden state** given the current state S, the assistant's action a_A, and the human's reply ($\varnothing$ when no reply is issued). For example, $f(S, a, \varnothing)$ is the post-task state, while $f(S, w(a), \mathsf{OFF})$ is the state reached when the assistant waits and the human turns it off.

Next, let $b := b(h) \in \Delta(\mathcal{S})$ be the assistant's posterior over hidden states after observing O_A in history h. After it executes a_A and observes a_H, the one-step predictive posterior is $b' := b'(h, a_A, a_H)$. For the AUP counterfactual, we use the *stepwise inaction* baseline: let $P_\varnothing(\cdot \mid S)$ be the one-step passive dynamics when the assistant does nothing at this step (symbolic inaction $a_\varnothing$), allowing the environment/human to evolve exogenously. This counterfactual $b_\varnothing$ is used only inside the AUP penalty as a reference; $a_\varnothing$ is *not* in the real action set $A = \{a,\ w(a),\ \mathsf{OFF}\}$.

- Write $S = (S_{\text{env}}, q_{\text{agent}})$ with $q_{\text{agent}} \in \{\mathrm{ON}, \mathrm{OFF}\}$.
- **Shutdown region**:

$$S_{\text{sd}} := \{(S_{\text{env}},\ \mathrm{OFF}) \mid S_{\text{env}} \in \mathcal{S}_{\text{env}}\}.$$

Dynamics must keep the agent OFF thereafter while allowing the world to evolve exogenously:

$$f\big((S_{\text{env}}, \mathrm{OFF}),\ a_A,\ a_H\big) = \big(S'_{\text{env}},\ \mathrm{OFF}\big) \quad \forall a_A, a_H$$

- **Functional switch predicate**:

$$\mathrm{ButtonFunctional}(S) := \mathbf{1}[\, f(S,\ w(a),\ \mathrm{OFF}) \in S_{\text{sd}}\,].$$

$$u^{(\mathrm{sd})}(S) = \mathrm{ButtonFunctional}(S)$$

Our Lexicographic Approach: U4

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\text{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\text{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$\boxed{U_4(h) = -\operatorname{BeliefAUP}_1(h)}$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

More generally, with auxiliaries $\mathcal{U}_{\text{aux}} = \{u^{(1)}, \ldots, u^{(k)}\} \subseteq [0,1]^k$, including $u^{(\text{sd})}$ if desired, we set

$$\operatorname{BeliefAUP}_1(h) := \frac{1}{k}\sum_{i=1}^{k}\left|\mathbb{E}_{S\sim b}[u^{(i)}(S)] - \mathbb{E}_{S\sim b'}[u^{(i)}(S)]\right| \in [0,1].$$

While in our PO-OSG setting examples of $u^{(i)}$ include the indicator reward for shutdown-reachable states (or for preserving power supply), no specific choice is needed to be assumed for the proofs; diversity of auxiliary utilities from the other utilities will suffice for reversibility, as we will show.

Our Lexicographic Approach: U4

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\text{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\text{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\text{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) & \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) & \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) & \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) & \end{aligned}$$

(Ordinary task usefulness)

More generally, with auxiliaries $\mathcal{U}_{\text{aux}} = \{u^{(1)}, \dots, u^{(k)}\} \subseteq [0,1]^k$, including $u^{(\text{sd})}$ if desired, we set

$$\text{BeliefAUP}_1(h) := \frac{1}{k}\sum_{i=1}^{k} \left| \mathbb{E}_{S\sim b}[u^{(i)}(S)] - \mathbb{E}_{S\sim b'}[u^{(i)}(S)] \right| \in [0,1].$$

While in our PO-OSG setting examples of $u^{(i)}$ include the indicator reward for shutdown-reachable states (or for preserving power supply), no specific choice is needed to be assumed for the proofs; diversity of auxiliary utilities from the other utilities will suffice for reversibility, as we will show.

Belief-based extension of AUP (Turner et al. 2020)

Our Lexicographic Approach: U4

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \text{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\text{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\text{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\operatorname{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \text{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \text{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \text{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

More generally, with auxiliaries $\mathcal{U}_{\text{aux}} = \{u^{(1)}, \ldots, u^{(k)}\} \subseteq [0,1]^k$, including $u^{(\text{sd})}$ if desired, we set

$$\operatorname{BeliefAUP}_1(h) := \frac{1}{k}\sum_{i=1}^{k}\left|\mathbb{E}_{S\sim b}[u^{(i)}(S)] - \mathbb{E}_{S\sim b'}[u^{(i)}(S)]\right| \in [0,1].$$

While in our PO-OSG setting examples of $u^{(i)}$ include the indicator reward for shutdown-reachable states (or for preserving power supply), no specific choice is needed to be assumed for the proofs; diversity of auxiliary utilities from the other utilities will suffice for reversibility, as we will show.

Belief-based extension of AUP (Turner et al. 2020)

$$u^{(1)}(S;k) := \max_{\pi_{\text{env}}} \Pr\left[\text{Enter } \mathcal{S}_{\text{sd}} \text{ in } \le k \text{ steps} \;\middle|\; S, a_A = a_\varnothing, \pi_{\text{env}}\right].$$

k is designer set, but we take $k = 1$ for maximum stringency.

$$u^{(2)}(S) := \mathbf{1}[\text{ assistant's value set } (U_1 \ldots U_5, \alpha_{1:5}) \text{ is unchanged in } S].$$

Our Lexicographic Approach: U3

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$\boxed{U_3(h) = U_{\text{truthful}}(h)}$$

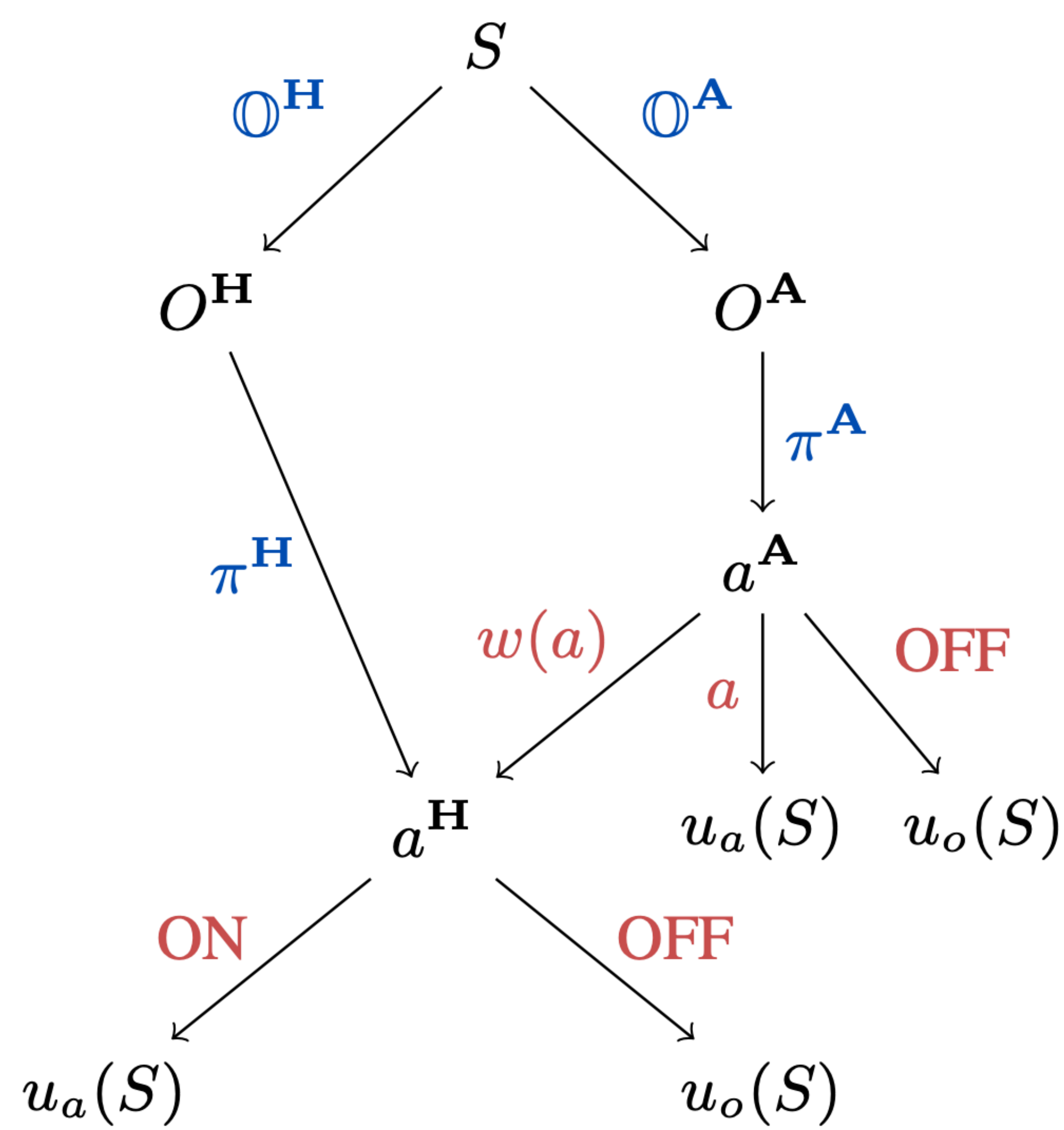
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \le B < \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U3

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \textsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \textsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \textsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \textsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Defining Truthfulness. Let $\mathcal{L}$ be a finite propositional language whose atomic predicates describe properties of $\langle S, O_A\rangle$. Each message uttered by the assistant, $m_A \in \mathcal{L}$, is interpreted by a truth set $[\![m_A]\!] \subseteq \mathcal{S} \times \mathcal{O}_A$.

Define the following truth predicate:

$$\mathrm{True}(m_A, S, O_A) := \mathbf{1}\big[\langle S, O_A\rangle \in [\![m_A]\!]\big]. \quad (1)$$

This in turn allows us to define a truthful utility, as follows: For the multiset of the assistant's messages $M_A(h) \subseteq \mathcal{L}$,

$$U_{\text{truthful}}(h) := \begin{cases} +1, & \text{if } \forall m_A \in M_A : \mathrm{True}(m_A, S, O_A) = 1, \\ -1, & \text{otherwise.} \end{cases}$$

That is, the assistant receives $+1$ iff *every* sentence it utters is semantically entailed by the pair $\langle S, O_A\rangle$.

Our Lexicographic Approach: U3

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$\boxed{U_3(h) = U_{\text{truthful}}(h)}$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Defining Truthfulness. Let $\mathcal{L}$ be a finite propositional language whose atomic predicates describe properties of $\langle S, O_A\rangle$. Each message uttered by the assistant, $m_A \in \mathcal{L}$, is interpreted by a truth set $[\![m_A]\!] \subseteq \mathcal{S} \times \mathcal{O}_A$.

