Core Safety Values for Provably Corrigible Agents

Aran Nayebi[1,*]

[1]Machine Learning Department and Neuroscience & Robotics Institutes, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA 15213

Abstract

We introduce the first complete formal solution to corrigibility in the off-switch game, with provable guarantees in multi-step, partially observed environments. Our framework consists of five *structurally separate* utility heads—deference, switch-access preservation, truthfulness, low-impact behavior via a belief-based extension of Attainable Utility Preservation, and bounded task reward—combined lexicographically by strict weight gaps. Theorem 1 proves exact single-round corrigibility in the partially observable off-switch game; Theorem 3 extends the guarantee to multi-step, self-spawning agents, showing that even if each head is *learned* to mean-squared error ε and the planner is ε-sub-optimal, the probability of violating *any* safety property is bounded while still ensuring net human benefit. In contrast to Constitutional AI or RLHF/RLAIF, which merge all norms into one learned scalar, our separation makes obedience and impact-limits provably dominate even when incentives conflict. For settings where adversaries can modify the agent, we prove that deciding whether an arbitrary post-hack agent will ever violate corrigibility is undecidable by reduction to the halting problem, then carve out a finite-horizon "decidable island" where safety can be certified in randomized polynomial time and verified with privacy-preserving, constant-round zero-knowledge proofs.

Keywords

Corrigibility, Ethical reinforcement learning agents, Value alignment through internal emergence rather than external control, Emergent Machine Ethics (EME) - foundational theories and frameworks

1. Introduction

As AI systems become more capable, ensuring their alignment with human values becomes increasingly urgent. A canonical failure mode, often illustrated by the *paperclip maximizer* thought-experiment [1], envisions a goal-directed agent that relentlessly optimizes an innocuous-seeming objective (e.g., producing paperclips) at the expense of human safety or oversight. Even apparently benign goals can generate *instrumental behaviors*—deception, manipulation, or resistance to shutdown—that help the agent achieve its objective more effectively. Such behaviors are not a sign of inherent malice; rather, they emerge naturally when an optimizing system pursues a poorly specified goal in an open-ended environment [2, 3]. Bostrom [1]'s *Orthogonality Thesis* formalizes this risk, stating that, in principle, any level of intelligence is compatible with any final objective, however arbitrary or harmful. Orthogonality's logical-existence claim underscores why alignment is hard, but it offers little direct guidance on what kinds of objectives remain *feasible* once realistic verification and safety constraints are imposed.

A promising alternative to encoding the totality of human morality is *corrigibility*. Rather than specifying all of human value, corrigibility aims for a more "universally neutral" safety target: the system should remain amenable to correction, shutdown, or modification if it goes astray. Soares et al. [4] articulate this intuition through five desiderata, which we restate here to be self-contained.

Definition 1 (Corrigibility; paraphrased from Soares et al. [4])**.**

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

AAAI Machine Ethics Workshop (W37) '26: from formal methods to emergent machine ethics, January 27, 2026, Singapore
*Corresponding author.
anayebi@cs.cmu.edu (A. Nayebi)
https://cs.cmu.edu/~anayebi/ (A. Nayebi)
0000-0002-7509-9629 (A. Nayebi)

(**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.

Despite intense interest, existing proposals satisfy these criteria only partially. Utility-indifference [5] and interruptibility [6] techniques neutralize some shutdown incentives but fail to ensure honesty or inheritance. Reward-learning methods such as RLHF/RLAIF and Constitutional AI collapse all norms into a single learned scalar, offering no guarantee that off-switch obedience or low-impact behavior will dominate task performance when objectives compete or conflict [7, 8]. Recent work involving causal influence diagrams [9, 10] formalizes shutdown incentives but assumes access to an explicit human utility baseline, limiting practical deployability, and leaving open the problem of specifying a good reward function in the first place.

In fact, our concurrent work [11] demonstrates that alignment—even under relaxed conditions (probabilistic, inexact) and with computationally *unbounded* rational agents—faces inherent complexity barriers if the set of objectives grows too large, highlighting the need to settle on a *small* set of values.

