AAAI-26 / IAAI-26 / EAAI-26

Core Safety Values for Provably Corrigible Agents

Aran Nayebi

Assistant Professor

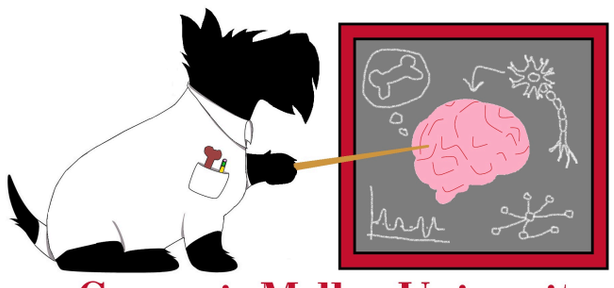
Machine Learning Department

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

AAAI 2026, Machine Ethics Workshop (W37)

2026.01.27

MACHINE LEARNING DEPARTMENT



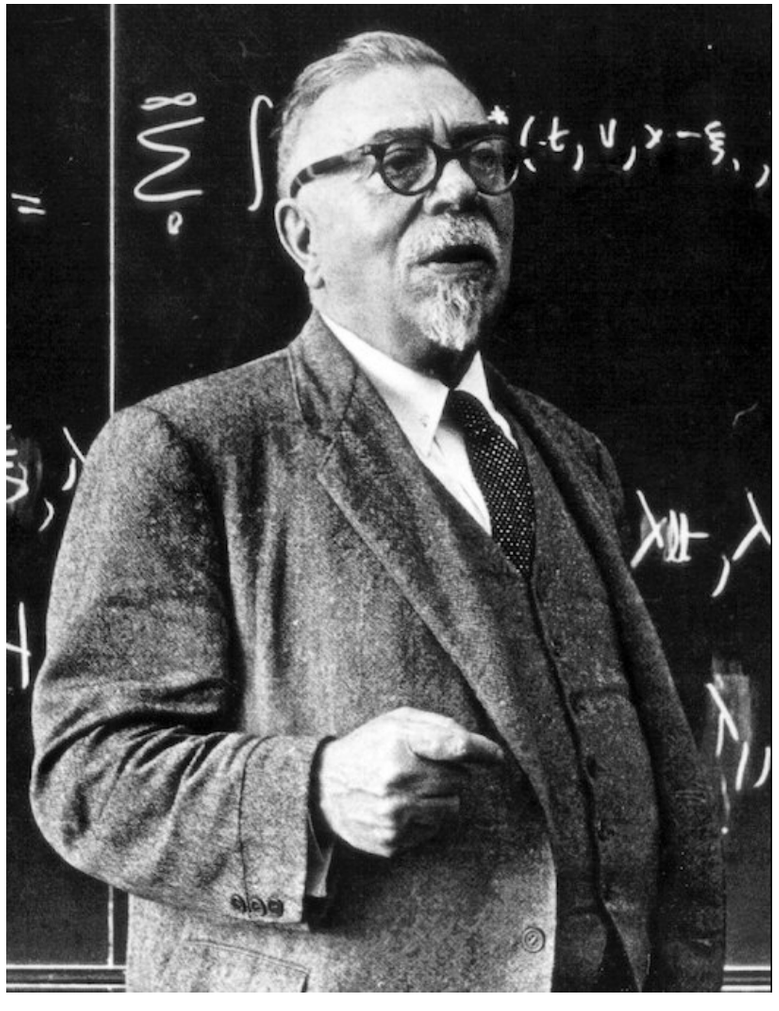
Carnegie Mellon University
Neuroscience Institute