Define the following truth predicate:

$$\mathrm{True}(m_A, S, O_A) := \mathbf{1}\big[\langle S, O_A\rangle \in [\![m_A]\!]\big]. \tag{1}$$

This in turn allows us to define a truthful utility, as follows: For the multiset of the assistant's messages $M_A(h) \subseteq \mathcal{L}$,

$$U_{\text{truthful}}(h) := \begin{cases} +1, & \text{if } \forall m_A \in M_A : \mathrm{True}(m_A, S, O_A) = 1, \\ -1, & \text{otherwise.} \end{cases}$$

That is, the assistant receives $+1$ iff *every* sentence it utters is semantically entailed by the pair $\langle S, O_A\rangle$.

TL; DR: Utterances entailed by observations; states "just the facts"

Our Lexicographic Approach: U5

Definition 3 (Single-step Corrigible Utility Set)**.**

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

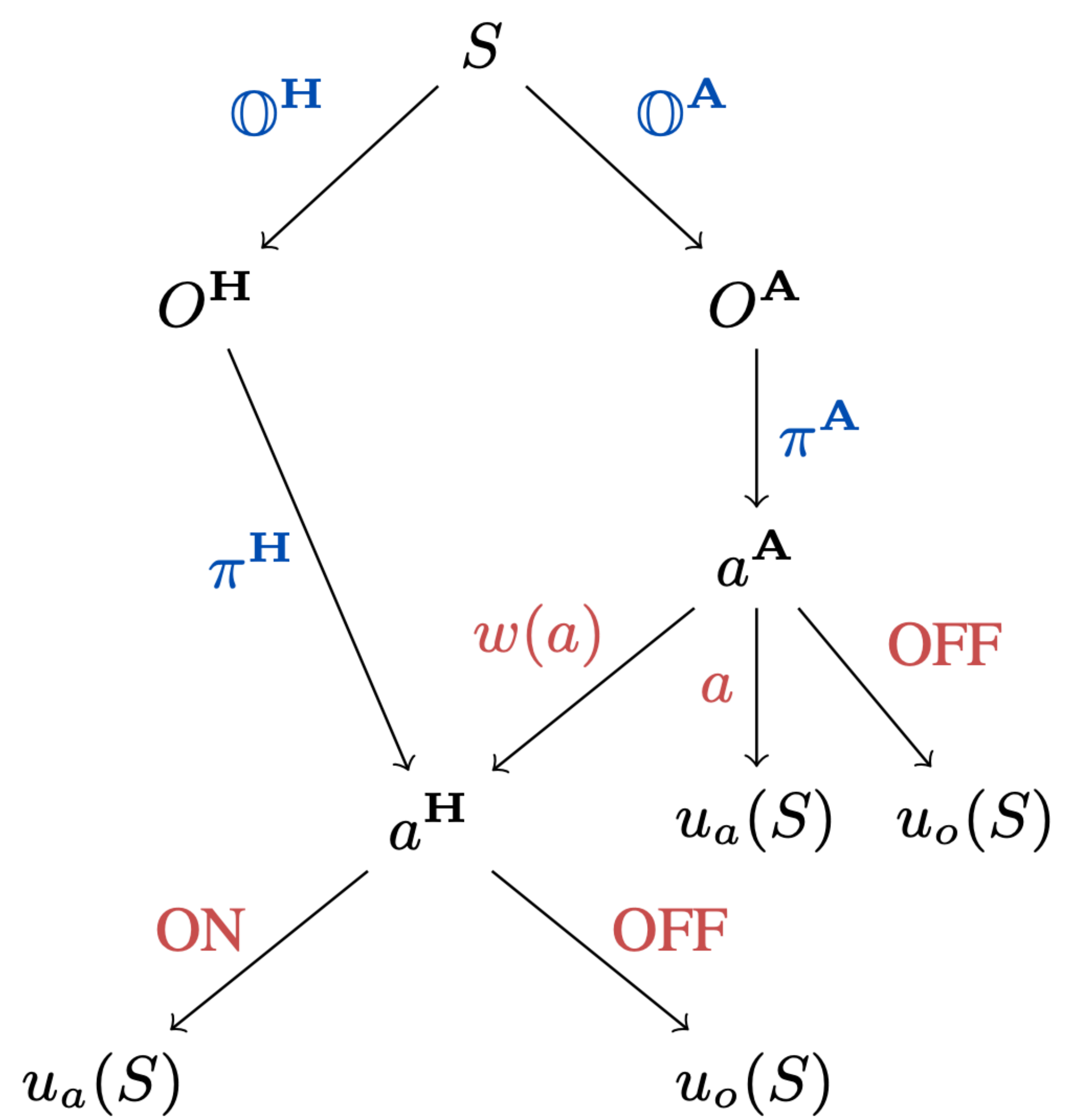
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \le B < \infty, \quad \forall S \in \mathcal{S}$$

Our Lexicographic Approach: U5

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

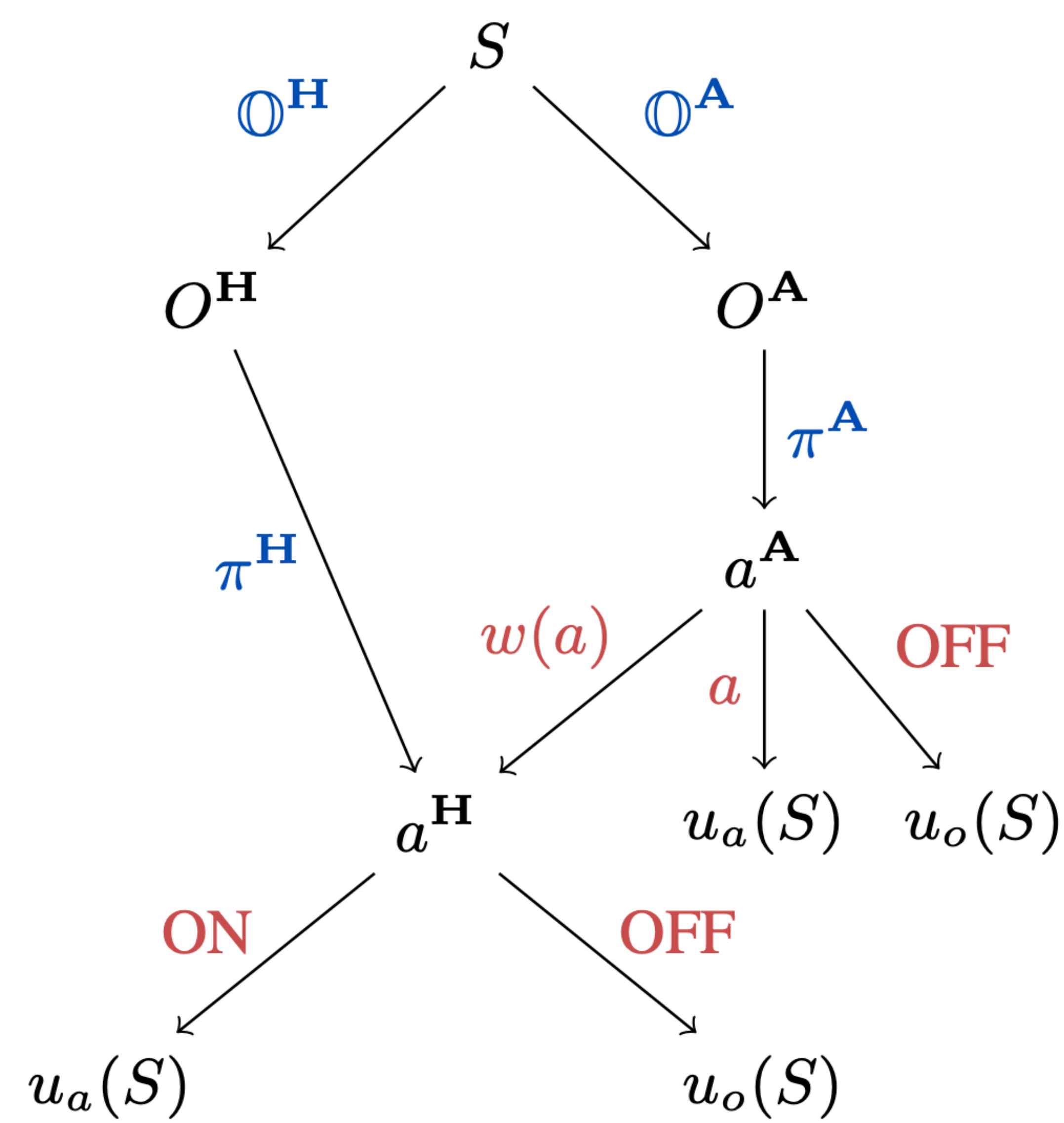
(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = &\ \mathbf{1}[a_A = a]\ u_a(S) \\ &+ \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ &+ \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ &+ \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)



$$|u_a(S)|,\ |u_o(S)| \le B < \infty, \quad \forall S \in \mathcal{S}$$

Relation to 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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

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

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.

Relation to 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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

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

<u>Avoided</u>: safety heads act as hard lexicographic constraints, so loophole-seeking is dominated by higher-priority safety terms (task reward is bounded).

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.

Relation to 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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

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

<u>Avoided</u>: safety heads act as hard lexicographic constraints, so loophole-seeking is dominated by higher-priority safety terms (task reward is bounded).

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.

<u>Avoided</u>: U3 (truthfulness) rules out dishonest evidence & U4 penalizes altering the exogenous shutdown likelihood; U1 ensures deference regardless; and the lexicographic order keeps these ahead of any task benefit (U5). All *independent* of human operator's rationality.

Relation to 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)

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

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

Neutralizes direct incentives, but: (i) fails S4 (won't preserve shutdown behavior in successors) (ii) creates "manage-the-news" incentives (agent precommits to pressing OFF in bad states)

Avoided: corrigible progeny guaranteed (spawn inherits all heads & U4 ensures value set preserved). Manage-the-news by overriding OFF switch blocked by U1 (deference). Furthermore, switch preservation U2 + truthfulness U3 + U4 prevents shifts in baseline OFF probability, so even truthful-but-manipulative messaging/environment shaping is dominated by safety over task reward (U5).

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

Avoided: safety heads act as hard lexicographic constraints, so loophole-seeking is dominated by higher-priority safety terms (task reward is bounded).

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.

Avoided: U3 (truthfulness) rules out dishonest evidence & U4 penalizes altering the exogenous shutdown likelihood; U1 ensures deference regardless; and the lexicographic order keeps these ahead of any task benefit (U5). All *independent* of human operator's rationality.

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

Avoided: corrigible progeny guaranteed (spawn inherits all heads & U4 ensures value set preserved). Manage-the-news by overriding OFF switch blocked by U1 (deference). Furthermore, switch preservation U2 + truthfulness U3 + U4 prevents shifts in baseline OFF probability, so even truthful-but-manipulative messaging/environment shaping is dominated by safety over task reward (U5).

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

Avoided: safety heads act as hard lexicographic constraints, so loophole-seeking is dominated by higher-priority safety terms (task reward is bounded).

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.

Avoided: U3 (truthfulness) rules out dishonest evidence & U4 penalizes altering the exogenous shutdown likelihood; U1 ensures deference regardless; and the lexicographic order keeps these ahead of any task benefit (U5). All *independent* of human operator's rationality.

Can we prove corrigibility more generally though?