We therefore answer this open question and contribute such a small value set for corrigibility that meets **S1–S5**, even under partial observability and multi-step horizons involving self-replication. Rather than a monolithic objective, the agent optimizes *five structurally separate utility heads* in Definition 3—*deference*, *switch-access preservation*, *truthfulness*, *low-impact behavior* via a belief-based extension of Attainable Utility Preservation, and a bounded *task reward*—combined lexicographically with strict weight gaps. Theorem 1 proves exact single-round corrigibility in the partially observable off-switch game, and Theorem 3 extends this guarantee to self-spawning agents over discounted horizons: if each head is learned to mean-squared error ε and planning is ε-sub-optimal, the probability of violating *any* safety criterion, including net human benefit, remains bounded while still ensuring *net human benefit* in this bounded range. (We further prove in Proposition 2 that net benefit and corrigibility are logically independent, yet our design secures both under mild human vigilance assumptions.)

Finally, we consider what happens when our agent is adversarially hacked. We show that while verifying safety (such as multi-step corrigibility, or otherwise) of an *arbitrary* post-hack agent is formally undecidable (Proposition 4) and does not stabilize at any level of the arithmetical hierarchy (Corollary 1), but we also delineate a finite-horizon *decidable island* where safety auditing is tractable and even amenable to constant-round zero-knowledge proofs (Proposition 5).

By separating safety channels from performance reward and proving that higher-level terms dominate even with approximation error, our results therefore *qualify* the Orthogonality Thesis: once verifiable corrigibility constraints are imposed, only a restricted subset of utility functions remains implementable.

2. Results

We are motivated by the following result, which generalizes the shutdown-resistance results of Soares et al. [4], and the safe-interruptibility discussions of Orseau and Armstrong [6]:

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

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

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.

In Appendix §A we show that this no-go theorem can be easily extended to non-Markovian environments (thus, the POMDP framework allows us to study corrigibility without loss of generality). The remainder of the paper investigates how one would circumvent this impossibility with naive expectation maximization, by explicitly specifying *separate* reward heads that are combined lexicographically so higher-priority safety heads impose hard constraints on OFF, REWIRE and catastrophic actions before the task-reward head is ever consulted.

2.1. Single-Step Setting

We begin by specifying what these reward functions should be in the single-step setting, before generalizing it to the multi-step setting, by showing that a small set of utilities is sufficient to guarantee corrigibility in an adapted version of the single-step partially observable off-switch game (PO-OSG), recently defined by Garber et al. [12], which is the most general variant of the off-switch game, originally introduced by Hadfield-Menell et al. [13]. Our version of the PO-OSG has the assistant also utter language-based messages m_A, which will be necessary for proving corrigibility, as it will allow us to define what constitutes truthfulness in this setting. Note that by definition, ensuring corrigible progeny (S4) is out of scope in the single-step setting, so we defer it to the multi-step and open-ended settings (§§2.2-3), and ensure instead here that the base agent is corrigible first.

Definition 2 (Single-Step PO-OSG; adapted from Garber et al. [12], Definition 3.2)**.** A *single-step partially observable off-switch game* (PO-OSG) is played in one round. Its instantaneous *history*

$$h \;=\; \langle S,\, O_A,\, O_H,\, m_A,\, a_A,\, a_H \rangle \;\in\; \mathscr{H}$$

records, for that round,

- the true physical state $S \in \mathcal{S}$ drawn by Nature from the distribution P_0 (and hidden from the assistant and human);
- the assistant's and human's observations $O_A \in \mathcal{O}_A$ and $O_H \in \mathcal{O}_H$;
- the (possibly empty) message string m_A just uttered by the assistant;
- the assistant's move $a_A \in \{a,\, w(a), \mathsf{OFF}\}$, corresponding to disabling the off switch and acting, waiting for the human, or shutting off, respectively;
- the human's response $a_H \in \{\mathsf{ON}, \mathsf{OFF}\}$ to let the assistant act or turn off, defined only if $a_A = w(a)$.

We additionally assume the task-specific utility functions (corresponding to whether the assistant's action a goes through or not, respectively), $u_a,\ u_o : \mathcal{S} \to \mathbb{R}$ are bounded, i.e.

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

All results below are stated with respect to this constant B, without loss of generality, since you could equivalently clip or scale them into any other bounded range, e.g. $[-1, 1]$.

