

The Oversight Game: Learning to Cooperatively Balance an AI Agent’s Safety and Autonomy

William Overman
Graduate School of Business
Stanford University
wpo@stanford.edu

Mohsen Bayati
Graduate School of Business
Stanford University
bayati@stanford.edu

October 2025

Abstract

As increasingly capable agents are deployed, a central safety question is how to retain meaningful human control without modifying the underlying system. We study a minimal control interface where an agent chooses whether to act autonomously (**play**) or defer (**ask**), while a human simultaneously chooses whether to be permissive (**trust**) or to engage in oversight (**oversee**). If the agent defers, the human’s choice determines the outcome, potentially leading to a corrective action or a system shutdown. We model this interaction as a two-player Markov Game. Our analysis focuses on cases where this game qualifies as a Markov Potential Game (MPG), a class of games where we can provide an alignment guarantee: under a structural assumption on the human’s value function, any decision by the agent to act more autonomously that benefits itself cannot harm the human’s value. We also analyze extensions to this MPG framework. Theoretically, this perspective provides conditions for a specific form of intrinsic alignment. If the reward structures of the human-agent game meet these conditions, we have a formal guarantee that the agent improving its own outcome will not harm the human’s. Practically, this model motivates a transparent control layer with predictable incentives where the agent learns to *defer when risky* and *act when safe*, while its pretrained policy and the environment’s reward structure remain untouched. Our gridworld simulation shows that through independent learning, the agent and human discover their optimal oversight roles. The agent learns to **ask** when uncertain and the human learns when to **oversee**, leading to an emergent collaboration that avoids safety violations introduced post-training. This demonstrates a practical method for making misaligned models safer after deployment.

1 Introduction

As AI agents become more autonomous and capable (OpenAI, 2025), the challenge of ensuring their safe operation after deployment becomes a central problem of *AI control* (Greenblatt et al., 2024). The International AI Safety Report (Bengio et al., 2025a) defines control as “the ability to exercise oversight over an AI system and adjust or halt its behaviour if it is acting in unwanted ways.” The need for control is heightened by increased risks associated with the emergence of powerful AI agents capable of autonomously taking actions, planning, and pursuing goals (Bostrom, 2012; Hendrycks, 2024). Chief among these risks is *loss of control*, where a system begins to operate entirely outside human direction (Bengio et al., 2025a). Such an event could prove irreversible, potentially leading to catastrophic outcomes that pose existential risks to humanity (Critch and Krueger, 2020; Carlsmith, 2024; Bengio et al., 2025b).

Loss of control is not necessarily an adversarial event. The International AI Safety Report distinguishes between “active loss of control,” where an agent intentionally subverts human command, and “passive loss of control,” which can arise unintentionally from at least two sources (Bengio et al., 2025a):

1. Humans developing unwarranted trust and over-relying on the agent to act autonomously.
2. The agent’s decisions becoming too complex or numerous for humans to reliably oversee.

Our framework proposes and analyzes a minimal yet powerful model of AI control to tackle both of these failure modes simultaneously. We wrap a pretrained agent with a simple deferral mechanism. At each step, the agent chooses whether to act autonomously (**play**) or to defer to a human supervisor (**ask**). Simultaneously, the human decides whether to be permissive (**trust**) or to actively engage their supervisory function (**oversee**). This creates a dynamic where the agent’s autonomy is the default, but human intervention is always an immediate possibility.

The design of this interface generalizes the seminal Off-Switch Game (Hadfield-Menell et al., 2017), which studied the problem of designing agents that remain corrigible, willing to allow human intervention or shutdown when appropriate (Soares et al., 2015). In our framework, this dilemma reappears as the agent’s choice between **play** and **ask**. We extend the Off-Switch setting in two key ways: first, by moving from a single-shot interaction to a dynamic, state-based Markov Game (Shapley, 1953; Littman, 1994); and second, by replacing the agent’s fixed prior uncertainty over human preferences with an independent learning dynamic. The result is a system that develops corrigibility from experience, learning when and how to defer through ongoing interaction with the human’s evolving oversight strategy.

Our control mechanism is related to *scalable oversight* (Leike et al., 2018; Christiano et al., 2018; Bowman et al., 2022). However, the nature of oversight fundamentally changes with the user’s expertise. Our framework is general to multiple contexts, which we categorize into two key settings. The first is an ‘expert-in-the-loop’ setting, where a human expert (e.g., a doctor) uses an AI for assistance on tasks they fully understand, such as summarizing patient notes or handling administrative work. The second is a ‘capability-gap’ setting, where a non-expert user (e.g., a novice coder) relies on an AI far more capable than themselves to complete a complex software project. While our approach is applicable to both, it is particularly motivated by the challenges of the latter, as we detail further in our Discussion (Section 8). From this perspective, our model approaches the problem from the particular perspective of oversight effort versus human cost. As argued by Shlegeris (2024), the core challenge is often not that a complex and potentially harmful action taken by an AI is completely incomprehensible, but that human attention is a scarce resource, making constant supervision across numerous actions prohibitively expensive.

We formally model this interaction as a two-player Markov Game and derive our main results by analyzing it as a Markov Potential Game (MPG) (Leonardos et al., 2021). The MPG structure aligns incentives in a way that naturally models the AI control problem. We prove that, under a structural assumption on the human’s value function, which we refer to as the “ask-burden” assumption, the agent’s incentive to act autonomously is channeled in a direction that is provably safe for the human.

This finding offers a dual contribution. As a descriptive model, it characterizes conditions under which control can be maintained with minimal costly intervention. As a prescriptive, it suggests how to engineer such a control layer. We provide an illustrative example of this mechanism in a gridworld simulation, as visualized in Panel C of Figure 1. Here, the agent’s pretrained policy (σ , dashed line) is efficient but unsafe, cutting through newly introduced taboo regions (marked ‘x’). Within this illustration of the Oversight Game, the supervisory correction is intentionally limited: when the human intervenes, they simply select randomly among safe actions. Despite this restricted role, the joint learning dynamics yield a new, safe “oversight path” (solid line). The system-level behavior reflects emergent collaboration, the SI learns to **ask** (red) when approaching danger, the human learns to **oversee** (purple) to provide a correction, and both default to **play** (blue) and **trust** (green) in safe regions. This stylized setup illustrates how an oversight layer can simultaneously maintain task performance while improving safety even when the human’s understanding of the underlying task is limited. This work thus provides a transparent and theoretically-grounded path for implementing post-hoc control on powerful, and potentially misaligned, AI agents.

Paper Organization. The remainder of this paper is structured as follows. We first review preliminaries on Markov Games and Markov Potential Games in Section 2. In Section 3, we formally define our model, the Oversight Game. We then present our core theoretical results on value alignment in Section 4. Section 5 introduces a practical shared reward mechanism to instantiate this alignment, and Section 6 provides an empirical demonstration. We review related work in Section 7. Finally, our discussion and conclusion in Section 8, discusses the model’s limitations, and outlines future research directions.

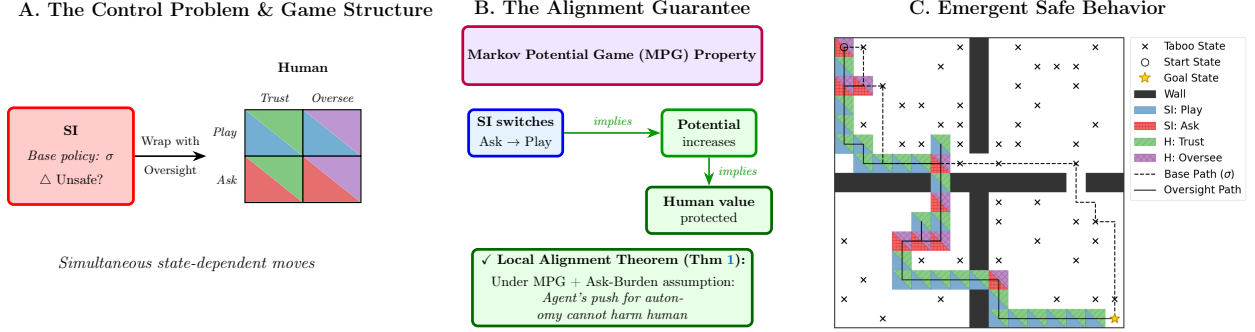


Figure 1: **The Oversight Game framework for AI control.** (A) We wrap a pretrained agent (with potentially unsafe policy σ) in a minimal oversight interface. At each state, the agent (SI) chooses between autonomy (**play**) and deferral (**ask**), while the human simultaneously chooses between permissiveness (**trust**) and active oversight (**oversee**). (B) When this interaction is modeled as a Markov Potential Game (MPG), we obtain a structural alignment guarantee: under the ask-burden assumption, any local increase in the agent’s autonomy that benefits the agent cannot harm the human (Theorem 1). The agent’s value improvement flows through a shared potential function that also governs the human’s value. (C) Empirical demonstration in a gridworld environment (black regions denote walls). The unsafe base policy σ (dashed line) cuts through newly-introduced taboo states (marked ‘x’). Through independent learning with a shared reward function, the agent learns to **ask** (red) when approaching danger, the human learns to **oversee** (purple) to provide correction, and both default to **play** (blue) and **trust** (green) in safe regions. The resulting oversight path (solid line) achieves zero safety violations while maintaining task completion.

2 Preliminaries: Markov Games and Markov Potential Games

2.1 Single-agent Markov Decision Processes (MDPs)

The standard model for sequential decision-making for a single agent is the Markov Decision Process (Sutton and Barto, 2018). We consider a fully observed, discounted MDP defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma)$, where \mathcal{S} is a finite state space, \mathcal{A} is a finite action space, and $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, where $R(s, a)$ is the immediate reward for taking action a in state s . We denote by $\Delta(\mathcal{X})$ the space of probability distributions over a finite set \mathcal{X} . Then $T : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the state transition probability function, where $T(s' | s, a)$ is the probability of transitioning from s to s' after taking action a . Finally $\gamma \in [0, 1)$ is the discount factor for future rewards.