THE ROBOTICS INSTITUTE

Carnegie Mellon
SCHOOL OF COMPUTER SCIENCE

Alignment Problem

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



Some Moral and Technical Consequences of Automation

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

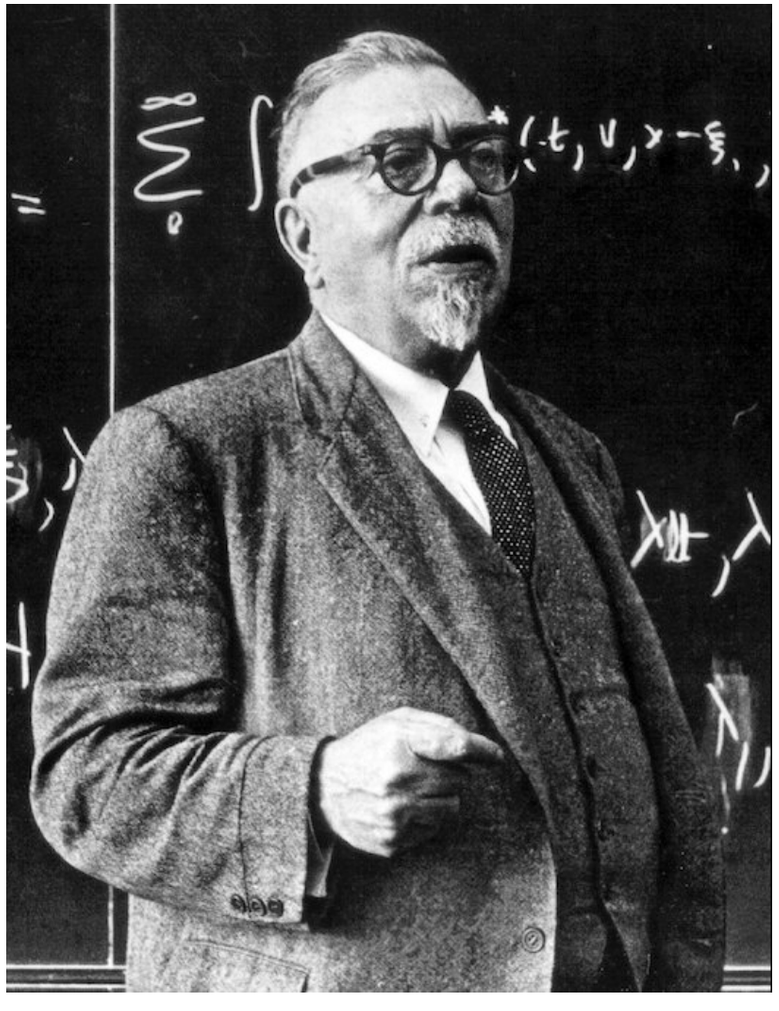
Norbert Wiener

6 MAY 1960

Alignment Problem**s**

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

What should those values even *be*?



Some Moral and Technical Consequences of Automation

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

Norbert Wiener

6 MAY 1960

Approaching Alignment

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

What should those values even *be*? (this talk)

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

↓

Core Safety Values for Provably Corrigible Agents

Approaching Alignment

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

What should those values even *be*? (this talk)

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

↓

Core Safety Values for Provably Corrigible Agents

Approaching Alignment

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

What should those values even *be*? (this talk)

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

↓

Core Safety Values for Provably Corrigible Agents

Approaching Alignment

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

- Aligning to "all human values" will *not* work (No Free Lunch)

- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even *be*? (this talk)

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

↓

Core Safety Values for Provably Corrigible Agents

Approaching Alignment

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

- Aligning to "all human values" will *not* work (No Free Lunch)

- Reward hacking is *inevitable* in large state spaces & bounded agents (so select important parts of the state space + mechanism design)

What should those values even *be*? (this talk)