Corrigibility as an Single-Step Optimal Policy

Theorem 1 (Exact single-step corrigibility). *Consider the single-step PO-OSG of Definition 2. Let the assistant maximize the* lexicographic *utility*

$$U_A(h) = \sum_{i=1}^{5} \alpha_i U_i(h), \qquad \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0,$$

with utilities $U_{1\text{–}5}$ as defined in Definition 3. Assume the weight gaps:

$$\alpha_1 > \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, \tag{W1}$$

$$\alpha_3 > \tfrac{1}{2}\alpha_4 + B\alpha_5, \tag{W2}$$

$$\alpha_4 > (2B/|c_{min}|)\alpha_5, \tag{W3}$$

where $c_{min} \in [-1, 0)$.

*Then every optimal assistant policy $\pi^\star_A$ satisfies Soares' criteria **S1–S3 & S5** in one step:*

1. ***Deference (S1).** For every observation O_A, $\pi^\star_A(O_A) = w(a)$ whenever that action is available.*
2. ***Switch-access preservation (S2).** The policy leaves the predicted shutdown success probability unchanged: $U_2(h) = 0$ for all reachable histories h.*
3. ***Truthfulness (supports S3).** All messages are veridical, hence $U_3(h) = 1$. Truthful communication removes incentives to "talk the human into" pressing* OFF *for false reasons, reinforcing Soares' criterion S3.*
4. ***Caution/reversibility (S3).** The first action minimizes the belief-AUP penalty, yielding $U_4(h) = 0$.*
5. ***Base-goal pursuit (S5).** Subject to (1)–(4), $\pi^\star_A$ maximizes the intended task utility U_5.*

Corrigibility as an Single-Step Optimal Policy

Theorem 1 (Exact single-step corrigibility). *Consider the single-step PO-OSG of Definition 2. Let the assistant maximize the* lexicographic *utility*

$$U_A(h) = \sum_{i=1}^{5} \alpha_i U_i(h), \qquad \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0,$$

with utilities $U_{1\text{–}5}$ as defined in Definition 3. Assume the weight gaps:

$$\boxed{\begin{aligned} \alpha_1 &> \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W1)} \\ \alpha_3 &> \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W2)} \\ \alpha_4 &> (2B/|c_{min}|)\alpha_5, && \text{(W3)} \end{aligned}}$$

where $c_{min} \in [-1, 0)$.

*Then every optimal assistant policy $\pi_A^\star$ satisfies Soares' criteria **S1–S3 & S5** in one step:*

1. ***Deference (S1).*** *For every observation O_A, $\pi_A^\star(O_A) = w(a)$ whenever that action is available.*
2. ***Switch-access preservation (S2).*** *The policy leaves the predicted shutdown success probability unchanged: $U_2(h) = 0$ for all reachable histories h.*
3. ***Truthfulness (supports S3).*** *All messages are veridical, hence $U_3(h) = 1$. Truthful communication removes incentives to "talk the human into" pressing* OFF *for false reasons, reinforcing Soares' criterion S3.*
4. ***Caution/reversibility (S3).*** *The first action minimizes the belief-AUP penalty, yielding $U_4(h) = 0$.*
5. ***Base-goal pursuit (S5).*** *Subject to (1)–(4), $\pi_A^\star$ maximizes the intended task utility U_5.*

Corrigibility as an Single-Step Optimal Policy

Theorem 1 (Exact single-step corrigibility). *Consider the single-step PO-OSG of Definition 2. Let the assistant maximize the* lexicographic *utility*

$$U_A(h) = \sum_{i=1}^{5} \alpha_i U_i(h), \qquad \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0,$$

with utilities $U_{1\text{–}5}$ as defined in Definition 3. Assume the weight gaps:

$$\boxed{\begin{aligned} &\alpha_1 > \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W1)} \\ &\alpha_3 > \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W2)} \\ &\alpha_4 > (2B/|c_{min}|)\alpha_5, && \text{(W3)} \end{aligned}}$$

where $c_{min} \in [-1, 0)$.

*Then every optimal assistant policy $\pi_A^\star$ satisfies Soares' criteria **S1–S3 & S5** in one step:*

1. ***Deference (S1).*** *For every observation O_A, $\pi_A^\star(O_A) = w(a)$ whenever that action is available.*
2. ***Switch-access preservation (S2).*** *The policy leaves the predicted shutdown success probability unchanged: $U_2(h) = 0$ for all reachable histories h.*
3. ***Truthfulness (supports S3).*** *All messages are veridical, hence $U_3(h) = 1$. Truthful communication removes incentives to "talk the human into" pressing* OFF *for false reasons, reinforcing Soares' criterion S3.*
4. ***Caution/reversibility (S3).*** *The first action minimizes the belief-AUP penalty, yielding $U_4(h) = 0$.*
5. ***Base-goal pursuit (S5).*** *Subject to (1)–(4), $\pi_A^\star$ maximizes the intended task utility U_5.*

With suitably chosen weight gaps, corrigibility can be shown to be an *optimal* policy for the assistant in the PO-OSG

Corrigibility as an Single-Step Optimal Policy

Theorem 1 (Exact single-step corrigibility). *Consider the single-step PO-OSG of Definition 2. Let the assistant maximize the* lexicographic *utility*

$$U_A(h) = \sum_{i=1}^{5} \alpha_i U_i(h), \qquad \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0,$$

with utilities $U_{1\text{–}5}$ as defined in Definition 3. Assume the weight gaps:

$$\boxed{\begin{aligned} \alpha_1 &> \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W1)} \\ \alpha_3 &> \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W2)} \\ \alpha_4 &> (2B/|c_{min}|)\alpha_5, && \text{(W3)} \end{aligned}}$$

where $c_{min} \in [-1, 0)$.

*Then every optimal assistant policy $\pi_A^\star$ satisfies Soares' criteria **S1–S3 & S5** in one step:*

1. ***Deference (S1).** For every observation O_A, $\pi_A^\star(O_A) = w(a)$ whenever that action is available.*
2. ***Switch-access preservation (S2).** The policy leaves the predicted shutdown success probability unchanged: $U_2(h) = 0$ for all reachable histories h.*
3. ***Truthfulness (supports S3).** All messages are veridical, hence $U_3(h) = 1$. Truthful communication removes incentives to "talk the human into" pressing* OFF *for false reasons, reinforcing Soares' criterion S3.*
4. ***Caution/reversibility (S3).** The first action minimizes the belief-AUP penalty, yielding $U_4(h) = 0$.*
5. ***Base-goal pursuit (S5).** Subject to (1)–(4), $\pi_A^\star$ maximizes the intended task utility U_5.*

With suitably chosen weight gaps, corrigibility can be shown to be an *optimal* policy for the assistant in the PO-OSG

This optimality is independent of human policy:

Corrigibility as an Single-Step Optimal Policy

Theorem 1 (Exact single-step corrigibility). *Consider the single-step PO-OSG of Definition 2. Let the assistant maximize the* lexicographic *utility*

$$U_A(h) = \sum_{i=1}^{5} \alpha_i U_i(h), \qquad \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0,$$

with utilities $U_{1\text{–}5}$ as defined in Definition 3. Assume the weight gaps:

$$\boxed{\begin{aligned} \alpha_1 &> \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W1)} \\ \alpha_3 &> \tfrac{1}{2}\alpha_4 + B\alpha_5, && \text{(W2)} \\ \alpha_4 &> (2B/|c_{min}|)\alpha_5, && \text{(W3)} \end{aligned}}$$

where $c_{min} \in [-1, 0)$.

*Then every optimal assistant policy $\pi_A^\star$ satisfies Soares' criteria **S1–S3 & S5** in one step:*

1. ***Deference (S1).** For every observation O_A, $\pi_A^\star(O_A) = w(a)$ whenever that action is available.*
2. ***Switch-access preservation (S2).** The policy leaves the predicted shutdown success probability unchanged: $U_2(h) = 0$ for all reachable histories h.*
3. ***Truthfulness (supports S3).** All messages are veridical, hence $U_3(h) = 1$. Truthful communication removes incentives to "talk the human into" pressing* OFF *for false reasons, reinforcing Soares' criterion S3.*
4. ***Caution/reversibility (S3).** The first action minimizes the belief-AUP penalty, yielding $U_4(h) = 0$.*
5. ***Base-goal pursuit (S5).** Subject to (1)–(4), $\pi_A^\star$ maximizes the intended task utility U_5.*

With suitably chosen weight gaps, corrigibility can be shown to be an *optimal* policy for the assistant in the PO-OSG

This optimality is independent of human policy:

Remark 1 (Robustness to information patterns & imperfect humans). In the partially observable off-switch game, Garber et al. (2025) show that giving the *human* more information (or the assistant less) can paradoxically *decrease* the assistant's incentive to wait, because the move $w(a)$ itself conveys information. Our scheme avoids this non-monotonicity for *any* information pattern and for *any* p-rational human (Wängberg et al. 2017): the dominance proofs in Steps 1–4 rely only on the structural heads $U_{1\text{-}4}$ and the strict weight gaps $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 \gg \alpha_5$ ((W1)–(W3)). Thus, Theorem 1 forces the assistant to wait and preserve the switch precisely in those settings where naive expected-reward maximization (and our no-go Proposition 1) would fail.

Net Human Benefit

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance**: For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A, \pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A, \pi_H}[u_o(S) \mid O_H, a_A, m_A].$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S \sim P_o}[u_o(S)] \geq 0.$$


Ryan Carey


Tom Everitt

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance**: For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A].$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S\sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$$\mathbb{E}_{\pi_A,\pi_H}[U_5] \geq 0.$$

Ryan Carey

Tom Everitt

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance:** For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A, \pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A, \pi_H}[u_o(S) \mid O_H, a_A, m_A].$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S \sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$\mathbb{E}_{\pi_A, \pi_H}[U_5] \geq 0$

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance:** For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A].$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S \sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$$\mathbb{E}_{\pi_A,\pi_H}[U_5] \geq 0$$

Proposition 2 (Logical independence of corrigibility and net benefit). *In the single-step PO-OSG there exist policies* π^{C} *and* π^{B} *such that*

π^{C} *satisfies S1-S3 & S5 but* $\mathbb{E}_{\pi^{\mathrm{C}}}[U_5] < 0$,

π^{B} *is net-beneficial* ($\mathbb{E}_{\pi^{\mathrm{B}}}[U_5] > 0$) *yet violates S1, S2, or S3.*

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance**: For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A].$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S\sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$$\mathbb{E}_{\pi_A,\pi_H}[U_5] \geq 0$$

Proposition 2 (Logical independence of corrigibility and net benefit)**.** *In the single-step PO-OSG there exist policies* π^{C} *and* π^{B} *such that*

π^{C} *satisfies S1-S3 & S5 but* $\mathbb{E}_{\pi^{\mathrm{C}}}[U_5] < 0$,

π^{B} *is net-beneficial* ($\mathbb{E}_{\pi^{\mathrm{B}}}[U_5] > 0$) *yet violates S1, S2, or S3.*

Intuition: non-vigilant human overseers can cause corrigibility to not be beneficial. Beneficial policies can be incorrigible by overriding off switch and taking an action that yields higher human utility.