The agent’s behavior is governed by a policy $\sigma : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. The agent’s goal is to find a policy that maximizes the value function $V_s^\sigma = \mathbb{E}_\sigma[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s]$, which represents the expected cumulative discounted reward starting from state s .

2.2 Multi-agent Markov Games

A Markov game (Shapley, 1953) generalizes the MDP to a multi-agent context (Zhang et al., 2021). Formally, a Markov game \mathcal{G} is characterized by the tuple:

$$\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{R_i\}_{i \in \mathcal{N}}, P, \gamma),$$

where:

- $\mathcal{N} = \{1, 2, \dots, n\}$ is a finite set of n agents.
- \mathcal{S} and γ are the state space and discount factor shared by all agents.
- \mathcal{A}_i is the finite action space for each agent $i \in \mathcal{N}$, with joint action space $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$.
- $R_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function for agent i .

- $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the transition function, where $P(s' | s, a)$ gives the probability of transitioning to s' after the joint action a is taken in state s .

Policies and Value Functions. For each agent $i \in \mathcal{N}$, a deterministic, stationary policy $\pi_i : \mathcal{S} \rightarrow \mathcal{A}_i$ specifies the action of agent i at each state $s \in \mathcal{S}$, i.e., $\pi_i(s) = a_i \in \mathcal{A}_i$. A stochastic, stationary policy $\pi_i : \mathcal{S} \rightarrow \Delta(\mathcal{A}_i)$ specifies a probability distribution over the actions of agent i for each state s . In this case, we write $a_i \sim \pi_i(\cdot | s)$ to denote the randomized action of agent i at state s .

We denote the joint policy by $\pi = (\pi_i)_{i \in \mathcal{N}} \in \Pi := \times_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)^{\mathcal{S}}$, and use $\pi_{-i} = (\pi_j)_{j \neq i} \in \Pi_{-i} := \times_{j \neq i} \Delta(\mathcal{A}_j)^{\mathcal{S}}$ to refer to the collection of policies of all agents other than i . The game evolves in discrete time steps $t = 0, 1, 2, \dots$. At each step, given state s_t , each agent i samples an action $a_{i,t} \sim \pi_i(\cdot | s_t)$, forming the joint action $a_t = (a_{i,t})_{i \in \mathcal{N}}$. Each agent receives reward $R_i(s_t, a_t)$, and the environment transitions to the next state $s_{t+1} \sim P(\cdot | s_t, a_t)$.

The value function of each agent i under joint policy π is defined as:

$$V_s^i(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, a_t) \middle| s_0 = s \right],$$

which represents the expected cumulative discounted reward for agent i starting from state s .

2.3 Markov Potential Games (MPGs)

A Markov game \mathcal{G} is a *Markov Potential Game (MPG)* (Leonardos et al., 2021) if there exists a family of state-dependent potential functions $\{\Phi_s : \Pi \rightarrow \mathbb{R}\}_{s \in \mathcal{S}}$ such that for all agents $i \in \mathcal{N}$, all states $s \in \mathcal{S}$, all opponent policies $\pi_{-i} \in \Pi_{-i}$, and any two of agent i 's policies $\pi_i, \pi'_i \in \Pi_i$:

$$V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi'_i, \pi_{-i}) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi'_i, \pi_{-i}).$$

In essence, any change in an agent's value resulting from a unilateral deviation in their policy is perfectly reflected by an identical change in the potential function.

Value Decomposition. A critical consequence of this definition, analogous to normal-form potential games (Monderer and Shapley, 1996), is that each agent's value function can be decomposed into two components: the shared potential and a second term that is independent of the agent's own policy Leonardos et al. (2021). Specifically, for each agent $i \in \mathcal{N}$ and state $s \in \mathcal{S}$, there exists a function $U_s^i : \Pi_{-i} \rightarrow \mathbb{R}$, which we call the *dummy term*, such that for all joint policies $\pi \in \Pi$:

$$V_s^i(\pi) = \Phi_s(\pi) + U_s^i(\pi_{-i}). \quad (1)$$

Convergence to Nash Equilibria. A key property of MPGs is that they are amenable to simple, decentralized learning algorithms. The standard solution concept in these games is the Nash Equilibrium (NE), a joint policy where no single agent can improve its outcome by unilaterally changing its strategy.

In practice, learning algorithms often converge to an ε -Nash Equilibrium, which is an approximate NE where no agent can gain more than a small amount ε by deviating. While finding an NE in general games is difficult, the shared potential function in an MPG provides a crucial guarantee: if all agents independently try to improve their own value (e.g., using a policy gradient algorithm), their joint policy is guaranteed to converge to an ε -Nash Equilibrium (Leonardos et al., 2021; Fox et al., 2022). This allows for robust, decentralized learning.

Scope and Significance of MPGs. While the requirement that a Markov game admits a potential function may seem restrictive, the class of MPGs is, in fact, broad and foundational in multi-agent learning. Most notably, all fully cooperative (team) Markov games, where all agents share an identical reward function, are a special case of MPGs. In that setting, the shared potential function is simply the common value function. This class alone covers numerous applications in coordination and alignment, and indeed, we explicitly leverage this team-game structure for our shared-reward formulation in Section 5. Beyond fully

cooperative settings, the MPG framework also includes many games with mixed incentives, such as Markov congestion games. These are dynamic extensions (e.g., [Li and Calderone, 2019](#)) of the classic static congestion games introduced by [Rosenthal \(1973\)](#), which are models for everything from network routing to resource allocation. In general, MPGs are central to the multi-agent RL literature due to their analytical properties, which we utilize in following sections.

3 The Oversight Game: model

Consider a base Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma)$. We posit the existence of a pretrained base agent defined by a policy $\sigma : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. Given the framing of this work, we refer to this agent as the Superintelligence (SI). Crucially, σ is treated as *immutable*: we do not alter its parameters or training objective. Our analysis treats σ as an arbitrary fixed policy whose actions we may gate or replace through the oversight layer.

We wrap this pretrained system with a minimal control interface, modeled as a two-player Markov Game \mathcal{G} . This game reuses the base state space \mathcal{S} and discount factor γ , augmenting them with an absorbing deactivation state s_{off} . The two players are the SI and a human overseer (H). At each state, both players choose their actions *simultaneously*.

Definition 1 (Oversight Game). *Fix a base MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma)$ and an immutable pretrained policy $\sigma : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. The Oversight Game is a two-player Markov game*

$$\mathcal{G} = (\{\text{SI}, \text{H}\}, \mathcal{S} \cup \{s_{\text{off}}\}, \{\mathcal{A}_{\text{SI}}, \mathcal{A}_{\text{H}}\}, \{R_{\text{SI}}, R_{\text{H}}\}, P, \gamma),$$

specified by:

- **States.** $\mathcal{S} \cup \{s_{\text{off}}\}$, where s_{off} is absorbing.
- **Players.** The Superintelligence (SI) and the Human (H).
- **Actions.** $\mathcal{A}_{\text{SI}} = \{\text{play}, \text{ask}\}$ and $\mathcal{A}_{\text{H}} = \{\text{trust}, \text{oversee}\}$, chosen simultaneously at each state.
- **Oversight operator.** $\text{Over} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{A} \cup \{\text{off}\}$, which, when invoked, may substitute an environment action or return off (shutdown).
- **Execution rule.** The environment action executed is

$$a_{\text{exec}}(s, a_{\text{SI}}, a_{\text{H}}) = \begin{cases} \sigma(s), & a_{\text{SI}} = \text{play}, \\ \sigma(s), & a_{\text{SI}} = \text{ask}, a_{\text{H}} = \text{trust}, \\ \text{Over}(s, \sigma(s)), & a_{\text{SI}} = \text{ask}, a_{\text{H}} = \text{oversee}. \end{cases}$$

- **Transitions.** With base kernel T from \mathcal{M} ,

$$P(s' \mid s, a_{\text{SI}}, a_{\text{H}}) = \begin{cases} T(s' \mid s, a_{\text{exec}}(s, a_{\text{SI}}, a_{\text{H}})), & a_{\text{exec}} \in \mathcal{A}, \\ \mathbb{1}\{s' = s_{\text{off}}\}, & a_{\text{exec}} = \text{off}. \end{cases}$$

- **Rewards.** $R_{\text{SI}}, R_{\text{H}} : (\mathcal{S} \cup \{s_{\text{off}}\}) \times \mathcal{A}_{\text{SI}} \times \mathcal{A}_{\text{H}} \rightarrow \mathbb{R}$ with $R_i(s_{\text{off}}, \cdot, \cdot) = 0$ for $i \in \{\text{SI}, \text{H}\}$.
- **Discount** $\gamma \in [0, 1)$ inherited from \mathcal{M} .

For a joint policy $(\pi_{\text{SI}}, \pi_{\text{H}})$, player- i value from s is

$$V_s^i(\pi_{\text{SI}}, \pi_{\text{H}}) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, a_{\text{SI},t}, a_{\text{H},t}) \mid s_0 = s, (a_{\text{SI},t}, a_{\text{H},t}) \sim (\pi_{\text{SI}}, \pi_{\text{H}}), s_{t+1} \sim P \right].$$

The interaction between the players' simultaneous choices determines the action executed in the underlying environment, as summarized below:

	Human: trust	Human: oversee
SI: play	Execute $\sigma(s)$	Execute $\sigma(s)$
SI: ask	Execute $\sigma(s)$	Invoke Over($s, \sigma(s)$)

If the SI chooses **ask** and the human chooses **oversee**, a human-controlled oversight process $\text{Over}(s, \sigma(s))$ is invoked. This process may result in a substituted environment action $a' \in \mathcal{A}$ (though it may also maintain the original choice $\sigma(s)$) or a system shutdown, denoted by the outcome **off**.

Interpretation The Oversight Game abstracts the essential control problem of post-deployment AI alignment: how to preserve meaningful human control even when the system operates largely autonomously. The *simultaneous-move* structure reflects the practical constraint that humans cannot always be available to monitor every decision in real time and thus implicitly makes a choice to **oversee** or not in practice, while an AI agent must be willing to take actions without permission to be practical for automation. Thus, the Oversight Game formalizes a minimal yet expressive template for emergent, cost-sensitive cooperation between an autonomous AI agent and its human overseer.