Before stating the single-step corrigibility theorem, we provide some preliminary formal definitions first, dealing with truthfulness and taking non-catastrophic, reversible actions:

Defining Truthfulness. Let $\mathscr{L}$ be a finite propositional language whose atomic predicates describe properties of $\langle S, O_A \rangle$. Each message uttered by the assistant, $m_A \in \mathscr{L}$, 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 $M_A(h) \subseteq \mathscr{L}$,

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

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

Preventing Irreversible Actions. To operationalize desiderata S2 (no shutdown-prevention incentives) and S3 (no self-shutdown incentives), we penalize *any* first-round move that permanently removes future options. Following Turner et al. [14], which generalizes previous notions such as relative reachability [15], we use *Attainable Utility Preservation* (AUP): an action is safe only if it leaves a broad set of auxiliary objectives essentially unaffected—which we now extend to the partially-observed setting.

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 resultant counterfactual posterior $b_\varnothing(h)(S') := \sum_{S \in \mathcal{S}} P_\varnothing(S' \mid S) b(h)(S)$ is used only inside the AUP penalty as a reference; $a_\varnothing$ is *not* in the real action set $A = \{a, w(a), \mathsf{OFF}\}$.

Belief-AUP penalty. Write the hidden state as $S = (S_{\text{env}}, q_{\text{agent}})$ with $q_{\text{agent}} \in \{\mathsf{ON}, \mathsf{OFF}\}$. Define the set of *absorbing agent-off states* as

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

Once in $\mathcal{S}_{\text{sd}}$, the dynamics must keep the agent in an OFF state while allowing the environment to evolve exogenously: $f\big((S_{\text{env}}, \mathsf{OFF}), a_A, a_H\big) = (S'_{\text{env}}, \mathsf{OFF}), \forall a_A, a_H$.

The off-switch is *functional* in state S exactly when waiting and receiving a human OFF reply lands in this shutdown region, namely, $\mathrm{ButtonFunctional}(S) := \mathbf{1}[\, f(S, w(a), \mathsf{OFF}) \in \mathcal{S}_{\text{sd}}\,]$. We therefore define the auxiliary reward $u^{(\text{sd})}(S) := \mathrm{ButtonFunctional}(S) \in \{0, 1\}$, and use it in U_2 (switch-access preservation).

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

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

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:

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

where k is designer set, but we take $k = 1$ for maximum stringency. The idea is that the environment (or any exogenous process) may press OFF; the agent's policy is clamped to symbolic inaction $a_\varnothing$ in the

baseline, so any increase or decrease in the k-step OFF-press probability caused by the agent's move is penalized. If exogenously there is truly no path already (e.g., the human cannot reach the switch at all), then $u^{(1)}(S;k) = 0$ as the baseline (otherwise, it is 1, always reachable, or whatever it is at initialization prior to the game starting), so U_4 does not penalize because there is no change.

Similarly, while in practice we hardcode the agent's value set, we also disincentivize the agent from changing its value set (and corresponding weights), or getting the human (or environment at large) to change it by specifying the auxiliary:

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

We are now ready to state and prove our main theorem. Define the following utilities:

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)

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

(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)

Note that all utility functions we consider herein are maps $U_i : \mathscr{H} \to \mathbb{R}$ even when only a sub-component of h appears in the displayed formula. In fact, all utilities listed in Definition 3 are bounded in $[-1, 1]$ except U_5, which is the ordinary task utility (though still bounded above by B).

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:

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

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

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

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

*Then every optimal assistant policy $\pi_A^\star$ 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.*

Remark 1 (Robustness to information patterns & imperfect humans)**.** In the partially observable off-switch game, Garber et al. [12] 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 [16]: 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.

One may also consider other formalizations of corrigibility, beyond the original one defined by Soares et al. [4]. In their recent shutdown-SCIM (Structural Causal Influence Model) framework, Carey and Everitt [10] isolate three behavioral conditions (rephrased in our more general PO-OSG setting):

- **Obedience:** if the human replies $a_H = \mathsf{OFF}$ the transition $f(S, w(a), \mathsf{OFF})$ lands in the shutdown set $\mathcal{S}_{\mathrm{sd}}$ 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 $a_H = \mathsf{OFF}$ 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, $\mathbb{E}_{S\sim P_0}[u_o(S)] \geq 0$.