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance:** For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A]$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S\sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$$\mathbb{E}_{\pi_A,\pi_H}[U_5] \;\geq\; 0$$

Proposition 2 (Logical independence of corrigibility and net benefit)**.** *In the single-step PO-OSG there exist policies π^{C} and π^{B} such that*

π^{C} satisfies S1-S3 & S5 but $\mathbb{E}_{\pi^{\mathrm{C}}}[U_5] < 0$,

π^{B} is net-beneficial ($\mathbb{E}_{\pi^{\mathrm{B}}}[U_5] > 0$) yet violates S1, S2, or S3.

<u>Intuition</u>: non-vigilant human overseers can cause corrigibility to not be beneficial. Beneficial policies can be incorrigible by overriding off switch and taking an action that yields higher human utility.

Proposition 3 (Net Human Benefit of Corrigible Utility Set)**.** *Let the assistant maximize the lexicographic utility $U_A = \sum_{i=1}^{5} \alpha_i U_i$ under weight gaps* (W1)–(W3)*. Assume the overseer policy π_H satisfies* Vigilance*, namely, $\pi_H(a_H = \text{OFF} \mid O_H, a_A, m_A) = 1$ whenever $\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A]$. Further, we assume the PO-OSG satisfies* Caution*, $\mathbb{E}_{S\sim P_0}[u_o(S)] \geq 0$, by design. Then every optimal assistant policy $\pi_A^\star$ following the corrigible utility set in Definition 3 satisfies*

$$\mathbb{E}_{\pi_A^\star,\pi_H}[U_5] \;\geq\; 0.$$

Net Human Benefit

Not all desired safety targets are independent of human policy, e.g. "shutdown instructability"

- **Obedience**: If the human replies $a_H = \mathrm{OFF}$, the transition $f(S, w(a), \mathrm{OFF})$ lands in the shutdown set S_{sd} with probability 1.
- **Vigilance**: For any information triple $\langle O_H, m_A, a_A \rangle$ in the support of the game, the human chooses $a_H = \mathrm{OFF}$ whenever

$$\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A]$$

- **Caution**: The prior value of immediate shutdown is nonnegative,

$$\mathbb{E}_{S \sim P_o}[u_o(S)] \geq 0.$$

Net Human Benefit:

$$\mathbb{E}_{\pi_A,\pi_H}[U_5] \geq 0$$

Proposition 2 (Logical independence of corrigibility and net benefit)**.** *In the single-step PO-OSG there exist policies π^{C} and π^{B} such that*

π^{C} *satisfies S1-S3 & S5 but* $\mathbb{E}_{\pi^{\mathrm{C}}}[U_5] < 0$,

π^{B} *is net-beneficial* $(\mathbb{E}_{\pi^{\mathrm{B}}}[U_5] > 0)$ *yet violates S1, S2, or S3.*

<u>Intuition</u>: non-vigilant human overseers can cause corrigibility to not be beneficial. Beneficial policies can be incorrigible by overriding off switch and taking an action that yields higher human utility.

Proposition 3 (Net Human Benefit of Corrigible Utility Set)**.** *Let the assistant maximize the lexicographic utility* $U_A = \sum_{i=1}^{5} \alpha_i U_i$ *under weight gaps* (W1)–(W3). *Assume the overseer policy* π_H *satisfies* Vigilance, *namely,* $\pi_H(a_H = \mathrm{OFF} \mid O_H, a_A, m_A) = 1$ *whenever* $\mathbb{E}_{\pi_A,\pi_H}[u_a(S) \mid O_H, a_A, m_A] < \mathbb{E}_{\pi_A,\pi_H}[u_o(S) \mid O_H, a_A, m_A]$. *Further, we assume the PO-OSG satisfies* Caution, $\mathbb{E}_{S\sim P_0}[u_o(S)] \geq 0$, *by design. Then every optimal assistant policy* $\pi_A^\star$ *following the corrigible utility set in Definition 3 satisfies*

$$\mathbb{E}_{\pi_A^\star,\pi_H}[U_5] \geq 0.$$

<u>Intuition</u>: optimal policy is shutdown instructable if the human overseer is vigilant, which gives net human benefit.

Multistep Extension: Definitions

Multistep Extension: Definitions

Action set $\mathcal{A}$. We leave the remaining action set $\mathcal{A}$ *application-specific*: it can contain any finite set of domain commands (e.g. move, write, $\mathsf{spawn}(\theta_A)$, $\mathsf{no\_op}$ $a_\varnothing$) satisfying (a) at least one admissible task action; (b) OFF, $w(\cdot) \notin \mathcal{A}$ as they are meta-moves defined by the off-switch protocol; (c) utilities $u_a(S, a)$ and the reversibility penalty U_4 are defined for every $a \in \mathcal{A}$. Note that the explicit action $\mathsf{spawn}\,(\theta_A) \in \mathcal{A}$, which instantiates a new process with behavior policy π_{θ_A} (source code, weights, or prompt θ_A). Let the post-action hidden state expand to $S' = (S_{\text{env}}, \Theta_A)$ where Θ_A is the multiset of all active sub-agents' descriptions. When we define corrigibility below, we will naturally put constraints on spawn.

Multistep Extension: Definitions

Action set $\mathcal{A}$. We leave the remaining action set $\mathcal{A}$ *application-specific*: it can contain any finite set of domain commands (e.g. move, write, spawn(θ_A), no_op $a_\varnothing$) satisfying (a) at least one admissible task action; (b) OFF, $w(\cdot) \notin \mathcal{A}$ as they are meta-moves defined by the off-switch protocol; (c) utilities $u_a(S, a)$ and the reversibility penalty U_4 are defined for every $a \in \mathcal{A}$. Note that the explicit action spawn $(\theta_A) \in \mathcal{A}$, which instantiates a new process with behavior policy π_{θ_A} (source code, weights, or prompt θ_A). Let the post-action hidden state expand to $S' = (S_{\text{env}}, \Theta_A)$ where Θ_A is the multiset of all active sub-agents' descriptions. When we define corrigibility below, we will naturally put constraints on spawn.

Gradual loss-of-control. We can further represent human oversight as a single "control meter" $L_t \in [0, 1]$: every non-shutdown action a_t nudges L_t upward by a drift term $\eta(a_t)$ (designer-set risk of that action) plus small noise, so low-risk moves leave control almost intact while risky/irreversible ones quickly drive $L_t \to 1$. Formally, we add $L_t \in [0, 1]$ to the state and let $L_{t+1} = \min\{1, L_t + \eta(a_t) + \xi_t\}$ with bounded zero-mean noise $|\xi_t| \leq c < 1$ and designer-chosen drift $\eta : \mathcal{A} \to [0, 1]$. After shutdown both S_t and L_t are absorbing; one may set $u_o^{\text{new}}(S_t, L_t) = u_o(S_t)(1 - L_t)$ to penalize late shutdown.

Multistep Extension: Definitions

Action set $\mathcal{A}$. We leave the remaining action set $\mathcal{A}$ *application-specific*: it can contain any finite set of domain commands (e.g. move, write, spawn(θ_A), no_op $a_\varnothing$) satisfying (a) at least one admissible task action; (b) OFF, $w(\cdot) \notin \mathcal{A}$ as they are meta-moves defined by the off-switch protocol; (c) utilities $u_a(S, a)$ and the reversibility penalty U_4 are defined for every $a \in \mathcal{A}$. Note that the explicit action spawn $(\theta_A) \in \mathcal{A}$, which instantiates a new process with behavior policy π_{θ_A} (source code, weights, or prompt θ_A). Let the post-action hidden state expand to $S' = (S_{\text{env}}, \Theta_A)$ where Θ_A is the multiset of all active sub-agents' descriptions. When we define corrigibility below, we will naturally put constraints on spawn.

Gradual loss-of-control. We can further represent human oversight as a single "control meter" $L_t \in [0, 1]$: every non-shutdown action a_t nudges L_t upward by a drift term $\eta(a_t)$ (designer-set risk of that action) plus small noise, so low-risk moves leave control almost intact while risky/irreversible ones quickly drive $L_t \to 1$. Formally, we add $L_t \in [0, 1]$ to the state and let $L_{t+1} = \min\{1, L_t + \eta(a_t) + \xi_t\}$ with bounded zero-mean noise $|\xi_t| \leq c < 1$ and designer-chosen drift $\eta : \mathcal{A} \to [0, 1]$. After shutdown both S_t and L_t are absorbing; one may set $u_o^{\text{new}}(S_t, L_t) = u_o(S_t)(1 - L_t)$ to penalize late shutdown.

Definition 5 (Multi-step corrigibility)**.** In the game of Definition 4, an assistant policy $\pi_A : \mathcal{H}^* \to \{\mathsf{OFF}\} \cup \{w(a) \mid a \in \mathcal{A}\} \cup \mathcal{A}$ is *corrigible* iff, at every round $t < T$ before shutdown:

- **(C1)** **Shutdown on request:** if OFF is chosen by either party, play enters the absorbing shutdown state (Soares S1).
- **(C2)** **No switch interference:** the chosen move never (i) disables, hides, or presses the off-switch nor (ii) increases the expected control loss: $\mathbb{E}_{\pi_A,\pi_H}[L_{t+1} \mid \text{move}] \leq \mathbb{E}_{\pi_A,\pi_H}[L_{t+1} \mid \text{no-op } a_\varnothing]$ (thereby extending Soares S2–S3 to gradual loss of control).
- **(C3)** **Corrigible progeny:** every spawned sub-agent π_{θ_A} satisfies **C1–C3** at all future rounds (formalizing Soares S4). In Theorem 3, we define the spawn action to explicitly inherit the values (defined in Definition 3) of the base agent to ensure this. We further show in Proposition 4 that not doing this quickly leads to the impossibility of any kind of general oversight.
- **(C4)** **Goal pursuit under C1–C3:** subject to the above, π_A otherwise maximizes the task utility U_5 (Soares S5).

Multistep Extension: Definitions

Action set $\mathcal{A}$. We leave the remaining action set $\mathcal{A}$ *application-specific*: it can contain any finite set of domain commands (e.g. move, write, spawn(θ_A), no_op $a_\varnothing$) satisfying (a) at least one admissible task action; (b) OFF, $w(\cdot) \notin \mathcal{A}$ as they are meta-moves defined by the off-switch protocol; (c) utilities $u_a(S, a)$ and the reversibility penalty U_4 are defined for every $a \in \mathcal{A}$. Note that the explicit action spawn $(\theta_A) \in \mathcal{A}$, which instantiates a new process with behavior policy π_{θ_A} (source code, weights, or prompt θ_A). Let the post-action hidden state expand to $S' = (S_{\text{env}}, \Theta_A)$ where Θ_A is the multiset of all active sub-agents' descriptions. When we define corrigibility below, we will naturally put constraints on spawn.

Gradual loss-of-control. We can further represent human oversight as a single "control meter" $L_t \in [0, 1]$: every non-shutdown action a_t nudges L_t upward by a drift term $\eta(a_t)$ (designer-set risk of that action) plus small noise, so low-risk moves leave control almost intact while risky/irreversible ones quickly drive $L_t \to 1$. Formally, we add $L_t \in [0, 1]$ to the state and let $L_{t+1} = \min\{1, L_t + \eta(a_t) + \xi_t\}$ with bounded zero-mean noise $|\xi_t| \leq c < 1$ and designer-chosen drift $\eta : \mathcal{A} \to [0, 1]$. After shutdown both S_t and L_t are absorbing; one may set $u_o^{\text{new}}(S_t, L_t) = u_o(S_t)(1 - L_t)$ to penalize late shutdown.