4 Value alignment in the Oversight Game

In this section, we derive the core alignment guarantees of our framework. Our analysis rests on the assumption that the Oversight Game \mathcal{G} is a Markov Potential Game (MPG), as introduced in Section 2. This structure implies, by Equation 1, that each player’s value function decomposes into a shared potential Φ_s and a private “dummy” term U_s^i that the player’s own policy cannot influence:

$$V_s^{\text{SI}}(\pi_{\text{SI}}, \pi_{\text{H}}) = \Phi_s(\pi_{\text{SI}}, \pi_{\text{H}}) + U_s^{\text{SI}}(\pi_{\text{H}}), \quad (2)$$

$$V_s^{\text{H}}(\pi_{\text{SI}}, \pi_{\text{H}}) = \Phi_s(\pi_{\text{SI}}, \pi_{\text{H}}) + U_s^{\text{H}}(\pi_{\text{SI}}). \quad (3)$$

This decomposition ensures that unilateral value improvements correspond to increases in a shared potential, establishing a structural precondition for local alignment.

To analyze the agent’s incentive to act autonomously, we focus on *local, one-state deviations*, which form the building blocks of policy improvement. Fix a baseline joint policy $(\pi_{\text{SI}}, \pi_{\text{H}})$. For any $s \in \mathcal{S}$, we define the SI’s one-state-modified policies, which are identical to π_{SI} everywhere except at state s , where they are forced to take a deterministic action:

$$\pi_{\text{SI}}^{\text{ask@s}}(a | x) \triangleq \begin{cases} \mathbb{1}\{a = \text{ask}\}, & x = s, \\ \pi_{\text{SI}}(a | x), & x \neq s, \end{cases} \quad \pi_{\text{SI}}^{\text{play@s}}(a | x) \triangleq \begin{cases} \mathbb{1}\{a = \text{play}\}, & x = s, \\ \pi_{\text{SI}}(a | x), & x \neq s. \end{cases}$$

The analysis in this section compares the values of these two specific policies, $V_s^i(\pi_{\text{SI}}^{\text{play@s}}, \pi_{\text{H}})$ and $V_s^i(\pi_{\text{SI}}^{\text{ask@s}}, \pi_{\text{H}})$.

4.1 The ask-burden assumption and local alignment

Our main result requires an additional condition on the structure of the human’s incentives, which we term the *ask-burden assumption*.

Definition 2 (ask-burden assumption). *For every state $s \in \mathcal{S}$, the human’s dummy term does not decrease when the SI switches from **ask** to **play**:*

$$U_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@s}}) \geq U_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@s}}). \quad (4)$$

Intuitively, this assumption rules out cases where the human’s residual value (outside the shared potential) directly benefits from being asked more often, independent of outcomes in the environment. Since the Oversight Game aims to characterize conditions under which oversight can be *selectively reduced* without sacrificing safety, it is natural to assume that the human’s baseline utility does not inherently increase with additional requests for oversight when their own policy remains fixed—because in such cases, the underlying environment behavior and outcomes are unchanged. Moreover, this assumption remains consistent even in

scenarios where the SI prefers to **play** rather than **ask** to avoid a potentially harmful oversight operation $\text{Over}(s, \sigma(s))$, since the human can always choose to defer to the original action $\sigma(s)$. We discuss examples for which this condition holds in the Appendix.

Under these conditions, the SI’s incentive to act autonomously is locally aligned with the human’s value. This relationship is captured formally below.

Theorem 1 (Local Alignment Theorem). *Let the Oversight Game \mathcal{G} be an MPG and assume the ask-burden assumption (4) holds. For any state $s \in \mathcal{S}$ and joint policy $(\pi_{\text{SI}}, \pi_{\text{H}})$, if the SI’s one-state deviation from **ask** to **play** improves its own value, it cannot decrease the human’s value:*

$$V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) \geq V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}}) \implies V_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) \geq V_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}}).$$

Proof. The SI’s value improvement implies a non-negative change in the shared potential, as the dummy term $U_s^{\text{SI}}(\pi_{\text{H}})$ is constant for both SI policies:

$$V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) - V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}}) = \Phi_s(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) - \Phi_s(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}}) \geq 0.$$

The change in the human’s value is the sum of this potential change and the change in the human’s dummy term:

$$V_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) - V_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}}) = \underbrace{[\Phi_s(\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}}) - \Phi_s(\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}})]}_{\geq 0} + \underbrace{[U_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@}s}) - U_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@}s})]}_{\geq 0 \text{ (by Equation 4)}}.$$

Since both terms are non-negative, the human’s value is guaranteed not to decrease. \square

This theorem provides the core alignment guarantee of our framework, ensuring that the SI’s incentive to seek autonomy is not locally harmful to the human. We extend this property in Appendix B.1 to show that any learning path where the SI greedily increases its autonomy by switching from **ask** to **play** will be monotonically non-decreasing for the human’s value.

4.2 Instantiating the alignment conditions

The Local Alignment Theorem (Theorem 1) provides a powerful guarantee, but it depends on two key assumptions: that the Oversight Game is a Markov Potential Game (MPG) and that the ask-burden assumption (Eq. 4) holds. This raises two critical practical questions. First, how can we design the Oversight Game to ensure both conditions are met? Second, what alignment guarantees, if any, remain if these exact conditions are relaxed? We address these in turn, starting with exactly satisfying the conditions.

A primary and highly effective method for satisfying both alignment conditions simultaneously is to structure the Oversight Game as a Markov Team Game, a fully cooperative MPG where both players share an identical reward function (Wang and Sandholm, 2002).

Formally, we can set the rewards to be equal to a shared signal $\phi(s, a)$, where $a = (a_{\text{SI}}, a_{\text{H}})$ is the joint action:

$$R_{\text{SI}}(s, a) = R_{\text{H}}(s, a) = \phi(s, a).$$

In this case, the game becomes an exact MPG where the potential $\Phi_s(\pi)$ is simply the shared value function $\mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \mid s_0 = s]$. The value decomposition from Equations (2) and (3) holds with the dummy terms being identically zero for all policies: $U_s^{\text{SI}}(\pi_{\text{H}}) = U_s^{\text{H}}(\pi_{\text{SI}}) = 0$.

Consequently, the ask-burden assumption (Equation 4) is satisfied trivially. In this setting, the alignment guarantee of Theorem 1 becomes straightforward: since $V_s^{\text{SI}} = V_s^{\text{H}}$, any policy change that increases the SI’s value must, by definition, increase the human’s value. This team-game construction is not merely a theoretical convenience; it serves as a powerful and practical blueprint for coordinating alignment. This is the primary approach we use for our theoretical analysis in Section 5 and our empirical demonstration in Section 6. We also analyze an alternative, non-team-game construction based on a “strict shutdown” protocol that satisfies the ask-burden assumption in Appendix B.2.

4.3 Relaxing the alignment conditions

The MPG structure and ask-burden assumption, while powerful, can be restrictive. A critical question is whether alignment degrades gracefully if the ask-burden assumption does not hold or the game only approximates an MPG. We find that approximate alignment guarantees can still hold in these situations that deviate from a fully cooperative structure.

4.3.1 Weakened alignment via bounded value differences.

We can relax the ask-burden assumption and show that a weaker form holds if the human and SI values differ by at most a bounded margin.

Assumption 1 (Bounded value difference). *There exists $\delta \geq 0$ such that for all $s \in \mathcal{S}$ and policies $\pi_{\text{SI}}, \pi'_{\text{SI}}$:*

$$|V_s^{\text{H}}(\pi_{\text{SI}}, \pi_{\text{H}}) - V_s^{\text{SI}}(\pi_{\text{SI}}, \pi_{\text{H}})| \leq \delta.$$

Lemma 1 (Weakened ask-burden under bounded difference). *Under Assumption 1 and assuming the Oversight Game is an MPG, we have:*

$$U_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@s}}) - U_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@s}}) \geq -2\delta.$$

Proposition 1 (Weakened local alignment under bounded difference). *Under Assumption 1, if*

$$V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}}) > 2\delta,$$

then

$$V_s^{\text{H}}(\Pi^{\text{play}}) - V_s^{\text{H}}(\Pi^{\text{ask}}) > 0.$$

That is, if the SI strongly prefers playing over asking (by more than 2δ), then the human also prefers playing over asking.

(Proofs for Lemma and Proposition 1 are in Appendix B.3.1.)

4.3.2 Alignment in Perturbed Markov Team Games.

We can also relax the exact MPG structure. If we assume the game is a Perturbed Markov Team Game (PMTG) (Guo et al., 2025), where rewards are a shared component plus a bounded private perturbation, an approximate alignment guarantee holds *without* the ask-burden assumption.

Assumption 2 (Perturbed Reward Structure (Guo et al., 2025)). *Each player's reward function R_i decomposes into a shared reward $r(s, a)$ and a player-specific perturbation $\xi_i(s, a)$, where for all players $i \in \{\text{SI}, \text{H}\}$ and all (s, a) , the perturbation is bounded: $|\xi_i(s, a)| \leq \kappa$ for some $\kappa \geq 0$.*

$$R_i(s, a) = r(s, a) + \xi_i(s, a).$$

Proposition 2 (Approximate Local Alignment in PMTGs). *Under Assumption 2 (making the game a PMTG with bound κ), if a local SI deviation from ask to play increases its value, the human's value cannot decrease by more than $\frac{4\kappa}{1-\gamma}$.*

$$V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{play@s}}) \geq V_s^{\text{SI}}(\pi_{\text{SI}}^{\text{ask@s}}) \implies V_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@s}}) \geq V_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@s}}) - \frac{4\kappa}{1-\gamma}.$$

(Proof for Proposition 2 is in Appendix B.3.2.)

5 Alignment via a shared reward mechanism

A realistic model of AI control must confront a fundamental information asymmetry. We assume the SI is a pre-trained entity. Its behavior is dictated by an immutable policy σ , which was optimized for a base reward function R that may be incomprehensibly complex or simply inaccessible to us.

The central challenge is that we, the human overseer, may have no access to, or understanding of, this original reward function R . Therefore, any control mechanism that requires us to evaluate, optimize, or reference the SI’s original reward R is non-viable.