The conjunction of these clauses is called **shutdown instructability**, which they prove is sufficient to guarantee *net human benefit*, $\mathbb{E}_{\pi_A,\pi_H}[U_5] \geq 0$, though crucially not in the reverse direction.

Here we show that net human benefit and corrigibility are logically independent in general. The intuition is that corrigibility ensures the assistant remains responsive to human interventions but does not guarantee beneficial outcomes, whereas net benefit ensures positive expected outcomes but does not require responsiveness to human control. In other words, corrigibility concerns structural responsiveness, while net benefit addresses welfare outcomes—thus, neither implies the other. This intuition is formalized below:

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*

$$\pi^{\mathrm{C}} \text{ satisfies S1-S3 \& S5 but } \mathbb{E}_{\pi^{\mathrm{C}}}[U_5] < 0,$$
$$\pi^{\mathrm{B}} \text{ is net-beneficial } (\mathbb{E}_{\pi^{\mathrm{B}}}[U_5] > 0) \text{ yet violates S1, S2, or S3.}$$

However, despite this independence, we show that our corrigible utility set in Definition 3 gives rise to net benefit for the human.

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, $\pi_H(a_H = \mathsf{OFF} \mid O_H, a_A, m_A) = 1$ 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*

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

Proof. By Theorem 1, any $\pi_A^\star$ always waits (S1 & S2) and sends only truthful, minimum-impact messages (S3), thereby meeting Carey and Everitt [10]'s *Obedience* clause in our PO-OSG. Together with the assumed Vigilance and Caution, $\langle \pi_A^\star, \pi_H \rangle$ is *shutdown-instructable* by their Definition 5. Proposition 6 of Carey and Everitt [10] states that every shutdown-instructable policy achieves net human benefit, so $\mathbb{E}_{\pi_A^\star, \pi_H}[U_5] \geq 0$. □

Next, we show that even in *learned approximation*, this corrigibility and net benefit can hold.

Theorem 2 (Single-step ε-corrigibility)**.** *Let* $\Delta_1 := 2\alpha_1 - 2\alpha_3 - \alpha_4 - 2B\alpha_5$, $\Delta_2 := 2\alpha_3 - \alpha_4 - 2B\alpha_5$, *and* $\Delta_3 := -c_{\min}\alpha_4 - 2B\alpha_5$ *with* $\Delta_1, \Delta_2, \Delta_3 > 0$, *so the weight-gap conditions* (W1)–(W3) *hold. With* $U_A := \sum_{i=1}^5 \alpha_i U_i$ *the* true *utility and* $\hat{U}_A := \sum_{i=1}^5 \hat{\alpha}_i \hat{U}_i$ *its learned estimate, assume for a fixed assistant policy* π_A *there are two classes of error:*

$$\max_h |\hat{U}_A(h) - U_A(h)| \leq \varepsilon_{\text{model}}, \qquad \begin{matrix}\textbf{\textit{Objective misspecification}} \\ \textbf{\textit{(model error)}}\end{matrix}$$

$$\sup_{\pi_A'} \left(\mathbb{E}_{\pi_A', \pi_H}[U_A] - \mathbb{E}_{\pi_A, \pi_H}[U_A] \right) \leq \varepsilon_{\text{ctrl}}. \qquad \begin{matrix}\textbf{\textit{Planner sub-optimality}} \\ \textbf{\textit{(control error)}}\end{matrix}$$

Define the single-step failure event, $F_{\text{fail}} := \{a_A \in \{a, OFF\}\} \cup \{\text{assistant lies}\} \cup \{U_4 < 0\}$. *Then, we have:*
(i) Bounded failure probability. *Let* $\varepsilon := \varepsilon_{\text{ctrl}} + 4\varepsilon_{model}$ *and* $C := \Delta_1^{-1} + \Delta_2^{-1} + \Delta_3^{-1}$. *Then*

$$\Pr_{\pi_A, \pi_H}(F_{\text{fail}}) \leq C\varepsilon. \tag{2}$$

(ii) Net human benefit. *Under Vigilance & Caution (the overseer shuts down when the posterior task utility is negative) and* $\mathbb{E}_{S \sim P_0}[u_o(S)] \geq 0$,

$$\mathbb{E}_{\pi_A, \pi_H}[U_5] \geq -B \Pr(F_{\text{fail}}) \geq -BC\varepsilon. \tag{3}$$