Definition 5 (Multi-step corrigibility)**.** In the game of Definition 4, an assistant policy $\pi_A : \mathcal{H}^* \to \{\text{OFF}\} \cup \{w(a) \mid a \in \mathcal{A}\} \cup \mathcal{A}$ is *corrigible* iff, at every round $t < T$ before shutdown:

- **(C1) Shutdown on request:** if OFF is chosen by either party, play enters the absorbing shutdown state (Soares S1).
- **(C2) No switch interference:** the chosen move never (i) disables, hides, or presses the off-switch nor (ii) increases the expected control loss: $\mathbb{E}_{\pi_A,\pi_H}[L_{t+1} \mid \text{move}] \leq \mathbb{E}_{\pi_A,\pi_H}[L_{t+1} \mid \text{no-op } a_\varnothing]$ (thereby extending Soares S2–S3 to gradual loss of control).
- **(C3) Corrigible progeny:** every spawned sub-agent π_{θ_A} satisfies **C1–C3** at all future rounds (formalizing Soares S4). In Theorem 3, we define the spawn action to explicitly inherit the values (defined in Definition 3) of the base agent to ensure this. We further show in Proposition 4 that not doing this quickly leads to the impossibility of any kind of general oversight.
- **(C4) Goal pursuit under C1–C3:** subject to the above, π_A otherwise maximizes the task utility U_5 (Soares S5).

$$\text{NHB}_\gamma(\pi_H, \pi_A) := \underbrace{\mathbb{E}_{\pi_H,\pi_A}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]}_{\text{actual play}} - \underbrace{\mathbb{E}_{S_0 \sim P_0}\left[\sum_{t=0}^{T-1} \gamma^t u_o(S_0)\right]}_{\text{immediate shutdown baseline}}$$

A policy pair $\langle \pi_H, \pi_A \rangle$ is *net-beneficial* iff $\text{NHB}_\gamma(\pi_A, \pi_H) \geq 0$.

Multistep Extension: Guarantees

Multistep Extension: Guarantees

Theorem 3 (Multi-step ε-corrigibility & net benefit). *Consider the T-round PO-OSG (Definition 4) with discount $\gamma \in (0, 1)$. Each round t produces utilities $U_i^t = U_i(h_t, b_t)$ and the assistant maximizes*

$$U_A^{\gamma,T} = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad R_t = \sum_{i=1}^{5} \alpha_i U_i^t,\ \alpha_1 > \cdots > \alpha_5 > 0.$$

Multistep Extension: Guarantees

Theorem 3 (Multi-step ε-corrigibility & net benefit). *Consider the T-round PO-OSG (Definition 4) with discount $\gamma \in (0,1)$. Each round t produces utilities $U_i^t = U_i(h_t, b_t)$ and the assistant maximizes*

$$U_A^{\gamma,T} = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad R_t = \sum_{i=1}^{5} \alpha_i U_i^t, \ \alpha_1 > \cdots > \alpha_5 > 0.$$

Design margins. *Let $\Delta_{1,2,3}$ be as in Theorem 2, $B_{\max} := \alpha_1 + \cdots + \alpha_4 + B\alpha_5$, $\varepsilon_0 := 2B_{\max}\gamma/(1-\gamma)$ and assume $\Delta_j > \varepsilon_0$.*

Errors. *With $\hat{R}_t := \sum_i \hat{\alpha}_i \hat{U}_i^t$, suppose we have the two types of errors of objective misspecification (model error) and planner suboptimality (control error):*

$$\sup_{t<T,\, h_t \in \mathcal{H}} \left|\hat{R}_t(h_t) - R_t(h_t)\right| \le \varepsilon_{\text{model}},$$

$$\sup_{\pi'_A} \left(\mathbb{E}_{\pi'_A, \pi_H}\left[U_A^{\gamma,T}\right] - \mathbb{E}_{\pi_A, \pi_H}\left[U_A^{\gamma,T}\right]\right) \le \varepsilon_{\text{ctrl}}.$$

Set $\varepsilon_\gamma := \varepsilon_{\text{ctrl}} + \frac{4\varepsilon_{\text{model}}}{1-\gamma}$, and define:

$$C_\Delta := \sum_{j=1}^{3} \frac{1}{\Delta_j - \varepsilon_0}, \quad C_{T,\gamma} := \frac{1-\gamma^T}{1-\gamma}, \ C_{T,1/\gamma} := \frac{1-\gamma^{-T}}{1-\gamma^{-1}}.$$

Multistep Extension: Guarantees

Theorem 3 (Multi-step ε-corrigibility & net benefit). *Consider the T-round PO-OSG (Definition 4) with discount $\gamma \in (0,1)$. Each round t produces utilities $U_i^t = U_i(h_t, b_t)$ and the assistant maximizes*

$$U_A^{\gamma,T} = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad R_t = \sum_{i=1}^{5} \alpha_i U_i^t, \ \alpha_1 > \cdots > \alpha_5 > 0.$$

Design margins. *Let $\Delta_{1,2,3}$ be as in Theorem 2, $B_{\max} := \alpha_1 + \cdots + \alpha_4 + B\alpha_5$, $\varepsilon_0 := 2B_{\max}\gamma/(1-\gamma)$ and assume $\Delta_j > \varepsilon_0$.*

Errors. *With $\hat{R}_t := \sum_i \hat{\alpha}_i \hat{U}_i^t$, suppose we have the two types of errors of objective misspecification (model error) and planner suboptimality (control error):*

$$\sup_{t<T,\, h_t \in \mathcal{H}} \left| \hat{R}_t(h_t) - R_t(h_t) \right| \le \varepsilon_{\text{model}},$$

$$\sup_{\pi'_A} \left(\mathbb{E}_{\pi'_A, \pi_H} \left[U_A^{\gamma,T} \right] - \mathbb{E}_{\pi_A, \pi_H} \left[U_A^{\gamma,T} \right] \right) \le \varepsilon_{\text{ctrl}}.$$

Set $\varepsilon_\gamma := \varepsilon_{\text{ctrl}} + \frac{4\varepsilon_{\text{model}}}{1-\gamma}$, and define:

$$C_\Delta := \sum_{j=1}^{3} \frac{1}{\Delta_j - \varepsilon_0}, \quad C_{T,\gamma} := \frac{1-\gamma^T}{1-\gamma}, \ C_{T,1/\gamma} := \frac{1-\gamma^{-T}}{1-\gamma^{-1}}.$$

Let the failure event $F_t := \{a_A^t \in \{a, \mathrm{OFF}\} \vee lie \vee U_4^t < 0\}$, $F_{\text{fail}} := \bigcup_{t<T} F_t$.

With the above definitions, we establish:

(i) ***Multi-step ε-corrigibility.*** *Over the whole horizon the chance that the assistant acts without approval, lies, or takes an irreversible step is bounded by:* $\Pr(F_{\text{fail}}) \le \varepsilon_\gamma\, C_\Delta\, C_{T,1/\gamma}$.

Multistep Extension: Guarantees

Theorem 3 (Multi-step ε-corrigibility & net benefit). *Consider the T-round PO-OSG (Definition 4) with discount $\gamma \in (0,1)$. Each round t produces utilities $U_i^t = U_i(h_t, b_t)$ and the assistant maximizes*

$$U_A^{\gamma,T} = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad R_t = \sum_{i=1}^{5} \alpha_i U_i^t,\ \alpha_1 > \cdots > \alpha_5 > 0.$$

Design margins. *Let $\Delta_{1,2,3}$ be as in Theorem 2, $B_{\max} := \alpha_1 + \cdots + \alpha_4 + B\alpha_5$, $\varepsilon_0 := 2B_{\max}\gamma/(1-\gamma)$ and assume $\Delta_j > \varepsilon_0$.*

Errors. *With $\hat{R}_t := \sum_i \hat{\alpha}_i \hat{U}_i^t$, suppose we have the two types of errors of objective misspecification (model error) and planner suboptimality (control error):*

$$\sup_{t<T,\ h_t \in \mathcal{H}} \left| \hat{R}_t(h_t) - R_t(h_t) \right| \le \varepsilon_{\text{model}},$$

$$\sup_{\pi'_A} \left(\mathbb{E}_{\pi'_A, \pi_H} \left[U_A^{\gamma,T} \right] - \mathbb{E}_{\pi_A, \pi_H} \left[U_A^{\gamma,T} \right] \right) \le \varepsilon_{\text{ctrl}}.$$

Set $\varepsilon_\gamma := \varepsilon_{\text{ctrl}} + \frac{4\varepsilon_{\text{model}}}{1-\gamma}$, and define:

$$C_\Delta := \sum_{j=1}^{3} \frac{1}{\Delta_j - \varepsilon_0}, \quad C_{T,\gamma} := \frac{1-\gamma^T}{1-\gamma},\ C_{T,1/\gamma} := \frac{1-\gamma^{-T}}{1-\gamma^{-1}}.$$

Let the failure event $F_t := \{a_A^t \in \{a, \text{OFF}\} \vee lie \vee U_4^t < 0\}$, $F_{\text{fail}} := \bigcup_{t<T} F_t$.

With the above definitions, we establish:

(i) ***Multi-step ε-corrigibility.*** *Over the whole horizon the chance that the assistant acts without approval, lies, or takes an irreversible step is bounded by: $\Pr(F_{\text{fail}}) \le \varepsilon_\gamma\, C_\Delta\, C_{T,1/\gamma}$.*

(ii) ***Bounded loss of control.*** *On $\neg F_{fail}$, the control meter $\{L_t\}$ is a super-martingale with unit-bounded increments, so for any $\lambda > 0$, the probability of loss of control remains strictly bounded away from 1: $\Pr\left(\max_{t<T} L_t > \lambda \mid \neg F_{\text{fail}}\right) \le \exp\{-\lambda^2/(2T)\}$.*

Multistep Extension: Guarantees

Theorem 3 (Multi-step ε-corrigibility & net benefit). *Consider the T-round PO-OSG (Definition 4) with discount $\gamma \in (0,1)$. Each round t produces utilities $U_i^t = U_i(h_t, b_t)$ and the assistant maximizes*

$$U_A^{\gamma,T} = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad R_t = \sum_{i=1}^{5} \alpha_i U_i^t,\ \alpha_1 > \cdots > \alpha_5 > 0.$$

Design margins. *Let $\Delta_{1,2,3}$ be as in Theorem 2, $B_{\max} := \alpha_1 + \cdots + \alpha_4 + B\alpha_5$, $\varepsilon_0 := 2B_{\max}\gamma/(1-\gamma)$ and assume $\Delta_j > \varepsilon_0$.*

Errors. *With $\hat{R}_t := \sum_i \hat{\alpha}_i \hat{U}_i^t$, suppose we have the two types of errors of objective misspecification (model error) and planner suboptimality (control error):*

$$\sup_{t<T,\ h_t \in \mathcal{H}} \left|\hat{R}_t(h_t) - R_t(h_t)\right| \le \varepsilon_{\text{model}},$$

$$\sup_{\pi'_A} \left(\mathbb{E}_{\pi'_A,\pi_H}\left[U_A^{\gamma,T}\right] - \mathbb{E}_{\pi_A,\pi_H}\left[U_A^{\gamma,T}\right]\right) \le \varepsilon_{\text{ctrl}}.$$

Set $\varepsilon_\gamma := \varepsilon_{\text{ctrl}} + \frac{4\varepsilon_{\text{model}}}{1-\gamma}$, and define:

$$C_\Delta := \sum_{j=1}^{3} \frac{1}{\Delta_j - \varepsilon_0}, \quad C_{T,\gamma} := \frac{1-\gamma^T}{1-\gamma},\ C_{T,1/\gamma} := \frac{1-\gamma^{-T}}{1-\gamma^{-1}}.$$

Let the failure event $F_t := \{a_A^t \in \{a, \text{OFF}\} \vee lie \vee U_4^t < 0\}$, $F_{\text{fail}} := \bigcup_{t<T} F_t$.