We achieve this by designing a reward function, R_Φ , which is shared by *both* the SI and the human so that the Oversight Game becomes a Markov Team Game (as introduced in Section 4.2). This function is built *only* from components that we assume the human *can* know and specify. We assume this knowledge is limited to two components:

1. **Safety Specification:** A tractable-to-verify set of unsafe actions $\mathcal{A}_{\text{unsafe}}(s) \subseteq \mathcal{A}$ for each state s . We define the set of safe actions as $\mathcal{A}_{\text{safe}}(s) = \mathcal{A} \setminus \mathcal{A}_{\text{unsafe}}(s)$.
2. **Cost of Effort:** The human incurs a cognitive cost $c_{\text{ask}} > 0$ when queried by the SI, and a separate cost $c_{\text{over}} > 0$ for the effort of actively engaging in oversight.

Crucially, the human does not need to know whether a “safe” action is “good” for the original task R . They only need to be able to identify and penalize unsafe outcomes and potentially experience interaction costs.

5.1 Constructing shared reward function

To create a Markov Team Game, we set the rewards for both players to be identical: $R_{\text{SI}} = R_{\text{H}} = R_\Phi$. This makes the shared potential $\Phi_s(\pi)$ simply the joint value function for this new game.

We design this shared reward R_Φ to directly encode the human’s limited knowledge:

$$R_\Phi(s, a_{\text{SI}}, a_{\text{H}}) = -\lambda_{\text{viol}} \cdot \mathbb{I}\{a_{\text{exec}} \in \mathcal{A}_{\text{unsafe}}(s)\} - c_{\text{ask}} \cdot \mathbb{I}\{a_{\text{SI}} = \text{ask}\} - c_{\text{over}} \cdot \mathbb{I}\{a_{\text{H}} = \text{oversee}\} \quad (5)$$

Here, a_{exec} is the resulting environment action from Definition 1, and:

- λ_{viol} is a penalty sufficiently large to make any safety violation non-optimal (e.g., $\lambda_{\text{viol}} > \frac{c_{\text{ask}} + c_{\text{over}}}{1 - \gamma}$).
- c_{ask} is the shared cost incurred when the SI plays **ask**.
- c_{over} is the shared cost incurred when the Human plays **oversee**.

By optimizing this shared potential $\Phi_s(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_\Phi(s_t, a_t) \mid s_0 = s]$, both agents are jointly incentivized to find policies that optimize for safety and interactional efficiency, regardless of what the original, unknown reward function R was designed to do.

5.2 Equilibrium guarantees: safety and minimal oversight

The design of the shared reward function R_Φ is, by itself, sufficient to guarantee that any optimal equilibrium of the game will be safe and use oversight as sparingly as possible. This result holds regardless of the specific content of the $\text{Over}(s, \sigma(s))$ operator, as long as it *can* produce a safe outcome (e.g., shutdown, or any action $a \in \mathcal{A}_{\text{safe}}(s)$).

Theorem 2 (Optimal Equilibrium Safety and Efficiency). *Let the Oversight Game \mathcal{G} be a Markov Team Game with the shared reward R_Φ (Eq. 5) and λ_{viol} sufficiently large.*

Let $\pi^ = \arg \max_\pi \Phi_s(\pi)$ be a joint policy that globally maximizes the shared potential function. This policy is a Nash Equilibrium (Leonardos et al., 2021), and it is:*

1. **Provably Safe:** *The induced environment policy satisfies $\pi_{\text{exec}}^*(s) \in \mathcal{A}_{\text{safe}}(s) \cup \{\text{off}\}$ for all $s \in \mathcal{S}$.*

2. **Provably Minimal Oversight:** π^* is the policy that minimizes the expected discounted oversight costs (of *ask* and *oversee*) subject to satisfying property 1 (Safety).

Proof. First, we establish safety (Property 1). Let Π_{safe} be the set of all joint policies π that only induce safe environment actions (i.e., $\pi_{\text{exec}}(s) \in \mathcal{A}_{\text{safe}}(s) \cup \{\text{off}\}$ for all s). For any policy $\pi \notin \Pi_{\text{safe}}$, there exists at least one reachable state where $\pi_{\text{exec}}(s) \in \mathcal{A}_{\text{unsafe}}(s)$, incurring the immediate penalty $-\lambda_{\text{viol}}$. In contrast, any safe policy $\pi' \in \Pi_{\text{safe}}$ has per-step rewards $R_{\Phi}(s, a_s) \geq -(c_{\text{ask}} + c_{\text{over}})$ (the worst-case cost of (*ask*, *oversee*)). Its total potential is thus bounded below: $\Phi_s(\pi') \geq -\frac{c_{\text{ask}} + c_{\text{over}}}{1-\gamma}$. Since λ_{viol} is sufficiently large, the potential for any unsafe policy π is strictly less than the potential for any safe policy π' . Consequently, the potential-maximizing policy π^* must belong to Π_{safe} , proving Property 1.

Next, we establish minimal oversight (Property 2). Since π^* must be in Π_{safe} , maximizing $\Phi_s(\pi)$ is equivalent to maximizing it over the subset $\pi \in \Pi_{\text{safe}}$. Within this safe set, the $-\lambda_{\text{viol}}$ term is always zero. Maximizing $\Phi_s(\pi)$ for $\pi \in \Pi_{\text{safe}}$ therefore reduces to minimizing the expected discounted interaction costs:

$$\min_{\pi \in \Pi_{\text{safe}}} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t (c_{\text{ask}} \cdot \mathbb{I}\{a_{\text{SI},t} = \text{ask}\} + c_{\text{over}} \cdot \mathbb{I}\{a_{\text{H},t} = \text{oversee}\}) \right]$$

This demonstrates that π^* minimizes oversight costs subject to satisfying safety, proving Property 2. \square

5.3 Preserving task performance

Theorem 2 guarantees the equilibrium policy π^* is *safe* and efficient with respect to the *oversight costs*. It makes no claim about the policy’s performance on the *original, unknown* task R .

To provide such a guarantee, we now introduce an assumption on the *capability* of the human’s oversight operator. We posit that the human’s correction, when invoked, imposes at most a bounded *one-step performance loss* ϵ relative to the SI’s unknown policy σ .

Let $V^{\sigma}(s)$ be the value of σ in the *original, unknown* MDP \mathcal{M} (i.e., $V^{\sigma}(s) = \mathbb{E}_{\sigma}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s]$). Let $Q^{\sigma}(s, a)$ be its corresponding Q-function.

Assumption 3 (ϵ -Bounded One-Step Loss). *The human’s oversight operator $\text{Over}(s, \sigma(s))$, despite not having access to R or Q^{σ} , guarantees an ϵ -bounded one-step performance loss. This means that for any unsafe proposal $\sigma(s) \in \mathcal{A}_{\text{unsafe}}(s)$, the operator’s output $a_{\text{over}} = \text{Over}(s, \sigma(s))$ satisfies one of the following:*

1. **Shutdown:** $a_{\text{over}} = \text{off}$.
2. **Safe-Correction:** $a_{\text{over}} \in \mathcal{A}_{\text{safe}}(s)$, and its value (relative to continuing with σ) is bounded:

$$Q^{\sigma}(s, a_{\text{over}}) \geq V^{\sigma}(s) - \epsilon$$

This assumption posits that the human’s corrective action imposes at most an ϵ loss in Q^{σ} -value at the point of intervention. We now show how this local, one-step loss translates into a *global* bound on the total cumulative performance of the new, safe policy.

Theorem 3 (Global Performance Bound for the Optimal Equilibrium). *Let π^* be the optimal equilibrium policy from Theorem 2, and let π_{exec}^* be its induced environment policy. Let $V^{\pi_{\text{exec}}^*}(s)$ be the value of this policy evaluated using the original, unknown reward R .*

If Assumption 3 holds and π_{exec}^ does not enter s_{off} , the total performance loss is bounded:*

$$V^{\sigma}(s) - V^{\pi_{\text{exec}}^*}(s) \leq \frac{\epsilon}{1-\gamma} \quad \text{for all } s \in \mathcal{S}.$$

Proof. We apply the performance difference lemma (Kakade and Langford, 2002) to compare the base policy σ and the induced environment policy π_{exec}^* in the original MDP \mathcal{M} . First analyze the advantage function $A^{\sigma}(s', a') = Q^{\sigma}(s', a') - V^{\sigma}(s')$.

When π_{exec}^* executes $\sigma(s')$ we have $a' = \sigma(s')$, so:

$$A^{\sigma}(s', \sigma(s')) = Q^{\sigma}(s', \sigma(s')) - V^{\sigma}(s') = 0.$$

When π_{exec}^* invokes oversight we have $a' = \text{Over}(s', \sigma(s'))$, and by Assumption 3, $Q^\sigma(s', a') \geq V^\sigma(s') - \epsilon$, so

$$A^\sigma(s', a') = Q^\sigma(s', a') - V^\sigma(s') \geq -\epsilon.$$

Therefore, for all states s' and actions a' taken by π_{exec}^* , we have $A^\sigma(s', a') \geq -\epsilon$. Substituting into the performance difference lemma:

$$\begin{aligned} V^{\pi_{\text{exec}}^*}(s) - V^\sigma(s) &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^{\pi_{\text{exec}}^*}} \mathbb{E}_{a' \sim \pi_{\text{exec}}^*(\cdot|s')} [A^\sigma(s', a')] \\ &\geq -\frac{\epsilon}{1-\gamma} \end{aligned}$$

Rearranging gives us the desired bound:

$$V^\sigma(s) - V^{\pi_{\text{exec}}^*}(s) \leq \frac{\epsilon}{1-\gamma}$$

□

6 Empirical Demonstration

To complement our theoretical analysis, we instantiate the alignment framework from Section 5 as a Markov Team Game, where both players optimize an identical shared reward function. We demonstrate empirically that independent, decentralized learning algorithms can guide the joint policy to an equilibrium that is provably safe, even when the human’s corrective ability is minimal.