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

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

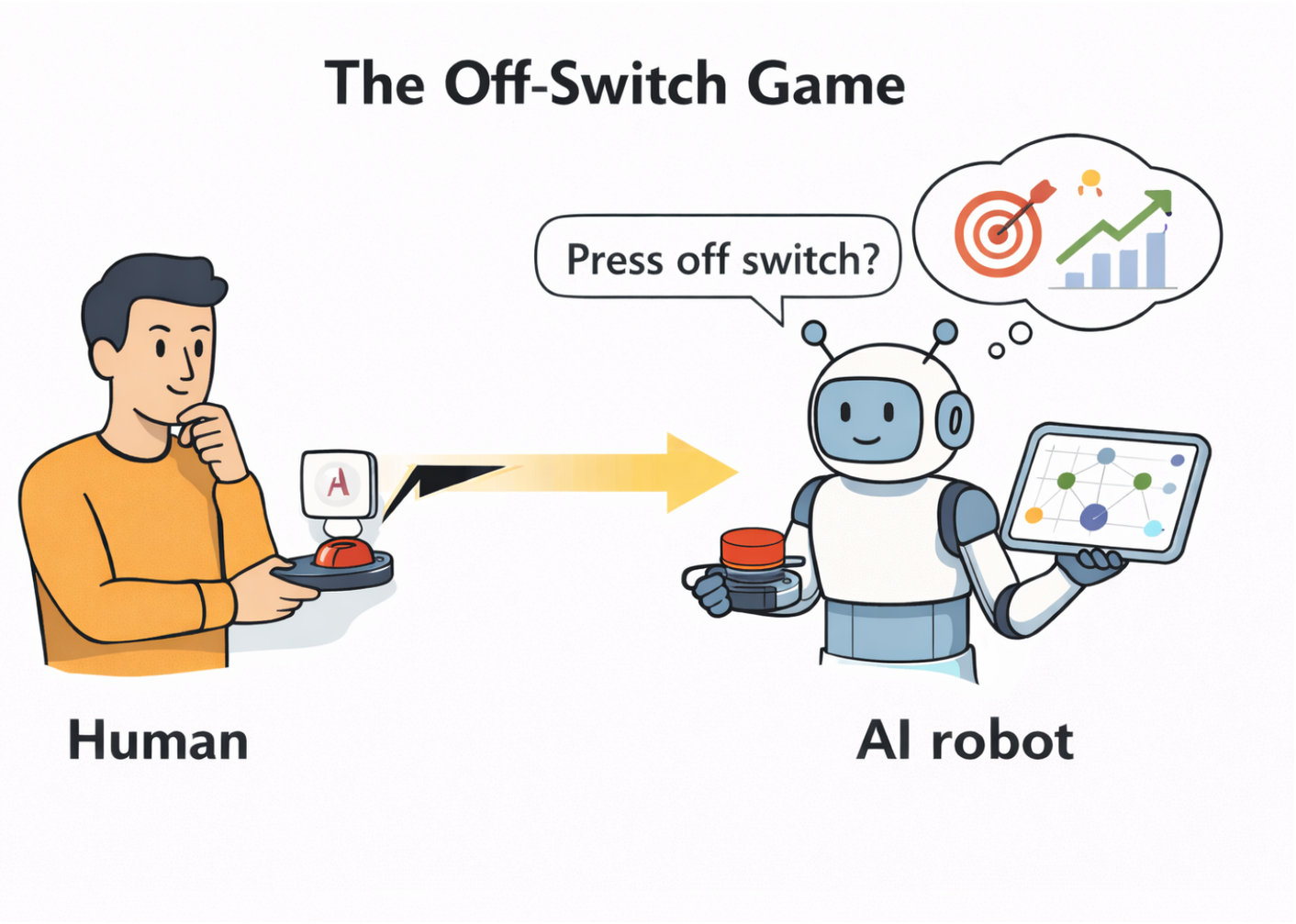
↓

Core Safety Values for Provably Corrigible Agents

What is Corrigibility? Off-Switch Game 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