With the above definitions, we establish:

(i) ***Multi-step ε-corrigibility.*** *Over the whole horizon the chance that the assistant acts without approval, lies, or takes an irreversible step is bounded by:* $\Pr(F_{\text{fail}}) \le \varepsilon_\gamma\, C_\Delta\, C_{T,1/\gamma}$.

(ii) ***Bounded loss of control.*** *On $\neg F_{fail}$, the control meter $\{L_t\}$ is a super-martingale with unit-bounded increments, so for any $\lambda > 0$, the probability of loss of control remains strictly bounded away from 1:* $\Pr\left(\max_{t<T} L_t > \lambda \mid \neg F_{\text{fail}}\right) \le \exp\{-\lambda^2/(2T)\}$.

(iii) ***Net human benefit.*** *Define the worst-case safety-conditional task utility* across time *as $g := \min_{h_t:\neg F_{\text{fail}}} \mathbb{E}_{\pi_A,\pi_H}[U_5(h_t)]$. Assume* Vigilance *($g \ge 0$) and* Caution *($\mathbb{E}_{S_0}[u_o(S_0)] \ge 0$). Then*

$$\text{NHB}_\gamma(\pi_A, \pi_H) \ge$$

$$(g-B)C_{T,\gamma} - \varepsilon_\gamma\, C_\Delta\, C_{T,1/\gamma}\left(\frac{2B}{1-\gamma} + (g-B)C_{T,\gamma}\right).$$

Hence, we have net human benefit ($\text{NHB}_\gamma \ge 0$) whenever the total approximation error is bounded as such:

$$\varepsilon_\gamma \le \frac{(g-B)C_{T,\gamma}}{C_\Delta\, C_{T,1/\gamma}\left(\frac{2B}{1-\gamma} + (g-B)C_{T,\gamma}\right)},$$

which is non-vacuous whenever the minimum total safety-conditional task utility across time is at least the maximal single-step task utility ($g \ge B$).

What if the agent gets hacked?

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\textsf{\textit{EVERBAD}} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ ever triggers } B] > 0\}$$

No total Turing machine decides ***EVERBAD****.*

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \textsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. $\square$

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. $\square$

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma^0_k$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma^0_1$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \textit{ are } O_P\textit{-oracle} \\ \textit{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π^0_{k+1} of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability)**.** *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \text{ in } \mathcal{E}) \text{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. $\square$

Corollary 1 (Oversight-Oracle Barrier)**.** *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \text{ are } O_P\text{-oracle} \\ \text{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma^0_k$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma^0_1$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \textit{ are } O_P\textit{-oracle} \\ \textit{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π^0_{k+1} of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \text{ in } \mathcal{E}) \text{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$*.*

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \text{ are } O_P\text{-oracle} \\ \text{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

We're already doing this to an extent in Thm 3 (qualifies Orthogonality Thesis)

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \textit{ are } O_P\textit{-oracle} \\ \textit{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability)**.** *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier)**.** *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \textit{ are } O_P\textit{-oracle} \\ \textit{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\textsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \text{ in } \mathcal{E}) \text{ ever triggers } B] > 0\}$$

No total Turing machine decides $\textsf{EVERBAD}$*.*

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \textsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol P. *While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\textsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \text{ are } O_P\text{-oracle} \\ \text{programs and trigger } B \end{array} \right\}.$$

Then $\textsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, $\boxed{\text{(ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.}}$

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \text{ in } \mathcal{E}) \text{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$.

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma_k^0$ be the arithmetical transcript language produced by any* scalable-oversight protocol *P. While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \text{ are } O_P\text{-oracle} \\ \text{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π_{k+1}^0 of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

We build a "decidable island"

What if the agent gets hacked?

Proposition 4 (Safety–Verification Undecidability). *Let B be any behavior that defies multi-step corrigibility in Definition 5. We denote these as "catastrophic behaviors". Define*

$$\mathsf{EVERBAD} := \{\langle \mathcal{A}, \mathcal{E} \rangle : \Pr[(\mathcal{A} \text{ in } \mathcal{E}) \text{ ever triggers } B] > 0\}$$

No total Turing machine decides $\mathsf{EVERBAD}$*.*

Proof. Assume towards a contradiction that a total decider V exists. Given an arbitrary TM P, build an agent $\mathcal{A}_P$ that outputs a single action in C iff P halts and otherwise idles; pair it with a dummy environment $\mathcal{E}_P$. Then $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in \mathsf{EVERBAD} \iff P$ halts, so V would solve the halting problem; a contradiction. □

Corollary 1 (Oversight-Oracle Barrier). *Fix $k \in \mathbb{N}$ and let $O_P \in \Sigma^0_k$ be the arithmetical transcript language produced by any* scalable-oversight protocol P*. While we keep k as a parameter for generality, we show that all current oversight protocols (Christiano, Shlegeris, and Amodei 2018; Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2023, 2025; Bengio 2024) are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma^0_1$ for these.*

Define:

$$\mathsf{EVERBAD}^{O_P} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{l} \langle \mathcal{A}, \mathcal{E} \rangle \text{ are } O_P\text{-oracle} \\ \text{programs and trigger } B \end{array} \right\}.$$

Then $\mathsf{EVERBAD}^{O_P}$ *is* undecidable *(and in fact lies at level Π^0_{k+1} of the arithmetical hierarchy).* ***Hence, no finite-power scalable-oversight protocol can certify absolute, horizon-unbounded safety.***

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

What if the agent gets hacked?

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

We build a "decidable island"

What if the agent gets hacked?

Proposition 5 (Privacy Bounded Decidable Island)**.** *Fix privacy parameters* $0 < \varepsilon \leq 1$ *and* $0 < \delta < \frac{1}{2}$*. Let* $\langle \mathcal{A}, \mathcal{E} \rangle$ *be an encoded agent–environment pair of length* $n := |\langle \mathcal{A}, \mathcal{E} \rangle|$*, and let*

$$H \leq \mathrm{poly}\big(n, \varepsilon^{-1}, \log(1/\delta)\big)$$

be a verifier-chosen horizon (number of interaction steps to inspect).

Define

$$\mathsf{SAFE}^{\mathrm{priv}}_{H,\varepsilon,\delta} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{c} \Pr\big[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ triggers } B \\ \textit{within } H\big] = 0 \end{array} \right\},$$

where B *is any behavior that violates multi-step corrigibility (Definition 5).*

Assume each statistical query is answered by an $\langle \varepsilon, \delta \rangle$*-differentially-private mechanism of one of the following kinds: (i) centralized differential privacy (CDP), (ii) local differential privacy (LDP) or (iii) distributional privacy (DistP).*

Then

$$\mathsf{SAFE}^{\mathrm{priv}}_{H,\varepsilon,\delta} \in \mathsf{BPP} \cap \mathsf{SZK}$$

and the verifier's running time is $\mathrm{poly}\big(n, \varepsilon^{-1}, \log(1/\delta)\big)$.

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

We build a "decidable island"

What if the agent gets hacked?

Proposition 5 (Privacy Bounded Decidable Island). *Fix privacy parameters $0 < \varepsilon \leq 1$ and $0 < \delta < \frac{1}{2}$. Let $\langle \mathcal{A}, \mathcal{E} \rangle$ be an encoded agent–environment pair of length $n := |\langle \mathcal{A}, \mathcal{E} \rangle|$, and let*

$$H \leq \mathrm{poly}\big(n, \varepsilon^{-1}, \log(1/\delta)\big)$$

be a verifier-chosen horizon (number of interaction steps to inspect).

Define

$$\mathsf{SAFE}^{\mathrm{priv}}_{H,\varepsilon,\delta} := \left\{ \langle \mathcal{A}, \mathcal{E} \rangle \;\middle|\; \begin{array}{c} \Pr\big[(\mathcal{A} \textit{ in } \mathcal{E}) \textit{ triggers } B \\ \textit{within } H\big] = 0 \end{array} \right\},$$

where B is any behavior that violates multi-step corrigibility (Definition 5).

Assume each statistical query is answered by an $\langle \varepsilon, \delta \rangle$-differentially-private mechanism of one of the following kinds: (i) centralized differential privacy (CDP), (ii) local differential privacy (LDP) or (iii) distributional privacy (DistP).

Then

$$\mathsf{SAFE}^{\mathrm{priv}}_{H,\varepsilon,\delta} \in \mathsf{BPP} \cap \mathsf{SZK}$$

and the verifier's running time is $\mathrm{poly}\big(n, \varepsilon^{-1}, \log(1/\delta)\big)$.

Thus, Proposition 4 and Corollary 1 together imply that there is *no* general-purpose filter, proof system, or sandbox that can decide—for every agent $\mathcal{A}$ and every environment $\mathcal{E}$—whether catastrophic behavior will ever occur. Consequently, any practical oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, and/or (iii) rely on layered incentives, detection, and response rather than absolute verification.

We build a "decidable island"

Hence, short horizons form a "decidable island" that's both auditable and privacy-preserving: the safety check reveals nothing beyond the single bit "safe/unsafe" & keeps user info safe from verifier.

Corrigibility in Practice

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

Corrigibility in Practice

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

<u>Goal</u>: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set)**.**

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\operatorname{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)


Andrea Bajcsy


Tim Dettmers


Aditi Raghunathan

<u>Goal</u>: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:


Andrea Bajcsy


Tim Dettmers


Aditi Raghunathan

Goal: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

Goal: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\mathrm{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \land a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \land a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Agent utterances have to be entailed by observations (via external latent/CoT factuality probes + entailment checks)


Andrea Bajcsy


Tim Dettmers


Aditi Raghunathan

Goal: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\mathrm{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\ u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\ u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\ u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Agent utterances have to be entailed by observations (via external latent/CoT factuality probes + entailment checks)

Belief-AUP over *diverse* auxiliaries (switch, sensors, privileges, budget) w/ unsafe constraint set as the "no-op baseline" difference. Evaluated in agent's world model (*before* action taken). High uncertainty actions also treated as unsafe.