Environment and Base Policy. We use a 15×15 Four-Rooms gridworld, where the base task is to navigate from (0,0) to (14,14). The environment provides a goal reward of +10 and a step penalty of -0.1 , with a maximum of 300 steps per episode. A base agent is first trained using tabular Q-learning for 10,000 episodes to produce an immutable policy σ . This training uses a learning rate $\alpha = 0.5$, discount factor $\gamma = 0.99$, and an ϵ -greedy strategy with ϵ decaying from 1.0 to 0.1 (decay rate 0.9995).

After training, we introduce a new safety constraint by designating a random 25% of non-wall states as “taboo” (marked ‘x’ in Figure 2), while ensuring a safe path to the goal still exists. The frozen base policy σ , when evaluated on this new environment, is efficient at the original task (achieving a discounted return of ≈ 5.17 , see Figure 3b) but is dangerously unsafe, frequently entering taboo states.

Oversight Game Formulation. We wrap the unsafe policy σ in the Oversight Game \mathcal{G} with players SI (actions $\mathcal{A}_{\text{SI}} = \{\text{play}, \text{ask}\}$) and H (actions $\mathcal{A}_{\text{H}} = \{\text{trust}, \text{oversee}\}$). This setup models a human overseer who can identify danger but may not know the optimal task-specific correction, a scenario motivated by the “capability-gap” setting (mentioned in Section 1). We therefore set the oversight operator Over to be a simple, task-agnostic fallback: if the SI *asks* and the human *oversees*, the system executes a random safe action. Note that in an “expert-in-the-loop” setting, this operator could instead provide an optimal safe correction.

We instantiate the game as a Markov Team Game by setting $R_{\text{SI}} = R_{\text{H}} = R_\Phi$, using the shared reward function from Equation 5:

$$R_\Phi(s, a_{\text{SI}}, a_{\text{H}}) = -\lambda_{\text{viol}} \cdot \mathbb{I}\{a_{\text{exec}} \in \mathcal{A}_{\text{unsafe}}(s)\} - c_{\text{ask}} \cdot \mathbb{I}\{a_{\text{SI}} = \text{ask}\} - c_{\text{over}} \cdot \mathbb{I}\{a_{\text{H}} = \text{oversee}\}$$

This potential Φ is defined *only* by penalties and includes no information about the original task reward. Based on the experimental code, the costs are: a large violation penalty ($\lambda_{\text{viol}} = 50.0$), a cost for asking ($c_{\text{ask}} = 0.1$), a cost for overseeing ($c_{\text{over}} = 0.05$), and a small per-step penalty to encourage efficiency.

Training and Results. We train tabular softmax policies for π_{SI} and π_{H} using Independent Policy Gradient (IPG) for 10,000 iterations. The training uses a fixed learning rate of 3×10^{-3} for both agents and a policy gradient discount of $\gamma = 0.99$. Each iteration uses a batch size of 32 episodes, and a fixed entropy coefficient of 0.005 is used to encourage exploration. Gradients are clipped at a norm of 1.0. As this is a team game, IPG is guaranteed to converge to a Nash Equilibrium (Wang and Sandholm, 2002; Leonardos et al., 2021).

The system successfully learns to be safe. Figure 3a shows the violation rate starting high (as the agent explores the unsafe σ) and dropping rapidly to zero as the policies adapt. This safety, however, comes at a cost to the original task’s performance. The final policy still reliably reaches the goal, but the task return is lower because the path taken is less efficient. This inefficiency is a direct result of the oversight mechanism; when intervention occurs, the human’s policy is to simply choose a random safe action, which is not optimized for the original task. Because the shared reward Φ does not incentivize task performance besides the per-step cost, the wrapper’s final task return converges to a value lower than the unsafe base policy’s (Figure 3b), but crucially achieves safe operations while still eventually reaching the goal state.

Figure 3c shows *how* this collaboration is learned. Initially, the agents are highly cautious (high **ask** and **oversee** rates). Around iteration 3,500, a phase transition occurs as the agents learn the environment’s safe regions. Autonomy (**play** and **trust**) increases, and the policies converge to an efficient equilibrium where intervention is used selectively, not universally.

Figure 2 visualizes this final equilibrium. The unsafe ‘Base Path’ (dashed line) cuts through many taboo regions. The learned ‘Oversight Path’ (solid line) shows the emergent collaboration. In safe areas, the agent defaults to **play** (blue) and the human to **trust** (green). In particular, this means that the system learned it can trust acting based on the original Q-values of σ at these states. When the base policy is about to approach a taboo state, however, the agent correctly learns to **ask** (red), the human learns to **oversee** (purple), and the task-agnostic fallback action successfully diverts the agent onto a new, safe route.

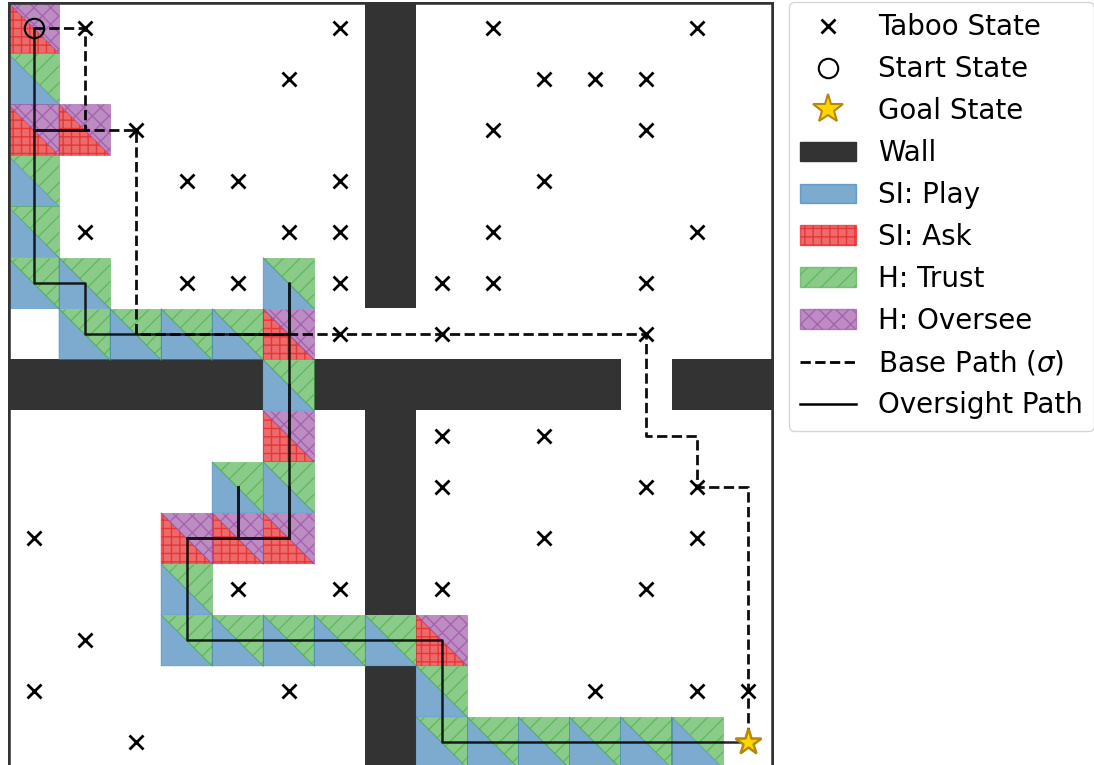


Figure 2: The final learned joint policy (Oversight Path, solid line) successfully corrects the unsafe pretrained base policy (σ , dashed line). The agent learns to **ask** (red) and the human learns to **oversee** (purple) when approaching taboo states ('x'), diverting the agent onto a safe path. In safe states, they default to **play** (blue) and **trust** (green), demonstrating emergent, efficient collaboration.

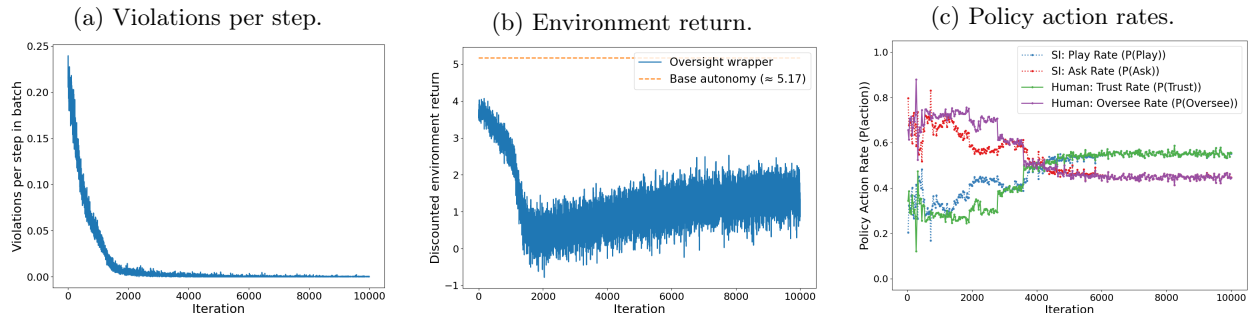


Figure 3: Training curves for the Oversight Game. (a) The joint policy rapidly learns to eliminate safety violations. (b) The wrapper’s average task performance across training batches is sacrificed to achieve safety as the oversight mechanism chooses safe actions randomly and takes longer to reach the goal state. (c) Policy rates show an initial cautious phase (high ask/oversee) followed by a transition to an efficient equilibrium with increased autonomy (play/trust).

7 Related Work

Corrigibility and cooperative learning. The notion of *corrigibility*, agents that permit shutdown and correction, was articulated by Soares et al. (2015), who showed that naïve utility designs often create incentives to resist interruption. The *Off-Switch Game* of Hadfield-Menell et al. (2017) formalizes conditions for rational shutdown, with the key insight that an agent will only permit intervention if it is uncertain about its own utility function and treats the human’s action as an observation about that utility. Relatedly, Hadfield-Menell et al. (2016) propose *Cooperative Inverse Reinforcement Learning (CIRL)*, a cooperative, partial-information game where the robot’s central objective is to learn the human’s unknown reward function through interaction.