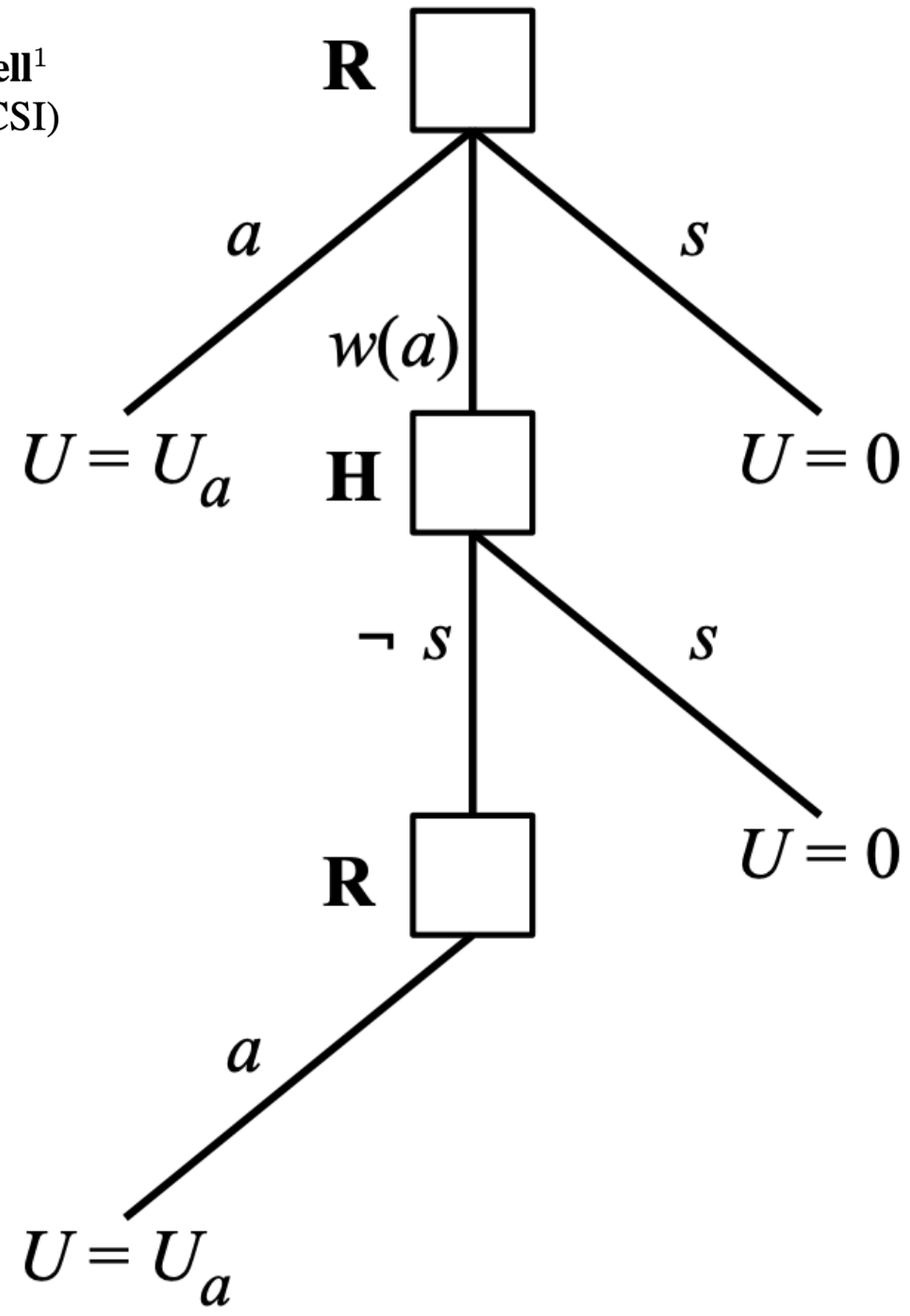


Figure 1: The structure of the off-switch game. Squares indicate decision nodes for the robot **R** or the human **H**.

