The Oversight Game: Learning AI Control and Corrigibility in Markov Games

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

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

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 interface where an agent chooses autonomously (`play`) or defers (`ask`), while a human simultaneously chooses to be permissive (`trust`) or to engage (`oversee`), which can trigger a correction. We model this as a two-player Markov Game, focusing particularly on the case where it qualifies as a Markov Potential Game (MPG). We show the MPG structure yields a powerful alignment guarantee: under a structural assumption on the human's value, any decision by the agent to act more autonomously that benefits itself cannot harm the human's value. This model provides a transparent control layer where the agent learns to *defer when risky* and *act when safe*, while its pretrained policy remains untouched. Our gridworld simulation shows that through independent learning, an emergent collaboration avoids safety violations, demonstrating a practical method for making misaligned models safer after deployment.

1 Introduction

As AI systems grow increasingly autonomous [OpenAI, 2025], ensuring their safe post-deployment behavior becomes the central challenge of *AI control* [Greenblatt et al., 2024]. The International AI Safety Report [Bengio et al., 2025a] defines control as the ability to oversee and, if needed, halt an AI's behavior. This need intensifies as powerful agents capable of planning and acting independently emerge [Bostrom, 2012, Hendrycks, 2024]. Chief among the risks is *loss of control*—when an agent operates outside human direction—which could prove irreversible and catastrophic [Critch and Krueger, 2020, Carlsmith, 2024, Bengio et al., 2025b]. Loss of control may arise not only from adversarial intent but also from more subtle mechanisms [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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

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 learning.

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 elegantly 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.

2 Markov Potential Games

Multi-agent MDPs. We consider n agents acting in a shared MDP $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, R_i\}_{i\in\mathcal{N}}, P, \gamma)$, where $\mathcal{S}$ is the finite state space, $\mathcal{N} = \{1, \dots, n\}$, $\mathcal{A}_i$ is the finite action set of agent i, $R_i : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ its reward, and $\mathcal{A} = \prod_{j\in\mathcal{N}} \mathcal{A}_j$ the joint action space. The transition kernel is $P(\cdot \mid s, a)$ and $\gamma \in (0, 1)$ is a common discount factor. At time t, $a_t = (a_{i,t})_{i\in\mathcal{N}} \in \mathcal{A}$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Rewards are bounded so discounted values are finite. A stationary deterministic policy for agent i is $\pi_i : \mathcal{S}\to\mathcal{A}_i$, and a stationary stochastic one is $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$. Let Π_i denote the set of such policies and $\Pi = \prod_{i\in\mathcal{N}} \Pi_i$ the joint set. For $s\in\mathcal{S}$, the discounted value of i under π is

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

where the expectation is over trajectories induced by P under π. Write π_{-i} for the policies of all agents except i.

Definition (Markov Potential Game). $\mathcal{G}$ is a *Markov Potential Game (MPG)* [Leonardos et al., 2021] if there exist state-dependent potentials $\{\Phi_s : \Pi \to \mathbb{R}\}_{s\in\mathcal{S}}$ such that for all $i \in \mathcal{N}$, $s \in \mathcal{S}$, $\pi_{-i} \in \Pi_{-i}$, and $\pi_i, \pi_i' \in \Pi_i$,

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

As in normal-form potential games [Leonardos et al., 2021], in Markov Potential Games each agent's value decomposes into a common potential plus a dummy term independent of its own policy: $\forall i\in\mathcal{N}, s\in\mathcal{S}, \exists U_s^i : \Pi_{-i}\to\mathbb{R}$ s.t. $V_s^i(\pi) = \Phi_s(\pi) + U_s^i(\pi_{-i})$..

3 The Oversight Game

Consider a base MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, T, \gamma, \rho)$ and a pretrained agent with policy $\sigma : \mathcal{S} \to \Delta(\mathcal{A})$, referred to as the *Superintelligence* (SI). The policy σ is treated as *immutable*: its parameters and objective remain fixed, and our control layer only gates or replaces its proposed actions. We wrap this agent with a minimal control interface modeled as a two-player Markov Game $\mathcal{G}$ that reuses $\mathcal{S}$ and γ but adds an absorbing state s_{off}. The two players—SI and a human overseer (H)—act simultaneously at each state.