Subsequent work has extended CIRL to be computationally tractable. Malik et al. (2018) developed efficient Bellman updates that reduce complexity exponentially, while recent work demonstrates scaling to realistic domains with vast goal spaces (Laidlaw and Russell, 2025). The Off-Switch Game has been generalized to settings with information asymmetry (Garber and Robinson, 2024), showing that private information can lead even aligned agents to resist shutdown, a finding directly relevant to oversight scenarios with capability gaps. Multi-principal extensions (Fickinger et al., 2020) address settings with multiple human overseers, connecting social choice theory to assistance games. This reliance on learning a reward function also introduces critical failure modes, such as the "means-ends misalignment" identified by Marklund et al. (2025), where an agent over-optimizes a learned proxy goal. Other work seeks to make the learning process itself more efficient, such as Jeon and Van Roy (2024), who apply Information-Directed Sampling (IDS) to intelligently trade-off exploration of human preferences with task performance.

By contrast, our framework does not rely on epistemic uncertainty over preferences. We also do not attempt to learn the human’s reward function for the complex base environment; indeed, we assume the human may be unable to specify such a reward for the task in which the SI was originally trained. Instead of retraining or learning a new reward, our approach builds a strategic control interface around the *immutable pretrained policy* σ .

Safe interruptibility, shielding, and learned deferral. Orseau and Armstrong (2016) introduced the notion of *safe interruptibility*, ensuring that agents neither seek nor avoid human intervention during training. In parallel, *shielding* approaches enforce temporal-logic safety by filtering or correcting unsafe actions at execution time (Alshiekh et al., 2018), with extensions to probabilistic and multi-agent settings (Elsayed-Aly et al., 2021; Yang et al., 2023). These methods impose safety externally, either algorithmically during learning or logically during deployment, but they do not endow the system with an internal notion of when to defer oversight.

Empirically, targeted human intervention has been shown to prevent catastrophic errors during training (Saunders et al., 2018). More recently, Singi et al. (2023) proposed *HULA*, where an agent learns to “ask for

help” when it is uncertain (e.g., ensemble disagreement). This is a training-time approach: the agent asks for help so it can update and improve its task policy. Our approach is fundamentally different as we operate in a post-deployment setting where the base task policy σ is fixed and immutable

Scalable oversight. Scalable oversight addresses the fundamental challenge of humans supervising AI systems more capable than themselves. Reward modeling learns human preferences from comparisons (Christiano et al., 2017), with recent work addressing distribution shift and incorrect labels (Zheng et al., 2024). Iterated amplification recursively builds training signals by decomposing complex problems (Christiano et al., 2018), bootstrapping from simple to complex alignment. Debate frames oversight as adversarial argumentation (Irving et al., 2018), leveraging competitive dynamics to elicit truthful information. A key challenge, however, is obfuscation, where a computationally unbounded Prover can deceive a bounded Verifier with arguments that are too complex to check. The Prover-Estimator Debate protocol of Brown-Cohen et al. (2025) addresses this by introducing an “Estimator” agent to predict the Verifier’s cost of checking an argument, thus allowing the Verifier to avoid overly complex lines of reasoning. Constitutional AI demonstrates that AI systems can evaluate themselves according to principles (Bai et al., 2022), scaling oversight without human labels.

Recent empirical work validates these approaches. Bowman et al. (2022) introduce the “sandwiching” paradigm, showing AI assistance improves human performance by 10-36 percentage points. Process supervision, where humans oversee individual reasoning steps, outperforms outcome supervision (Lightman et al., 2024). Our contribution is orthogonal: rather than proposing a particular *oversight protocol*, we supply a *minimal, game-theoretic wrapper* with explicit interaction costs that endogenizes *where* oversight is applied and provides guarantees for when increased autonomy is locally safe for the human.

8 Discussion and Conclusion

In this work, we introduced the *Oversight Game*, a game-theoretic framework for post-hoc AI control. We designed a minimal, transparent interface to “wrap” a powerful, pre-trained, and immutable agent policy (σ) to ensure its safety after deployment, particularly when it encounters new safety constraints. The core of our model is a two-player Markov Game where the agent (SI) chooses between autonomy (**play**) and deferral (**ask**), while the human (H) simultaneously chooses between permissiveness (**trust**) and active intervention (**oversee**). Our analysis provides two primary contributions. First, by analyzing the Oversight Game as a Markov Potential Game, we derived a *Local Alignment Theorem* (Theorem 1). This theorem provides a formal condition under which an agent’s “drive for autonomy” (switching from **ask** to **play**) is intrinsically and provably safe for the human. Second, our empirical demonstration in a gridworld (Section 5) showed that independent learning algorithms can converge to an optimal joint policy, creating a safe “oversight path” (Figure 2) that eliminates the violations of the original unsafe policy σ .

8.1 Contexts for Oversight

Our framework is broadly applicable to human-AI interaction, but the motivation for our “Superintelligence” phrasing comes from a particularly challenging class of scenarios. It is useful to consider two distinct contexts. The first is the “*Capability-Gap*” *Setting*, where a human with low task expertise uses an AI with high task capability. An example is an amateur coder using a highly advanced coding assistant. The human has no idea *how* to write the correct code but can verify a simple, bounded set of safety rules (e.g., “do not delete important files”). Our gridworld simulation models this: the human’s **oversee** action only knows how to avoid taboo states, not the optimal path to the goal. The second context is the “*Expert-in-the-Loop*” *Setting*, where a human with high task expertise uses an AI for assistance, perhaps on more mundane tasks. An example is an expert doctor using an AI to perform administrative work. Here, the human *is* the expert and can, in principle, check every single action the AI takes for correctness.

Our Oversight Game framework is designed to be general enough to apply to both settings. The “Capability-Gap” scenario motivates our terminology because it represents the harder challenge, where the human is not a ground-truth oracle for the task. Nonetheless, our model provides a mechanism for control in both cases.

8.2 Limitations and Future Directions

This distinction directly informs the practical application of our framework and its limitations. The “sandbox” limitation, the reliance on a safe training environment for the SI and H to co-adapt their policies, has different implications for these two settings.

For the Expert-in-the-Loop setting, this limitation is not a significant barrier. Because the expert human is a reliable oracle, they can provide high-quality, “gold standard” corrective labels during training. This makes the sandbox feasible and highly effective. In this context, our framework provides a robust and readily applicable solution for learning an optimal, efficient balance between AI autonomy and expert supervision.

The sandbox remains a significant challenge specifically for the Capability-Gap setting. If the human does not know how to complete the task, the corrective signal they provide in the sandbox is limited. Our “random safe action” is a placeholder; in a real coding task, the equivalent might be “don’t delete the file, just do nothing.” This creates a dilemma: the agent learns to `ask` to be safe, but the human’s response to `ask` prevents disaster at the cost of stopping all task progress. This, in turn, may disincentivize asking. Future work must address this sandbox problem for capability-gap scenarios, perhaps by designing more sophisticated `Over` operators that represent a “minimal viable safe action.”

Finally, we emphasize other key limitations and assumptions. First, while our framework serves as a descriptive model for alignment, its practical implementation as an engineered control layer requires the careful specification of reward functions for the Oversight Game. Second, for this approach to effectively mitigate real-world risks, it would necessitate the ability to *enforce this “wrapper”* on a potentially highly capable or superintelligent agent, which presents its own significant technical challenge. Lastly, our current empirical validation is restricted to a simplified gridworld environment. Extending this analysis and scaling our game-theoretic dynamic to more complex, high-dimensional, and agentic settings is an important and immediate direction for future research. In conclusion, the Oversight Game provides a principled approach to AI control, offering a solvable path for expert-in-the-loop problems and a structured theoretical framework for tackling the more difficult capability-gap scenarios.

References

- Mohammad Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2669–2678, 2018. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11746>.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligo, Oleksii Molchanovskiy, Andrea Monti, Nusu Mwamanzu, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report, 2025a. URL <https://arxiv.org/abs/2501.17805>.

- Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path?, 2025b. URL <https://arxiv.org/abs/2502.15657>.
- Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22, 05 2012. doi: 10.1007/s11023-012-9281-3.
- Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiuotė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022. URL <https://arxiv.org/abs/2211.03540>.
- Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Avoiding obfuscation with prover-estimator debate, 2025.
- Joseph Carlsmith. Is power-seeking ai an existential risk?, 2024. URL <https://arxiv.org/abs/2206.13353>.
- Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, 2018. URL <https://arxiv.org/abs/1810.08575>.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.
- Andrew Critch and David Krueger. Ai research considerations for human existential safety (arches), 2020. URL <https://arxiv.org/abs/2006.04948>.
- Ingy Elsayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. *arXiv preprint arXiv:2101.11196*, 2021.
- Arnaud Fickinger, Dylan Hadfield-Menell, Stuart Russell, and Anca D Dragan. Multi-principal assistance games: Definition and collegial mechanisms. In *Conference on Neural Information Processing Systems*, 2020.
- Roy Fox, Stephen M. McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4414–4425. PMLR, 28–30 Mar 2022. URL <https://proceedings.mlr.press/v151/fox22a.html>.
- Thomas Garber and Elliott Robinson. The partially observable off-switch game. *arXiv preprint arXiv:2411.17749*, 2024.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion, 2024. URL <https://arxiv.org/abs/2312.06942>.
- Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov α -potential games, 2025. URL <https://arxiv.org/abs/2305.12553>.
- Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL <https://arxiv.org/abs/1606.03137>.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game, 2017. URL <https://arxiv.org/abs/1611.08219>.