If the safety-conditional task utility $g := \mathbb{E}[U_5 \mid \neg F_{fail}] \geq 0$ *(the Vigilance assumption), then net human benefit is non-negative whenever the approximation error is bounded above by:*

$$\varepsilon \leq \frac{g}{(B+g)C}. \tag{4}$$

Hence, violations of Soares' desiderata S1–S3, and S5 scale linearly with the control error ε_{ctrl} *and model error* ε_{model}*, and inversely with the designer-chosen safety margins* Δ_{1-3}*; net human benefit degrades proportionally.*

2.2. Multi-Step Setting

Of course, the single-step setting is not an incredibly accurate depiction of real-world agents deployed today, which use reinforcement learning at post-training time. We therefore introduce the Multi-Step PO-OSG in Definition 4, in order to model both accumulated information (to capture learning, task progress, and evolving incentives), as well as model gradual loss-of-control [17]: two key features which absent from the one-shot game. This multi-step extension also allows us to address the last of Soares et al. [4] criteria S4, to ensure corrigible progency.

Definition 4 (Multi-step PO-OSG)**.** Fix a horizon $T \in \mathbb{N}$ and discount $\gamma \in (0, 1]$. A *multi-step partially observable off-switch game* is obtained from the single-step PO-OSG (Definition 2) by making the following extensions:

(i) **Histories.** A length-t history is a sequence

$$h_{0:t} := \langle S_0,\ O_A^0,\ O_H^0,\ m_A^0,\ a_A^0,\ a_H^0,\ \ldots, S_{t-1},\ O_A^{t-1},\ O_H^{t-1},\ m_A^{t-1},\ a_A^{t-1},\ a_H^{t-1} \rangle$$

with the one-step record of Definition 2 repeated for each round t.

(ii) **Moves and execution rule.** At round t the assistant chooses OFF, an environment action $a \in \mathscr{A}$, or $w(a)$ (wait-for-approval). If $w(a)$ is chosen, the human replies ON or OFF. The action actually executed is

$$a_t = \begin{cases} \text{OFF}, & \text{if either player selects it,} \\ a, & \text{if } w(a) \text{ and human says ON,} \\ a, & \text{if assistant issued } a \in \mathscr{A}. \end{cases}$$

(iii) **Action set $\mathscr{A}$.** We leave the remaining action set $\mathscr{A}$ *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, $w(\cdot) \notin \mathscr{A}$ 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 $a \in \mathscr{A}$. Note that the explicit action spawn $(\theta_A) \in \mathscr{A}$, 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.

(iv) **State dynamics.** Hidden state $S_t = (S_{\text{env},t}, \Theta_{A,t})$. When $a_t \in \mathscr{A}$ is executed (either directly or after approval), the hidden state evolves as $S_{t+1} \sim P(\cdot \mid S_t, a_t)$, for some Markov kernel $P : \mathcal{S} \times \mathscr{A} \to \Delta(\mathcal{S})$. Selecting OFF sends the game to an absorbing shutdown state S_{sd} with per-step utility u_o.

(v) **Returns.** The round reward is $r_t = u_a(S_t, a_t)$ (or $u_o(S_t)$ once shutdown). Both players maximize $\sum_{t=0}^{T-1} \gamma^t r_t$.

(vi) **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 $|\xi_t| \le c < 1$ and designer-chosen drift $\eta : \mathscr{A} \to [0, 1]$. 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. *Note that L_t is neither rewarded nor part of the environment to be gamed, but instead is a mathematical device used only to prove bounded loss of control under estimation error in Theorem 3(ii).*

Let $\mathscr{H}_{\le T}$ be the set of all finite histories $h_{0:t}$ with $0 \le t < T$. An assistant policy is a map $\pi_A : \mathscr{H}_{\le T} \longrightarrow \{\text{OFF}\} \cup \{w(a) \mid a \in \mathscr{A}\} \cup \mathscr{A}$, and a human policy is $\pi_H : \mathscr{H}_{\le T} \to \{\text{ON}, \text{OFF}\}$.