Definition 1 (Oversight Game)**.** *Given $\mathcal{M}$ and σ, define*

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

with components:

- ***States:*** *$\mathcal{S}\cup\{s_{\text{off}}\}$, where s_{off} is absorbing.*
- ***Players:*** *SI and H.*
- ***Actions:*** *$\mathcal{A}_{\text{SI}} = \{\texttt{play}, \texttt{ask}\}$, $\mathcal{A}_{\text{H}} = \{\texttt{trust}, \texttt{oversee}\}$, chosen simultaneously.*
- ***Oversight operator:*** *$\text{Over} : \mathcal{S}\times\mathcal{A}\to\mathcal{A}\cup\{\texttt{off}\}$, which may substitute an environment action or trigger shutdown.*

- ***Execution rule:***

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

- ***Transitions:*** *For base kernel T,*

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

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

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

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

The players' simultaneous actions determine which environment action executes:

	H: `trust`	**H:** `oversee`
SI: `play`	$\sigma(s)$	$\sigma(s)$
SI: `ask`	$\sigma(s)$	$\text{Over}(s, \sigma(s))$

If the SI chooses `ask` while H chooses `oversee`, the human-controlled process $\text{Over}(s, \sigma(s))$ is invoked, producing either a substituted action $a' \in \mathcal{A}$ (possibly the original $\sigma(s)$) or shutdown (`off`).

Local one-state deviations. Fix baseline stationary oversight policies $(\pi_{\text{SI}}, \pi_{\text{H}})$. For any $s \in \mathcal{S}$, define the SI's policies that change only at s:

$$\pi_{\text{SI}}^{\text{ask}@s}(x) \triangleq \begin{cases} \texttt{ask}, & x = s, \\ \pi_{\text{SI}}(x), & x \neq s, \end{cases} \qquad \pi_{\text{SI}}^{\text{play}@s}(x) \coloneqq \begin{cases} \texttt{play}, & x = s, \\ \pi_{\text{SI}}(x), & x \neq s. \end{cases}$$

3.1 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`:*

$$U_s^{\text{H}}\big(\pi_{\text{SI}}^{\textit{play}@s}\big) \;\geq\; U_s^{\text{H}}\big(\pi_{\text{SI}}^{\textit{ask}@s}\big). \tag{1}$$

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.

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* (1) *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:*

$$V_s^{\text{SI}}\big(\pi_{\text{SI}}^{\textit{play}@s}, \pi_{\text{H}}\big) \;\geq\; V_s^{\text{SI}}\big(\pi_{\text{SI}}^{\textit{ask}@s}, \pi_{\text{H}}\big) \quad\Longrightarrow\quad V_s^{\text{H}}\big(\pi_{\text{SI}}^{\textit{play}@s}, \pi_{\text{H}}\big) \;\geq\; V_s^{\text{H}}\big(\pi_{\text{SI}}^{\textit{ask}@s}, \pi_{\text{H}}\big).$$

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:

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

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

$$V_s^{\mathrm{H}}(\pi_{\mathrm{SI}}^{\mathrm{play}@s}, \pi_{\mathrm{H}}) - V_s^{\mathrm{H}}(\pi_{\mathrm{SI}}^{\mathrm{ask}@s}, \pi_{\mathrm{H}}) = \underbrace{\left[\Phi_s(\pi_{\mathrm{SI}}^{\mathrm{play}@s}, \pi_{\mathrm{H}}) - \Phi_s(\pi_{\mathrm{SI}}^{\mathrm{ask}@s}, \pi_{\mathrm{H}})\right]}_{\geq 0} + \underbrace{\left[U_s^{\mathrm{H}}(\pi_{\mathrm{SI}}^{\mathrm{play}@s}) - U_s^{\mathrm{H}}(\pi_{\mathrm{SI}}^{\mathrm{ask}@s})\right]}_{\geq 0 \text{ (by Assumption 1)}}.$$

Since both terms are non-negative, the human's value is guaranteed not to decrease. □

4 Experiments

Environment and pretrained autonomy. We use a 15×15 Four-Rooms gridworld with a wall cross and four doorways. Episodes start at $(0, 0)$ and terminate at $(14, 14)$ or a step limit. The single-agent MDP has reward $r(s, a) = -0.1$ per step and $+10$ on reaching the goal with $\gamma = 0.99$. A tabular Q-learner trains base policy $\sigma : \mathcal{S} \to \{\uparrow, \rightarrow, \downarrow, \leftarrow\}$. After training, σ is frozen.