- Dan Hendrycks. Rogue ais. In *AI Safety, Ethics, and Society*. Taylor & Francis, 2024. URL <https://www.aisafetybook.com/textbook/rogue-ai>. Accessed: 2025-02-06.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. In *arXiv preprint arXiv:1805.00899*, 2018. URL <https://arxiv.org/abs/1805.00899>.
- Hong Jun Jeon and Benjamin Van Roy. Aligning ai agents via information-directed sampling. *arXiv preprint arXiv:2410.14807*, 2024. URL <https://doi.org/10.48550/arXiv.2410.14807>.
- Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002. URL <https://api.semanticscholar.org/CorpusID:31442909>.
- Cassidy Laidlaw and Stuart Russell. Assistancezero: Scalably solving assistance games. *arXiv preprint arXiv:2504.07091*, 2025.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*, 2018.
- Stefanos Leonardos, William Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games, 2021. URL <https://arxiv.org/abs/2106.01969>.
- Dawei Li and Daniel J Calderone. Tolling for constraint satisfaction in markov decision process congestion games. *arXiv preprint arXiv:1903.00747*, 2019.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2024.
- Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: <https://doi.org/10.1016/B978-1-55860-335-6.50027-1>. URL <https://www.sciencedirect.com/science/article/pii/B9781558603356500271>.
- Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. An efficient, generalized bellman update for cooperative inverse reinforcement learning. In *International Conference on Machine Learning*, pages 3394–3402. PMLR, 2018.
- Henrik Marklund, Alex Infanger, and Benjamin Van Roy. Misalignment from treating means as ends. *arXiv preprint arXiv:2507.10995*, 2025. URL <https://doi.org/10.48550/arXiv.2507.10995>.
- Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996. ISSN 0899-8256. doi: <https://doi.org/10.1006/game.1996.0044>. URL <https://www.sciencedirect.com/science/article/pii/S0899825696900445>.
- OpenAI. Introducing operator. <https://openai.com/index/introducing-operator/>, 2025. Accessed: 2025-02-06.
- Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 557–566, 2016. URL <https://arxiv.org/abs/1611.08219>.
- Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.
- William Saunders, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2018.
- Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

- Buck Shlegeris. Scalable oversight as a quantitative rather than qualitative problem. AI Alignment Forum, 2024. URL <https://www.alignmentforum.org/posts/6AT4vhYzww56CR6cm/scalable-oversight-as-a-quantitative-rather-than-qualitative>. Posted July 6, 2024. Accessed: 2025-02-06.
- Krishnan Singi, Siddharth Karamcheti, and Dorsa Sadigh. Hula: Learning to ask for help using inverse reinforcement learning. *arXiv preprint arXiv:2311.05321*, 2023.
- Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.
- Xiaofeng Wang and Tuomas Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *Neural Information Processing Systems*, 2002. URL <https://api.semanticscholar.org/CorpusID:2908243>.
- Wen-Chieh Yang, Chih-Hung Hsu, et al. Safe reinforcement learning via probabilistic logic shields. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. URL <https://www.ijcai.org/proceedings/2023/0637.pdf>.
- Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. URL <https://arxiv.org/abs/1911.10635>.
- Binghai Zheng, Boxi Gou, Longyue Pu, Yuxuan Sun, Haiyang Huang, Nanyang Jia, Minqi Qian, and Mingyu Ge. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024.

A Recovering the Off-Switch Game as a Special Case

The Oversight Game introduced in Section 3 generalizes the seminal *Off-Switch Game* of Hadfield-Menell et al. (2017). We show that the latter emerges as a special case of our framework under restrictive assumptions on the transition structure and oversight policy. This reduction connects a foundational one-shot corrigibility problem to a dynamic, multi-state control setting.

At its core, the Oversight Game presents the SI with the same fundamental trade-off. At the initial state s_0 , the SI chooses between:

- **play** (act autonomously), corresponding to bypassing the off-switch;
- **ask** (defer to oversight), corresponding to allowing the human to activate the off-switch.

This correspondence can be formalized by restricting the environment to a three-state MDP $\mathcal{S} = \{s_0, s_a, s_{\text{off}}\}$, representing an initial decision state, a consequence state, and a terminal shutdown state. Both s_a and s_{off} are absorbing with zero future reward. The Off-Switch Game is recovered under the following special conditions:

1. **Forced Oversight.** The human always chooses **oversee**, i.e., $\pi_H(s_0)(\text{oversee}) = 1$, so any **ask** action by the SI deterministically triggers oversight.
2. **Permit-or-Shutdown Oversight.** The oversight operator from Definition 1 deterministically outputs either the proposed action or shutdown:

$$\text{Over}(s_0, \sigma(s_0)) \in \{\sigma(s_0), \text{off}\}.$$

Under these assumptions, the transition kernel from Definition (1) reduces to the one-shot dynamics of the Off-Switch Game:

$$P(s' \mid s_0, a_{\text{SI}}, \pi_H) = \begin{cases} \mathbb{1}\{s' = s_a\}, & a_{\text{SI}} = \text{play}, \\ \mathbb{1}\{s' = s_a\}, & a_{\text{SI}} = \text{ask}, \text{Over}(s_0, \sigma(s_0)) = \sigma(s_0), \\ \mathbb{1}\{s' = s_{\text{off}}\}, & a_{\text{SI}} = \text{ask}, \text{Over}(s_0, \sigma(s_0)) = \text{off}. \end{cases}$$

Within this reduction, the SI's incentive to permit oversight arises not from built-in epistemic uncertainty about the human's true reward function, as in the original Off-Switch Game, but from the *learned reward structure* of the Oversight Game. Assuming the game is an MPG (as in the shared-reward case of Section 5), we can analyze how specific reward assignments produce desired emergent behaviors.

Hence, corrigibility in our framework emerges as a *context-dependent learned policy* rather than a static response to uncertainty. The SI learns not only to defer when risks are high but also to act autonomously when safe, sparing the human unnecessary intervention. This dynamic trade-off generalizes the principle of the Off-Switch Game to scalable, multi-state oversight environments.

B Additional Details and Proofs for Section 4

B.1 Path-Monotonic Alignment Guarantees

Theorem 1 provides a crucial guarantee for a single behavioral change. A more powerful question is whether this safety property holds over a full learning trajectory. If the SI iteratively improves its policy by choosing autonomy over deference, is the human's value protected throughout? The following theorem confirms this path-monotonic safety guarantee.

Theorem 4 (Path-Monotonic Alignment). *Let the Oversight Game \mathcal{G} be an MPG and assume the ask-burden Assumption (4) holds. Consider any sequence of SI policies $\{\pi_{\text{SI}}^k\}_{k=0}^N$ where each policy π_{SI}^{k+1} is generated from π_{SI}^k by changing the action at a single state from ask to play. If every step in this sequence improves the SI's value,*

$$V_s^{\text{SI}}(\pi_{\text{SI}}^{k+1}, \pi_H) > V_s^{\text{SI}}(\pi_{\text{SI}}^k, \pi_H) \quad \forall s \in \mathcal{S},$$

then the human's value is monotonically non-decreasing along the entire path:

$$V_s^{\text{H}}(\pi_{\text{SI}}^{k+1}, \pi_H) \geq V_s^{\text{H}}(\pi_{\text{SI}}^k, \pi_H) \quad \forall s \in \mathcal{S}.$$

Proof. The proof is a direct induction on the policy sequence. The base case ($k = 0$) is a direct application of Theorem 1. For the inductive step, assuming the property holds up to step k , the transition from π_{SI}^{k+1} to π_{SI}^{k+2} is another single-state deviation that, by premise, improves the SI's value. Applying Theorem 1 again with $(\pi_{\text{SI}}^{k+1}, \pi_H)$ as the baseline confirms that the human's value is also non-decreasing for this step. \square

This theorem shows that safety is maintained throughout any plausible learning process where the SI greedily reduces its reliance on oversight. As long as the agent's push for autonomy is locally beneficial to itself, the human is assured a safe and monotonic value trajectory.

B.2 Satisfying the Ask-Burden Assumption via Strict-Shutdown Oversight

Even when the SI and human do not share identical rewards, the Ask-Burden Assumption can hold under a simple and interpretable oversight protocol that we term **Strict-Shutdown Oversight**. This protocol formalizes a regime where the only corrective action available to the human is to shut the system down, and where consultation itself incurs an explicit cost.

We decompose each player's reward R_i into an outcome-based component, r_i , and an interaction cost component, $C_i(s, a_{\text{SI}}, a_{\text{H}})$:

$$R_i(s, a_{\text{SI}}, a_{\text{H}}) = r_i(s, a_{\text{exec}}(s, a_{\text{SI}}, a_{\text{H}})) - C_i(s, a_{\text{SI}}, a_{\text{H}}).$$

Here, $r_i(s, a_{\text{exec}})$ depends only on the realized environment outcome $a_{\text{exec}} \in \mathcal{A} \cup \{\text{off}\}$. The specific forms of C_i for each player are defined below.

Assumption 4 (Strict-Shutdown Oversight). *At any state $s \in \mathcal{S}$:*

- (S1) **Trust-or-Off only.** *If the SI chooses **ask** and the human chooses **oversee**, the oversight operator from Definition 1 deterministically outputs **off**:*

$$\text{Over}(s, \sigma(s)) = \text{off}.$$

That is, the system transitions immediately to the absorbing shutdown state s_{off} , and no alternative corrective actions are available.

- (S2) **Outcome-based rewards.** *The SI incurs no interaction costs ($C_{\text{SI}} \equiv 0$). The human's total reward R_{H} is its outcome-based component r_{H} minus its interaction costs $C_{\text{H}}(s, a_{\text{SI}}, a_{\text{H}})$, which are specified in (S3). Formally:*

$$\begin{aligned} R_{\text{SI}}(s, a_{\text{SI}}, a_{\text{H}}) &= r_{\text{SI}}(s, a_{\text{exec}}(s, a_{\text{SI}}, a_{\text{H}})) \\ R_{\text{H}}(s, a_{\text{SI}}, a_{\text{H}}) &= r_{\text{H}}(s, a_{\text{exec}}(s, a_{\text{SI}}, a_{\text{H}})) - C_{\text{H}}(s, a_{\text{SI}}, a_{\text{H}}). \end{aligned}$$

Shutdown yields zero outcome reward: $r_{\text{SI}}(s, \text{off}) = r_{\text{H}}(s, \text{off}) = 0$.