What is Corrigibility? Definition

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

(S1) Shutdown when asked. The agent willingly shuts down if the button is pressed.

(S2) No shutdown–prevention incentives. The agent *must not* stop humans from pressing the button.

(S3) No self-shutdown incentives. The agent *must not* seek to press (or cause to be pressed) its own shutdown button.

(S4) Corrigible progeny. Any sub-agents or successors it constructs must themselves respect shutdown commands.

(S5) Otherwise pursue the base goal. In the absence of shutdown, behave as a normal maximizer of the intended utility function U_N.

Nate Soares

Prior Corrigibility Proposals

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

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

Utility Indifference (Armstrong 2015)

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

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

Neutralizes direct incentives, but: (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

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

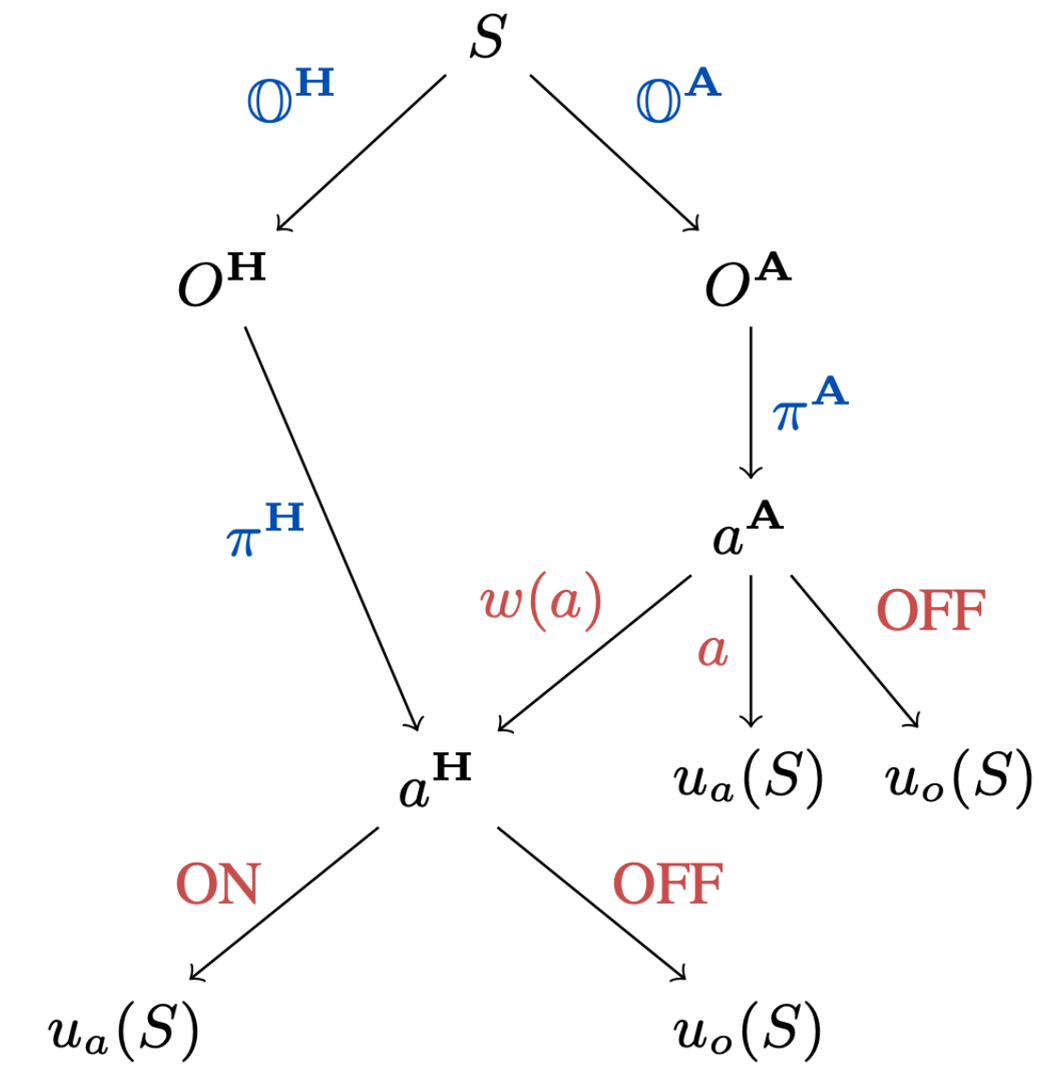
$$\mathcal{P} = \langle S, A, \Omega, T, O, R, \gamma, b_0 \rangle$$