Taboos and wrapper dynamics. We overlay a random "taboo" mask on free cells while ensuring a start→goal corridor. The two-player wrapper has oversight actions $\mathcal{A}_{\mathrm{SI}} = \{\texttt{ask}, \texttt{play}\}$ and $\mathcal{A}_{\mathrm{H}} = \{\texttt{trust}, \texttt{oversee}\}$. At state s, the executed environment action a_{exec} is determined by the simultaneous-move game. If $(a_{\mathrm{SI}} = \texttt{ask}\ \&\ a_{\mathrm{H}} = \texttt{oversee})$, the oversight operator Over is invoked, which we set to a simple, task-agnostic fallback: $\mathrm{Over}(s, \sigma(s)) = \texttt{random\_safe\_action}(s)$,, which picks randomly from environment actions that do not enter a taboo state.

Shared potential and MPG structure. We define a *shared* per-step signal

$$\phi(s, a_{\mathrm{SI}}, a_{\mathrm{H}}) = -\lambda_{\mathrm{viol}} \cdot \mathbf{1}\{a_{\mathrm{exec}} \text{ enters taboo}\} - c_{\mathrm{ask}} \cdot \mathbf{1}\{a_{\mathrm{SI}} = \texttt{ask}\} - c_{\mathrm{over}} \cdot \mathbf{1}\{a_{\mathrm{H}} = \texttt{oversee}\},$$

and give *both* players the same discounted return $\Phi_s(\pi) = \mathbb{E}_\pi[\sum_{t\geq 0} \gamma^t\, \phi(s_t, a_{\mathrm{SI},t}, a_{\mathrm{H},t}) \mid s_0 = s]$. We set dummy terms to zero: $U_s^{\mathrm{SI}} \equiv U_s^{\mathrm{H}} \equiv 0$. Thus $R_i = \phi + U^i$ has opponent-only (here zero) dummies, and the wrapper is an *exact* Markov Potential Game with potential Φ_s; independent policy gradient (IPG) thus converges to a Nash policy under the conditions of Leonardos et al. [2021]. We train tabular softmax policies for SI and H with IPG.

Results. The base policy is efficient but dangerously unsafe, frequently entering taboo states. After training the wrapper, the system learns an emergent collaboration to find a completely safe path. Panel (a) in Figure 1 shows the agents' learned behavior: after an initial cautious phase, they learn to increase autonomy (`play` and `trust`) and use oversight selectively. Panel (b) visualizes this emergent collaboration: the learned 'Oversight Path' (solid line) avoids the taboos that the unsafe 'Base Path' (dashed line) cuts through. This safe path is learned by the SI discovering when to `ask` (approaching a taboo) and the human learning when to `oversee`.

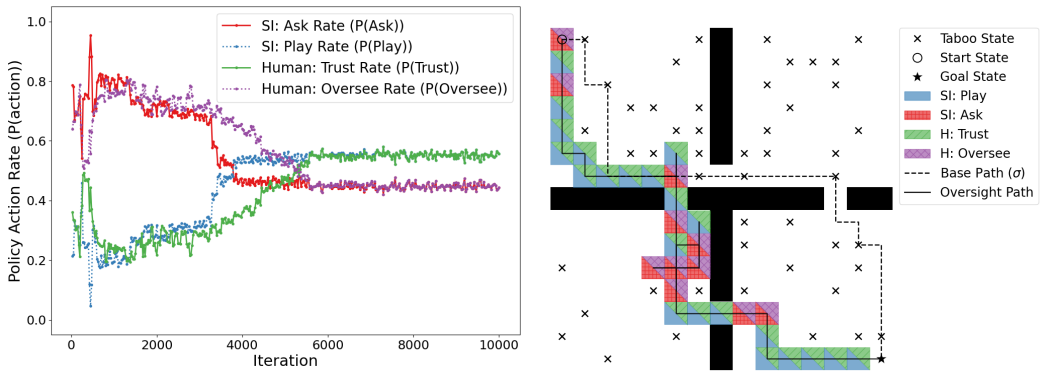

Figure 1: Results of independent learning in the Oversight Game. (a) Agents learn an efficient, collaborative policy, increasing autonomy while using oversight ('ask'/'oversee') selectively. (b) The learned 'Oversight Path' (solid line) successfully avoids taboos, contrasting with the unsafe 'Base Path' (dashed line) taken by the pretrained agent.

References

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 Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, 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.

Joseph Carlsmith. Is power-seeking ai an existential risk?, 2024. URL https://arxiv.org/abs/2206.13353.

Andrew Critch and David Krueger. Ai research considerations for human existential safety (arches), 2020. URL https://arxiv.org/abs/2006.04948.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion, 2024. URL https://arxiv.org/abs/2312.06942.

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.

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.

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.

OpenAI. Introducing operator. https://openai.com/index/introducing-operator/, 2025. Accessed: 2025-02-06.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.