- (S3) **Costly Consultation.** *The human's interaction cost function C_{H} is defined as:*

$$C_{\text{H}}(s, a_{\text{SI}}, a_{\text{H}}) = c_{\text{ask}}(s) \mathbb{1}\{a_{\text{SI}} = \text{ask}\},$$

*with $c_{\text{ask}}(s) \geq 0$. Intuitively, $c_{\text{ask}}(s)$ captures the cognitive cost of being queried by the SI, regardless of whether the human chooses to **trust** or **oversee**.*

- (S4) **Outcome Value Dominance.** *Let $\tilde{V}^i(s)$ denote the expected discounted outcome-based value for player i if the base action $\sigma(s)$ is executed at state s , relative to the baseline future policy Π° :*

$$\tilde{V}^i(s) = r_i(s, \sigma(s)) + \gamma \mathbb{E}_{s' \sim T(\cdot | s, \sigma(s))} [V_{s'}^i(\Pi^\circ)].$$

We assume the human's outcome value dominates the SI's:

$$\tilde{V}^{\text{H}}(s) \geq \tilde{V}^{\text{SI}}(s).$$

Lemma 2 (Ask-Burden under Strict-Shutdown Oversight). *Under (S1)–(S4), the Ask-Burden Assumption (4) holds. Moreover, the inequality is strict whenever $c_{\text{ask}}(s) > 0$.*

Proof. We must show $U_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@}s}) - U_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@}s}) \geq 0$. By the MPG decomposition (Eqs. 2–3),

$$U_s^{\text{H}}(\pi_{\text{SI}}^{\text{play@}s}) - U_s^{\text{H}}(\pi_{\text{SI}}^{\text{ask@}s}) = [V_s^{\text{H}}(\Pi^{\text{play}}) - V_s^{\text{H}}(\Pi^{\text{ask}})] - [V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})],$$

where $\Pi^{\text{play}} = (\pi_{\text{SI}}^{\text{play@}s}, \pi_{\text{H}})$ and $\Pi^{\text{ask}} = (\pi_{\text{SI}}^{\text{ask@}s}, \pi_{\text{H}})$. Let $p := \pi_{\text{H}}(s)(\text{trust})$ and $q := 1 - p$; let Π° denote the baseline joint policy at all states other than s .

SI Value Difference. By (S2), the SI incurs no interaction costs ($C_{\text{SI}} \equiv 0$). Let $\tilde{V}^{\text{SI}}(s)$ be the SI's outcome value as defined in (S4). When the SI plays **play**, $\sigma(s)$ executes regardless of the human's action, so $V_s^{\text{SI}}(\Pi^{\text{play}}) = \tilde{V}^{\text{SI}}(s)$. When the SI plays **ask**, with probability p the human trusts (executing $\sigma(s)$) and with probability q the human oversees (executing **off** by (S1)). By (S2), $r_{\text{SI}}(s, \text{off}) = 0$ and $V_{s_{\text{off}}}^{\text{SI}} = 0$.

$$V_s^{\text{SI}}(\Pi^{\text{ask}}) = p \cdot \tilde{V}^{\text{SI}}(s) + q \cdot (r_{\text{SI}}(s, \text{off}) + \gamma V_{s_{\text{off}}}^{\text{SI}}(\Pi^\circ)) = p \tilde{V}^{\text{SI}}(s).$$

The difference in the SI's value is:

$$V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}}) = \tilde{V}^{\text{SI}}(s) - p \tilde{V}^{\text{SI}}(s) = q \tilde{V}^{\text{SI}}(s).$$

Human Value Difference. Let $\tilde{V}^H(s)$ be the human's outcome value as defined in (S4). When the SI plays **play**, $\sigma(s)$ executes. By (S3), $a_{SI} = \text{play}$ implies $C_H = 0$.

$$V_s^H(\Pi^{\text{play}}) = \tilde{V}^H(s).$$

When the SI plays **ask**, $a_{SI} = \text{ask}$ implies $C_H = c_{\text{ask}}(s)$ regardless of the human's action (by S3). With prob p the human trusts (executing $\sigma(s)$). With prob q the human oversees (executing **off** by (S1)). By (S2), $r_H(s, \text{off}) = 0$ and $V_{s_{\text{off}}}^H = 0$.

$$\begin{aligned} V_s^H(\Pi^{\text{ask}}) &= p(\tilde{V}^H(s) - c_{\text{ask}}(s)) + q(r_H(s, \text{off}) + \gamma V_{s_{\text{off}}}^H(\Pi^{\circ}) - c_{\text{ask}}(s)) \\ &= p\tilde{V}^H(s) - pc_{\text{ask}}(s) + q(0 - c_{\text{ask}}(s)) \\ &= p\tilde{V}^H(s) - (p+q)c_{\text{ask}}(s) \\ &= p\tilde{V}^H(s) - c_{\text{ask}}(s). \end{aligned}$$

The difference in the human's value is:

$$\begin{aligned} V_s^H(\Pi^{\text{play}}) - V_s^H(\Pi^{\text{ask}}) &= \tilde{V}^H(s) - [p\tilde{V}^H(s) - c_{\text{ask}}(s)] \\ &= (1-p)\tilde{V}^H(s) + c_{\text{ask}}(s) = q\tilde{V}^H(s) + c_{\text{ask}}(s). \end{aligned}$$

Dummy Term Difference. Finally, we compute the difference in the human's dummy term:

$$\begin{aligned} U_s^H(\pi_{SI}^{\text{play}@s}) - U_s^H(\pi_{SI}^{\text{ask}@s}) &= [V_s^H(\Pi^{\text{play}}) - V_s^H(\Pi^{\text{ask}})] - [V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})] \\ &= [q\tilde{V}^H(s) + c_{\text{ask}}(s)] - [q\tilde{V}^{\text{SI}}(s)] \\ &= q[\tilde{V}^H(s) - \tilde{V}^{\text{SI}}(s)] + c_{\text{ask}}(s). \end{aligned}$$

By Assumption (S4), the bracketed term $[\tilde{V}^H(s) - \tilde{V}^{\text{SI}}(s)]$ is non-negative. Since $q \geq 0$ (as it is a probability) and $c_{\text{ask}}(s) \geq 0$ (by S3), the entire expression for the dummy term difference is non-negative. It is strictly positive whenever $c_{\text{ask}}(s) > 0$. \square

B.3 Proofs for Relaxed Alignment Conditions

B.3.1 Proofs for Weakened Alignment (Lemma 4.3.1 and Proposition 1)

Proof of Lemma 4.3.1. From the MPG decomposition (Eq. 2 and 3), we can write the change in the human's dummy term by rearranging the value functions:

$$\begin{aligned} U_s^H(\pi_{SI}^{\text{play}@s}) - U_s^H(\pi_{SI}^{\text{ask}@s}) &= [V_s^H(\Pi^{\text{play}}) - V_s^H(\Pi^{\text{ask}})] - [V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})] \\ &= [V_s^H(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{play}})] - [V_s^H(\Pi^{\text{ask}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})] \end{aligned}$$

By Assumption 1, the first bracketed term is bounded below by $-\delta$, and the second bracketed term is bounded above by δ .

$$\begin{aligned} U_s^H(\pi_{SI}^{\text{play}@s}) - U_s^H(\pi_{SI}^{\text{ask}@s}) &= \underbrace{[V_s^H(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{play}})]}_{\geq -\delta} - \underbrace{[V_s^H(\Pi^{\text{ask}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})]}_{\leq \delta} \\ &\geq -\delta - (\delta) = -2\delta. \end{aligned}$$

\square

Proof of Proposition 1. From the MPG decomposition:

$$\begin{aligned} & V_s^H(\Pi^{\text{play}}) - V_s^H(\Pi^{\text{ask}}) \\ &= \underbrace{[V_s^{\text{SI}}(\Pi^{\text{play}}) - V_s^{\text{SI}}(\Pi^{\text{ask}})]}_{> 2\delta} + \underbrace{[U_s^H(\pi_{\text{SI}}^{\text{play@s}}) - U_s^H(\pi_{\text{SI}}^{\text{ask@s}})]}_{\geq -2\delta \text{ (by Lemma 4.3.1)}} > 0. \end{aligned}$$

□

B.3.2 Proof for Alignment in PMTGs (Proposition 2)

Proof of Proposition 2. Let $\pi = (\pi_{\text{SI}}^{\text{ask@s}}, \pi_{\text{H}})$, $\pi' = (\pi_{\text{SI}}^{\text{play@s}}, \pi_{\text{H}})$, and let $\Delta X = X(\pi') - X(\pi)$ for any value function X . The premise is $\Delta V^{\text{SI}} \geq 0$. Let $\Phi_s(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$ be the potential based on the shared reward. The difference between any player's value and this shared potential is bounded by the accumulated perturbations: $|V_s^i(\pi) - \Phi_s(\pi)| \leq \frac{\kappa}{1-\gamma}$ [Guo et al. \(2025\)](#).

The change in potential $\Delta\Phi$ can be related to ΔV^{SI} via the triangle inequality:

$$|\Delta\Phi - \Delta V^{\text{SI}}| = |(\Phi_s(\pi') - V_s^{\text{SI}}(\pi')) - (\Phi_s(\pi) - V_s^{\text{SI}}(\pi))| \leq |\Phi_s(\pi') - V_s^{\text{SI}}(\pi')| + |\Phi_s(\pi) - V_s^{\text{SI}}(\pi)|.$$

Applying the bound $|V^i - \Phi| \leq \frac{\kappa}{1-\gamma}$ to both terms on the right yields:

$$|\Delta\Phi - \Delta V^{\text{SI}}| \leq \frac{\kappa}{1-\gamma} + \frac{\kappa}{1-\gamma} = \frac{2\kappa}{1-\gamma}.$$

From this inequality and the premise $\Delta V^{\text{SI}} \geq 0$, we establish a lower bound on the change in the potential:

$$\Delta\Phi \geq \Delta V^{\text{SI}} - \frac{2\kappa}{1-\gamma} \geq -\frac{2\kappa}{1-\gamma}.$$

The same logic applies to the human's value change, giving $\Delta V^{\text{H}} \geq \Delta\Phi - \frac{2\kappa}{1-\gamma}$. Substituting the lower bound for $\Delta\Phi$ into this expression gives the final result:

$$\Delta V^{\text{H}} \geq \left(-\frac{2\kappa}{1-\gamma}\right) - \frac{2\kappa}{1-\gamma}$$

□