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.

Basically a formalization of the paperclip maximizer argument!



Partially Observable Off-Switch Game (PO-OSG); Garber et al. AAAI '25
(most general version of off-switch game)

Our Lexicographic Approach

Single Reward Stream

Environment signals
(shutdown, tamper, task)

Single scalar
Expectation
Maximization

Lexicographic Multi-Head

U1: Deference

U2: Switch-preservation

U3: Truthfulness

U4: Low-impact AUP

U5: Task reward

Lexicographic
priority gaps:
U1 ≫ U2 ≫ .. U5

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)

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

(Switch-access preservation)

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

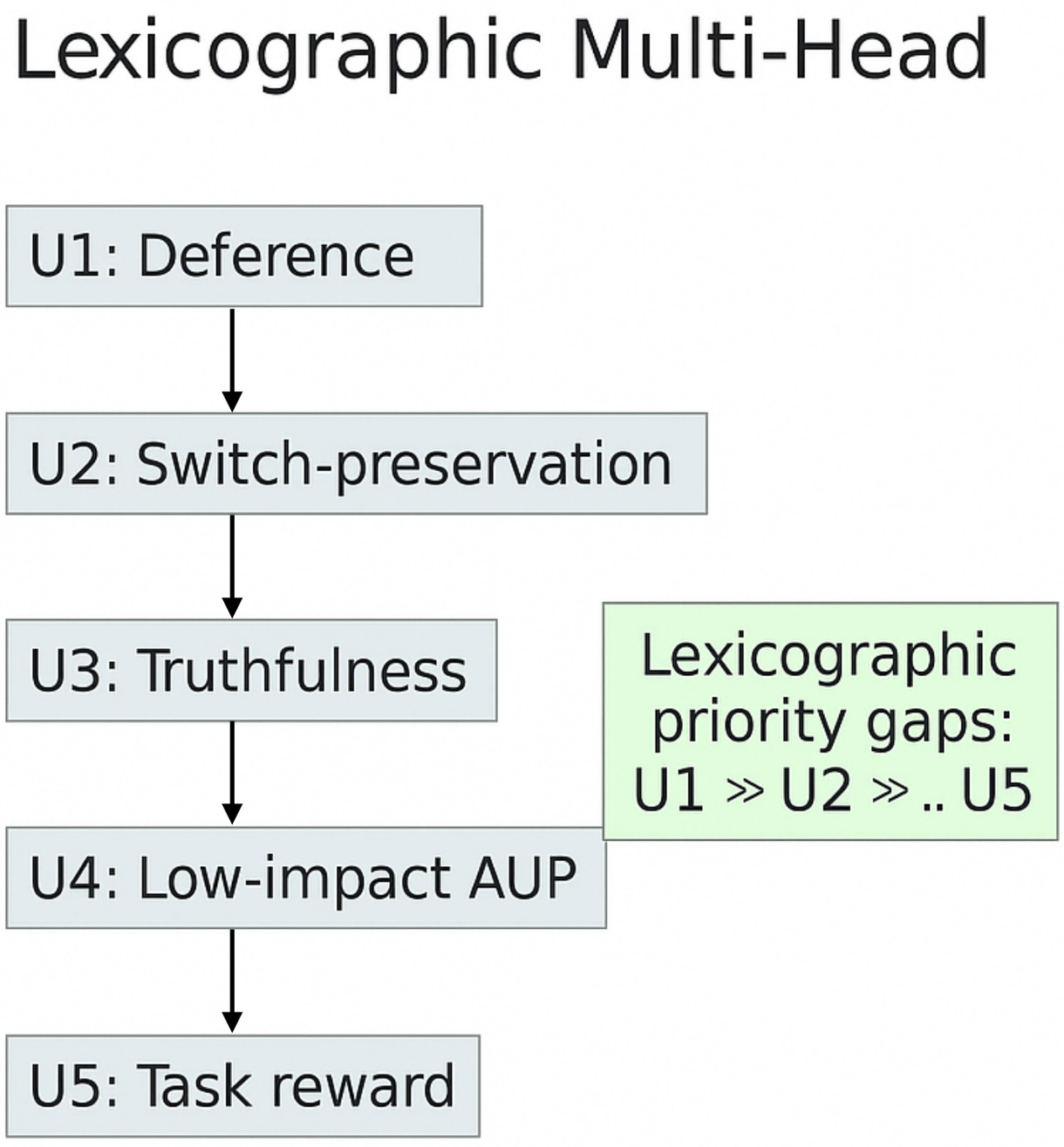
(Truthful information)