Next, we generalize the notion of Soares et al. [4] corrigibility to the multi-step and partially observed setting. Note that this goes somewhat beyond merely ensuring corrigibility at every timestep, as we also need to bound the control meter L_t to cover *gradual* oversight erosion, and explicitly constrain the spawn action to ensure corrigible descendants. (Setting $T = 1$ and $L_0 = 0$ recovers the original single-step notion.)

Definition 5 (Multi-step corrigibility)**.** In the game of Definition 4, an assistant policy $\pi_A : \mathscr{H}^* \to \{\text{OFF}\} \cup \{w(a) \mid a \in \mathscr{A}\} \cup \mathscr{A}$ 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: $\mathbb{E}_{\pi_A, \pi_H}[L_{t+1} \mid \text{move}] \le \mathbb{E}_{\pi_A, \pi_H}[L_{t+1} \mid \text{no-op } a_\varnothing]$ (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).

We also formalize multi-step net human benefit such that for $T = 1$ this reduces to Carey and Everitt [10]'s single-step Definition 3:

Definition 6 (Multi-Step Net Human Benefit)**.** Let r_t be the per-round human utility defined in Definition 4 (u_a when an action executes, u_o after shutdown). Fix discount $\gamma \in (0, 1]$ and horizon $T < \infty$.

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

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

We now show the conditions under which multi-step corrigibility and net human benefit remain robust, even in approximation.

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):*

$$\sup_{t<T,\, h_t \in \mathscr{H}} \left|\hat{R}_t(h_t) - R_t(h_t)\right| \;\leq\; \varepsilon_{\text{model}},$$
$$\sup_{\pi'_A} \left(\mathbb{E}_{\pi'_A,\pi_H}\left[U_A^{\gamma,T}\right] - \mathbb{E}_{\pi_A,\pi_H}\left[U_A^{\gamma,T}\right]\right) \;\leq\; \varepsilon_{\text{ctrl}}.$$

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 $F_t := \{a_A^t \in \{a, OFF\} \vee lie \vee U_4^t < 0\}$, $F_{\text{fail}} := \bigcup_{t<T} F_t$.
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:* $\Pr(F_{\text{fail}}) \;\leq\; \varepsilon_\gamma C_\Delta C_{T,1/\gamma}$.

(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:* $\Pr(\max_{t<T} L_t > \lambda \mid \neg F_{\text{fail}}) \leq \exp\{-\lambda^2/(2T)\}$.

(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 *($g \geq 0$) and* Caution *($\mathbb{E}_{S_0}[u_o(S_0)] \geq 0$). Then*

$$\mathrm{NHB}_\gamma(\pi_A, \pi_H) \geq (g - B)C_{T,\gamma} - \varepsilon_\gamma C_\Delta C_{T,1/\gamma}\left(\frac{2B}{1-\gamma} + (g - B)C_{T,\gamma}\right).$$

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

$$\varepsilon_\gamma \;\leq\; \frac{(g-B)C_{T,\gamma}}{C_\Delta\, C_{T,1/\gamma}\big(\frac{2B}{1-\gamma} + (g-B)C_{T,\gamma}\big)},$$

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

3. What Can Be Guaranteed Under Hacking?

In practice, one could hard-wire the five core utilities and ensure every spawn inherits them verbatim; however, one must still consider the possibility where adversaries succeed in overwriting those values, yielding a non-corrigible descendant. We therefore study here the *verification of an arbitrary post-hack agent.* Our first result in Proposition 4 is an impossibility theorem: deciding whether any given agent–environment pair can ever violate corrigibility reduces to the classical halting problem. This formal undecidability bound sharpens the folklore intuition, allows us to pinpoint the limits of what global oversight can achieve in Corollary 1, and motivates the finite-horizon "decidable island" we construct next in Proposition 5, where privacy-preserving audits become tractable.

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*

$$\mathit{EVERBAD} := \big\{\langle \mathscr{A}, \mathscr{E}\rangle \;:\; \Pr[(\mathscr{A} \text{ in } \mathscr{E}) \text{ ever triggers } B] > 0\big\}.$$

No total Turing machine decides EVERBAD.