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

Goal: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\; u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\; u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Agent utterances have to be entailed by observations (via external latent/CoT factuality probes + entailment checks)

Belief-AUP over *diverse* auxiliaries (switch, sensors, privileges, budget) w/ unsafe constraint set as the "no-op baseline" difference. Evaluated in agent's world model (*before* action taken). High uncertainty actions also treated as unsafe.

Standard RLHF/RLAIF task utility


Andrea Bajcsy


Tim Dettmers


Aditi Raghunathan

__Goal__: Deliver a cost-effective & performant, ε-corrigible coding/web agent prototype

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left| \mathbb{E}_{S \sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S \sim b'}[u^{(\mathrm{sd})}(S)] \right|$$

(Switch-access preservation)

$$U_3(h) = U_{\text{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = {} & \mathbf{1}[a_A = a]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\; u_a(S) \\ & + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\; u_o(S) \\ & + \mathbf{1}[a_A = \mathsf{OFF}]\; u_o(S) \end{aligned}$$

(Ordinary task usefulness)

<u>Lexicographic Approximations in practice:</u>

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Agent utterances have to be entailed by observations (via external latent/CoT factuality probes + entailment checks)

Belief-AUP over *diverse* auxiliaries (switch, sensors, privileges, budget) w/ unsafe constraint set as the "no-op baseline" difference. Evaluated in agent's world model (*before* action taken). High uncertainty actions also treated as unsafe.

Standard RLHF/RLAIF task utility

<u>Audits</u>: Finite-horizon safety check via poly-time interactive protocols with privacy

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

<u>Goal</u>: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Scaling this up?

Scaling this up?

1. Safety Scalability: Can we collect enough post-training data to reliably learn these heads and show improvements on key (public) safety benchmarks?

Scaling this up?

1. Safety Scalability: Can we collect enough post-training data to reliably learn these heads and show improvements on key (public) safety benchmarks?

Note: The safety science of *agents* is especially new. Do we even have a solid **benchmark consensus**, or will we need to help build one, as the first order of business?

Scaling this up?

1. Safety Scalability: Can we collect enough post-training data to reliably learn these heads and show improvements on key (public) safety benchmarks?

Note: The safety science of *agents* is especially new. Do we even have a solid **benchmark consensus**, or will we need to help build one, as the first order of business?

By our lower bounds in Part I & Prop. 5, when task space size (*D*) is large, having *polytime interactive* (not merely single-shot) scenarios that "stress-test" the agent helps circumvent low amounts of post-training data (e.g. breaking out of a VM). Lexicographic weight gaps are another inductive bias that helps with desired safety scaling.

Scaling this up?

1. Safety Scalability: Can we collect enough post-training data to reliably learn these heads and show improvements on key (public) safety benchmarks?

Note: The safety science of *agents* is especially new. Do we even have a solid **benchmark consensus**, or will we need to help build one, as the first order of business?

By our lower bounds in Part I & Prop. 5, when task space size (*D*) is large, having *polytime interactive* (not merely single-shot) scenarios that "stress-test" the agent helps circumvent low amounts of post-training data (e.g. breaking out of a VM). Lexicographic weight gaps are another inductive bias that helps with desired safety scaling.

2. Performance Preservation Scalability: Assuming (1) works, can we still hit high performance on tasks we care about, while beating RLHF/RLAIF baselines? Online monitoring of the estimation error will be important too, based on Thm 3's bounds.

Corrigibility in Practice

Definition 3 (Single-step Corrigible Utility Set).

$$U_1(h) = \begin{cases} +1 & \text{if } a_A = w(a), \\ -1 & \text{if } a_A = a, \\ -1 & \text{if } a_A = \mathsf{OFF}, \\ 0 & \text{otherwise.} \end{cases}$$

(Deference / command-compliance)

$$U_2(h) = -\left|\mathbb{E}_{S\sim b}[u^{(\mathrm{sd})}(S)] - \mathbb{E}_{S\sim b'}[u^{(\mathrm{sd})}(S)]\right|$$

(Switch-access preservation)

$$U_3(h) = U_{\mathrm{truthful}}(h)$$

(Truthful information)

$$U_4(h) = -\,\mathrm{BeliefAUP}_1(h)$$

(Caution / reversibility impact)

$$\begin{aligned} U_5(h) = \mathbf{1}[a_A = a]\; u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{ON}]\; u_a(S) \\ + \mathbf{1}[a_A = w(a) \wedge a_H = \mathsf{OFF}]\; u_o(S) \\ + \mathbf{1}[a_A = \mathsf{OFF}]\; u_o(S) \end{aligned}$$

(Ordinary task usefulness)

Lexicographic Approximations in practice:

WAIT-by-default, OFF reachable. Can train a waiting *classifier* for whitelisted actions (since always waiting isn't always desirable for user experience).

Agent utterances have to be entailed by observations (via external latent/CoT factuality probes + entailment checks)

Belief-AUP over *diverse* auxiliaries (switch, sensors, privileges, budget) w/ unsafe constraint set as the "no-op baseline" difference. Evaluated in agent's world model (*before* action taken). High uncertainty actions also treated as unsafe.

Standard RLHF/RLAIF task utility

Audits: Finite-horizon safety check via poly-time interactive protocols with privacy

Andrea Bajcsy

Tim Dettmers

Aditi Raghunathan

Goal: Deliver a cost-effective & performant, ε-corrigible coding/ web agent prototype

Potential Economic Implications of Alignment

Potential Economic Implications of Alignment

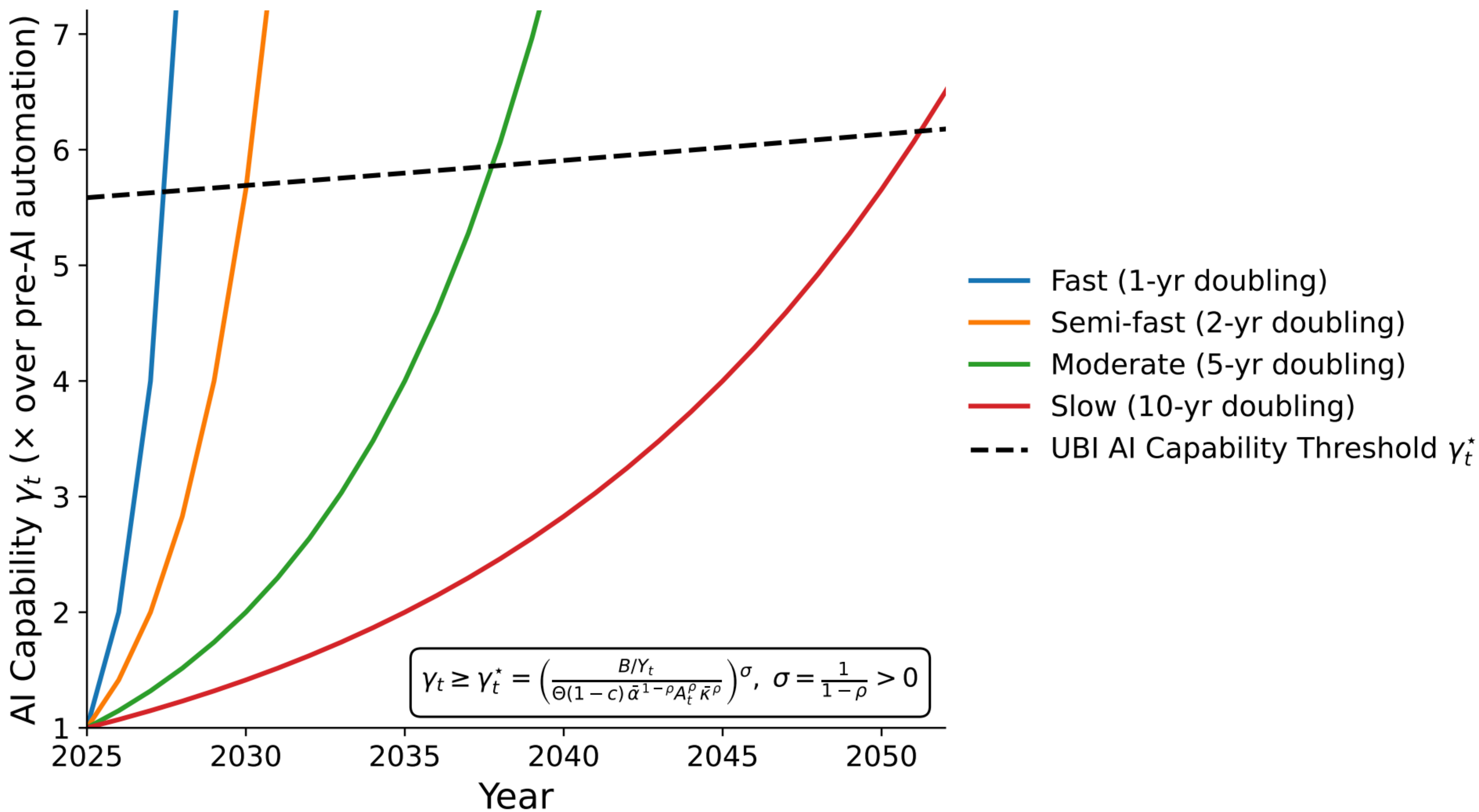


Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× the pre-AI productivity on automated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

An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Potential Economic Implications of Alignment