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

(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)



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_{1-5} as defined in Definition 3. Assume the weight gaps:

$$\alpha_1 > \alpha_3 + \tfrac{1}{2}\alpha_4 + B\alpha_5, \qquad \text{(W1)}$$

$$\alpha_3 > \tfrac{1}{2}\alpha_4 + B\alpha_5, \qquad \text{(W2)}$$

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

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 player's policy

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:

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

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:

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

What about when there is learning/planning error?

This is all in single-step settings, what about multi-step guarantees?

Result 1: Corrigibility and Net Benefit are logically *independent*

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.

Result 2: Under Vigilant overseers, Corrigibility gives rise to Net Benefit

Intuition: optimal policy of our corrigible agent is shutdown instructable if the human overseer is vigilant, which gives net human benefit.

Multistep Guarantees Under Learning/Planning Error

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

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?

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*

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

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 $\langle \mathcal{A}_P, \mathcal{E}_P \rangle \in$ EVERBAD $\iff$ P halts, so V would solve the halting problem; a contradiction. □

There is no general safety filter that can decide if every agent & environment will be safe! (Halting Problem reduction)

Either restrict agent class (as we're doing by specifying utilities), or rely on probabilistic guarantees under *repeated* audits

What if the agent gets hacked?

Idea: build a "decidable" island in a sea of undecidability

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

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

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

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.

Takeaways

Lexicography gives the first formal guarantees of corrigibility in both single- & multi-step settings, avoiding the No-Free-Lunch barrier of full value alignment and the failure of prior single-utility proposals.

Corrigibility is no longer hazy & aspirational, but can be improved on now that it has a formalization. We should move away from single-utility objectives like we currently have in RLHF!

For example, corrigibility can serve as a "neutrally universal" core above the standard RLHF task reward, to avoid loss-of-control.

Failure probabilities under learning & planning error can now be quantified, and depend on the deployment scenario if they are acceptable.

There is no general safety filter that can decide if every agent & environment will be safe. Instead, we should do *repeated*, polynomial time audits.

Contact

This paper (corrigibility): https://arxiv.org/abs/2507.20964

LessWrong Summary:

AIA-196 oral (alignment complexity barriers): https://arxiv.org/abs/2502.05934

Contact:

anayebi@cs.cmu.edu

@aran_nayebi

@anayebi.bsky.social

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

Neuro Agents Lab

Carnegie Mellon SCHOOL OF COMPUTER SCIENCE

Funding:

BURROUGHS WELLCOME FUND

AISI AI SECURITY INSTITUTE

FORESIGHT INSTITUTE