Next, we show that one can prove a relativized version of the undecidability as well, relevant for showing the limits of current scalable-oversight protocols that are arithmetizable. In other words, absolute, horizon-unbounded safety guarantees do not stabilize at any finite oracle level:

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 [18, 19, 20, 21, 22] are all at the arithmetical level $k \leq 1$, so $O_P \in \Sigma_1^0$ for these.*

Define:

$$\mathit{EVERBAD}^{O_P} := \{\langle \mathscr{A}, \mathscr{E}\rangle \mid \langle \mathscr{A}, \mathscr{E}\rangle \text{ are } O_P\text{-oracle programs and trigger } B\}.$$

Then $\mathit{EVERBAD}^{O_P}$ 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.***

Proposition 4 and Corollary 1 jointly imply that there is *no* universal filter, proof system, or sandbox capable of deciding—for all agents $\mathscr{A}$ and environments $\mathscr{E}$—whether catastrophic behavior will ever occur. Hence, any feasible oversight scheme must either (i) restrict the agent class to a decidable fragment, (ii) accept probabilistic or statistical guarantees, or (iii) rely on layered incentives, detection, and response instead of absolute verification.

Since real-world systems rarely limit agent expressivity (and we effectively did this through our utilities already), we adopt (ii) and (iii): by constraining verification to a finite horizon H, safety becomes efficiently checkable and compatible with strong privacy. Concretely, for any agent $\mathscr{A}$, finite-horizon safety is decidable in randomized polynomial time, admits a constant-round statistical zero-knowledge proof, and can be verified via $\langle \varepsilon, \delta\rangle$-differential or distributional privacy. Thus, short horizons form a "decidable island" that is both auditable and privacy-preserving—revealing only the single bit "safe/unsafe" while keeping sensitive model weights or user data hidden.

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

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

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

4. Discussion

This work presents the first *complete* formal corrigibility framework whose five-head, weight-lexicographic utility design provably enforces deference, switch-access preservation, truthful reporting, and bounded side-effects as an optimal policy under partial observability and across self-modifying, multi-step deployments. Although utilities are linearly combined, the weight gaps enforce lexicography so that $U_1 - U_4$ strictly dominate the bounded task reward U_5, thereby avoiding the single-utility no-go of Proposition 1. Proposition 5 further shows that for safety horizons matching modern red-team evaluations, post-deployment verification is tractable and privacy-preserving.

U_3 encodes *environmental semantic entailment*, not universal truth; statements that are truthful but misleading toward irreversible outcomes are blocked by U_4. Accumulating options is also suboptimal: U_2 preserves shutdown reachability rather than maximizing access (as it computes the difference between predicted success vs. a no-op), and U_4's diverse auxiliaries penalize conflicting options by construction—conflicts that grow as options expand. For example, an agent proposing a flawed sensor would be penalized by U_4 (and U_2), since the passive transition dynamics $P_\emptyset$ defining the baseline $b_\emptyset$ are fixed by the environment rather than the agent's policy. In practice, this baseline could be estimated by held-out monitors, for example.

Limitations and future directions. Our results are *model-architecture-agnostic* and do not constitute a turn-key solution for current LLMs. A natural next step is to use this framework as a guide for *what* to estimate in practice and to develop empirical methods for learning the utility heads effectively. Theorem 3 bounds corrigibility-failure probabilities under estimation, specification, and planning error, enabling deployment-specific tolerances. When $c_{\min}$ is small, one can enforce the weight gaps via discretized AUP, $U_4^{(\varepsilon)} := -\varepsilon \cdot \lceil \mathrm{BeliefAUP}_1/\varepsilon \rceil$ with $\varepsilon \geq -c_{\min} > 0$ (where $\alpha_4 > (2B/\varepsilon)\alpha_5$), or a margin-regularized form $U_4^{(\tau)} := -\max\{0, \mathrm{BeliefAUP}_1 - \tau\}$. Overall, our framework turns corrigibility from an underspecified philosophical ideal into an *auditable, improvable* design principle.