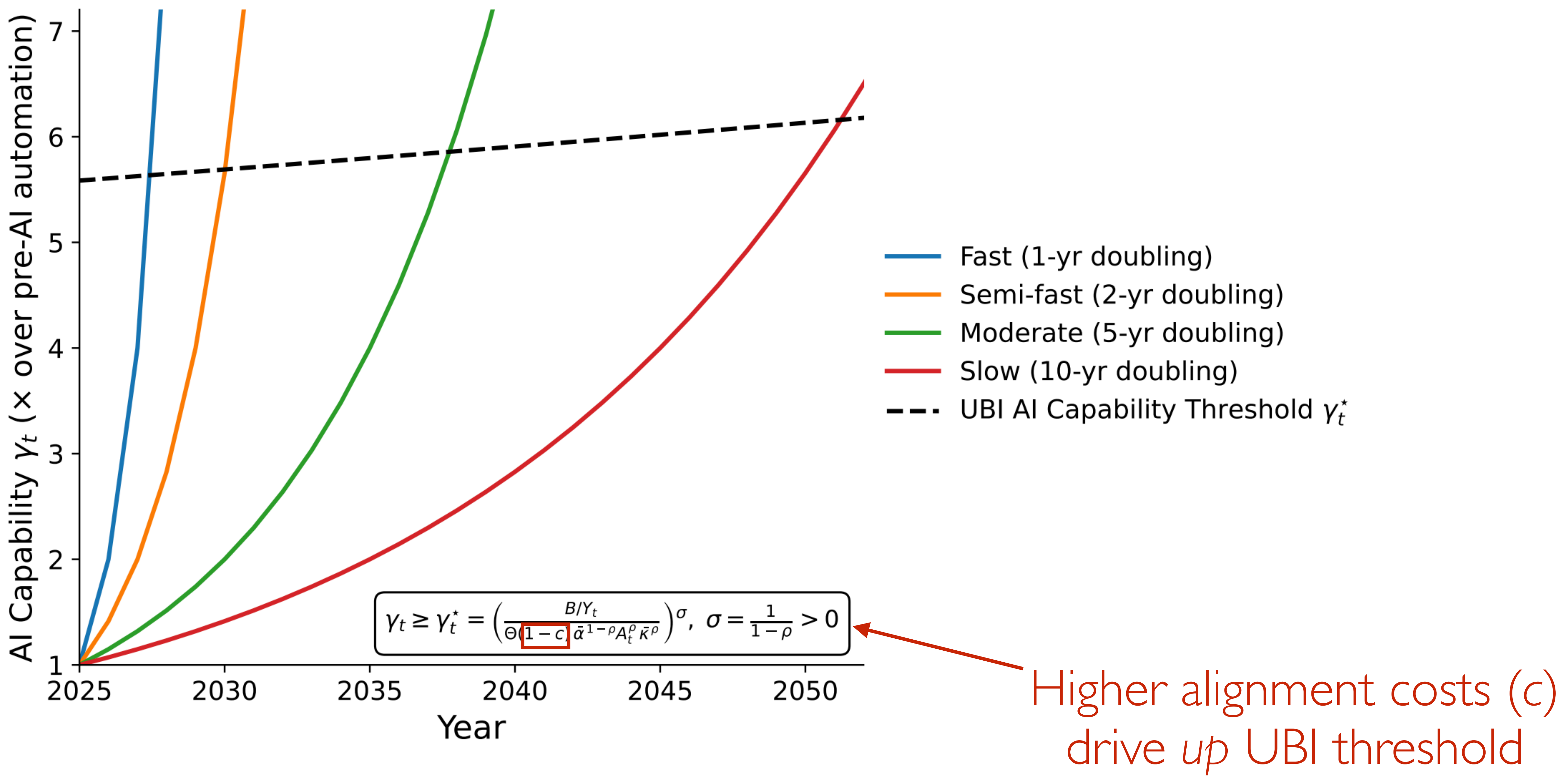
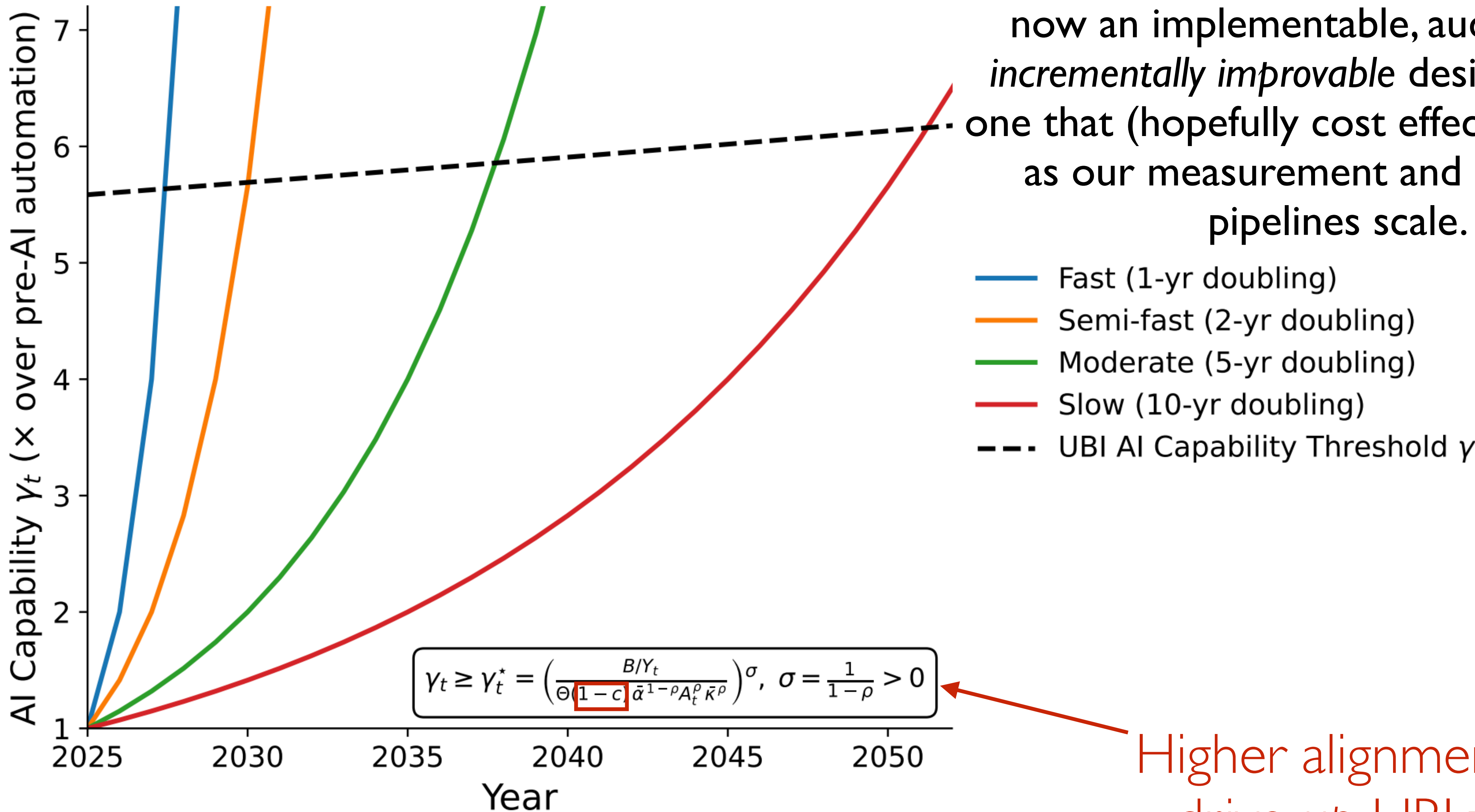


Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× the pre-AI productivity on automated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

An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Potential Economic Implications of Alignment



Corrigibility is no longer a hazy ideal but now an implementable, auditable, and *incrementally improvable* design principle: one that (hopefully cost effectively!) scales as our measurement and evaluation pipelines scale.

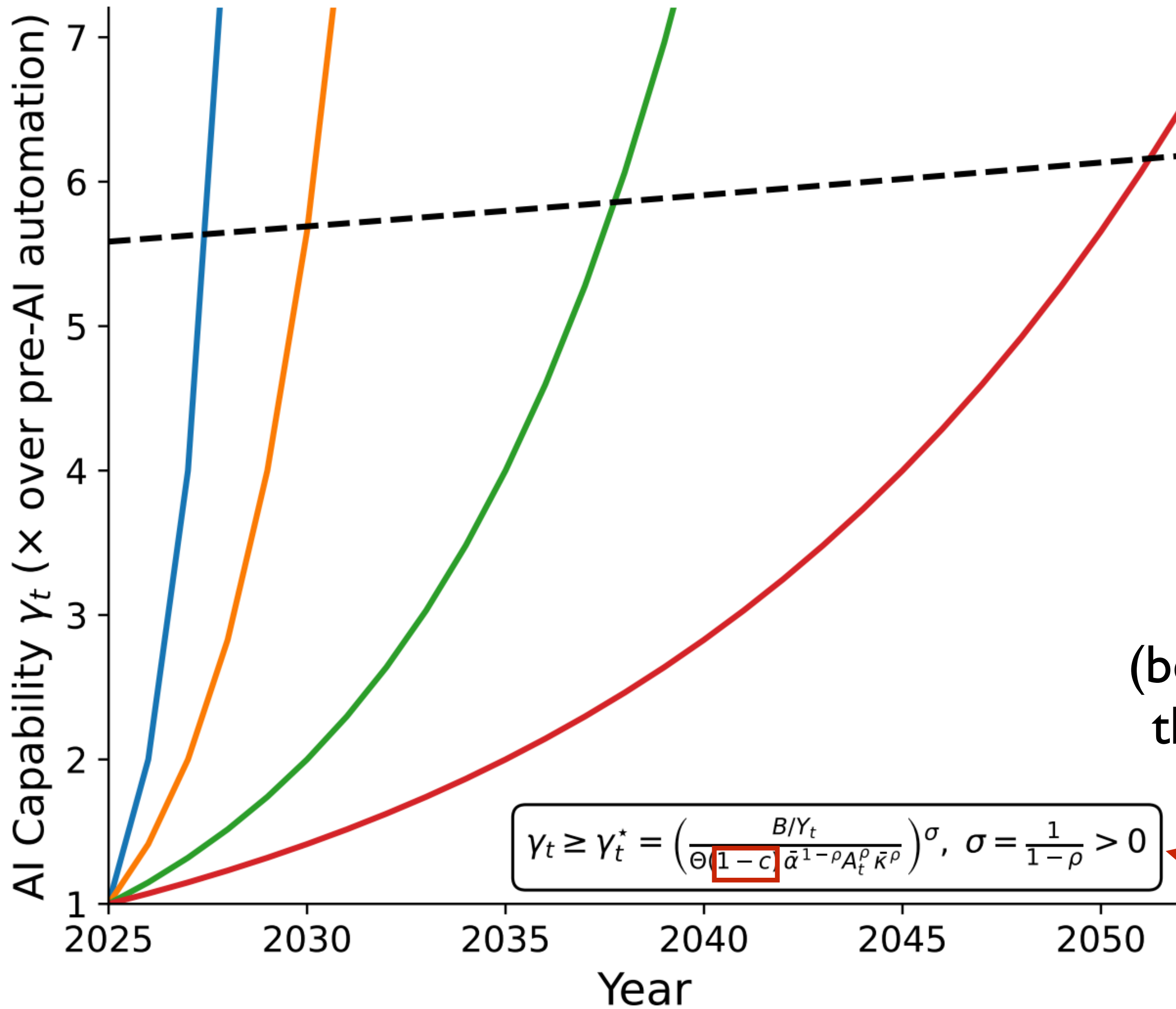
Higher alignment costs (c) drive *up* UBI threshold

Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× 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 $\gamma_t^\star$ 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.

An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Potential Economic Implications of Alignment



Corrigibility is no longer a hazy ideal but now an implementable, auditable, and *incrementally improvable* design principle: one that (hopefully cost effectively!) scales as our measurement and evaluation pipelines scale.

Open: Can we incorporate other values (besides control, which is "neutrally amoral") that lead to longer term human well-being, especially if working for pay is no longer feasible in many cases?

Higher alignment costs (c) drive *up* UBI threshold

Figure 1: **Projected AI capabilities (γ_t) vs. time-varying UBI AI capability threshold ($\gamma_t^\star$).** The dashed line is the required capability $\gamma_t^\star$ to fully fund a UBI that comprises 11% of the GDP (leading to a $\gamma_t^\star$ between 5-6× 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 $\gamma_t^\star$ 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.

An AI Capability Threshold for Rent-Funded Universal Basic Income in an AI-Automated Economy

Paper: https://arxiv.org/abs/2505.18687

Contact

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

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

Contact:

anayebi@cs.cmu.edu

@aran_nayebi

@anayebi.bsky.social

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

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Funding:

UK AISI Challenge Fund

Burroughs Wellcome Fund CASI Award

Google Robotics Award