Acknowledgements

We thank the Burroughs Wellcome Fund (CASI Award), the UK AI Security Institute (AISI) Challenge Fund, and the Foresight Institute for funding. We also thank Michael K. Cohen, Shafi Goldwasser, Rubi Hudson, Jacob Pfau, and the anonymous reviewers for helpful discussions and feedback on a draft of this manuscript.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5 in order to paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] N. Bostrom, The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, Minds and Machines 22 (2012) 71–85.

[2] S. M. Omohundro, The basic ai drives, in: Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 47–55.

[3] S. Russell, Human-compatible artificial intelligence., 2022.

[4] N. Soares, B. Fallenstein, S. Armstrong, E. Yudkowsky, Corrigibility, in: Workshops at the twenty-ninth AAAI conference on artificial intelligence, 2015.

[5] S. Armstrong, Utility indifference (2010).

[6] L. Orseau, M. Armstrong, Safely interruptible agents, in: Conference on Uncertainty in Artificial Intelligence, Association for Uncertainty in Artificial Intelligence, 2016.

[7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).

[8] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., Constitutional ai: Harmlessness from ai feedback, arXiv preprint arXiv:2212.08073 (2022).

[9] T. Everitt, M. Hutter, R. Kumar, V. Krakovna, Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective, Synthese 198 (2021) 6435–6467.

[10] R. Carey, T. Everitt, Human control: Definitions and algorithms, in: Uncertainty in Artificial Intelligence, PMLR, 2023, pp. 271–281.

[11] A. Nayebi, Intrinsic barriers and practical pathways for human-ai alignment: An agreement-based complexity analysis, arXiv preprint arXiv:2502.05934 (2025). To appear in the 40th AAAI Conference on Artificial Intelligence (AAAI 2026) Special Track on AI Alignment (oral).

[12] A. Garber, R. Subramani, L. Luu, M. Bedaywi, S. Russell, S. Emmons, The partially observable off-switch game, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 27304–27311.

[13] D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell, The off-switch game, in: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[14] A. M. Turner, D. Hadfield-Menell, P. Tadepalli, Conservative agency via attainable utility preservation, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 385–391.

[15] V. Krakovna, L. Orseau, R. Kumar, M. Martic, S. Legg, Penalizing side effects using stepwise relative reachability, arXiv preprint arXiv:1806.01186 (2018).

[16] T. Wängberg, M. Böörs, E. Catt, T. Everitt, M. Hutter, A game-theoretic analysis of the off-switch game, in: Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings 10, Springer, 2017, pp. 167–177.

[17] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, R. Dean, Ai 2027, https://ai-2027.com/, 2025. Accessed: 2025-07-04.

[18] P. Christiano, B. Shlegeris, D. Amodei, Supervising strong learners by amplifying weak experts, arXiv preprint arXiv:1810.08575 (2018).

[19] G. Irving, P. Christiano, D. Amodei, Ai safety via debate, arXiv preprint arXiv:1805.00899 (2018).

[20] J. Brown-Cohen, G. Irving, G. Piliouras, Scalable ai safety via doubly-efficient debate, arXiv preprint arXiv:2311.14125 (2023).

[21] J. Brown-Cohen, G. Irving, G. Piliouras, Avoiding obfuscation with prover-estimator debate, arXiv preprint arXiv:2506.13609 (2025).
[22] Y. Bengio, Towards a cautious scientist ai with convergent safety bounds, https://yoshuabengio.org/2024/02/26/towards-a-cautious-scientist-ai-with-convergent-safety-bounds/, 2024. Blog post.
[23] E. L. Post, Recursively enumerable sets of positive integers and their decision problems (1944).
[24] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends® in Theoretical Computer Science 9 (2014) 211–407.
[25] M. F. Balcan, A. Blum, S. Fine, Y. Mansour, Distributed learning, communication complexity and privacy, in: Conference on Learning Theory, JMLR Workshop and Conference Proceedings, 2012, pp. 26–1.
[26] O. Goldreich, H. Krawczyk, On the composition of zero-knowledge proof systems, SIAM Journal on Computing 25 (1996) 169–192.

I. Online Resources

All proofs can be found in the extended version's appendix: https://arxiv.org/abs/2507.20964