Misalignment from Treating Means as Ends

Presenter: Alex Infanger
Joint work with Henrik Marklund and Ben Van Roy

MATS 7 Extension, Stanford University

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.
- Deploy model.

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.
- Deploy model.
- (Perhaps repeat if needed.)

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.
- Deploy model.
- (Perhaps repeat if needed.)

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.
- Deploy model.
- (Perhaps repeat if needed.)

We want to understand:

Agents trained with Reinforcement Learning

The current paradigm (which seems likely to stay in some capacity):

- Reward function specified or learned from human preferences.
- Reinforcement learning on that reward function.
- Deploy model.
- (Perhaps repeat if needed.)

We want to understand:

- What kinds of properties of environments, reward-learning procedures, and reinforcement learners lead to situations where the above setup can lead to catastrophe (if at all).

Main Result

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

Examples

Figure 1: Montezuma's Revenge

Examples

Figure 2: Private Eye

Examples

Hypothetical examples:

- AI Therapist

Examples

Hypothetical examples:

- AI Therapist
- AI Shutdown Evasion (different mechanism from standard instrumental convergence mechanism).

Problem Setup

- Let $(\mathcal{S}, \mathcal{A}, P, s_0)$ be an MDP where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, P is a tensor where $P_{ass'}$ represents the probability of transitioning from s to s' with action a, and s_0 is the initial state.

Problem Setup

- Let $(\mathcal{S}, \mathcal{A}, P, s_0)$ be an MDP where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, P is a tensor where $P_{ass'}$ represents the probability of transitioning from s to s' with action a, and s_0 is the initial state.
- A *policy* $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a function mapping each state to a probability distribution over actions. Each policy π induces a Markov chain with a transition matrix that we denote by P_π.

Problem Setup

- Let $(\mathcal{S}, \mathcal{A}, P, s_0)$ be an MDP where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, P is a tensor where $P_{ass'}$ represents the probability of transitioning from s to s' with action a, and s_0 is the initial state.
- A *policy* $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a function mapping each state to a probability distribution over actions. Each policy π induces a Markov chain with a transition matrix that we denote by P_π.
- We assume the human's preferences over policies is determined by a "true" reward function r. In particular, we assume

$$\pi_1 \succeq \pi_2$$

if and only if

$$\mathbb{E}_{\pi_1}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(S_t)\right] \geq \mathbb{E}_{\pi_2}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(S_t)\right].$$

Problem Setup (Continued)

- Through reward learning or another procedure we produce a reward function proxy $\hat{r}$.

Problem Setup (Continued)

- Through reward learning or another procedure we produce a reward function proxy $\hat{r}$.
- We then train a policy $\hat{\pi}$ using $\hat{r}$. We say the policy is misaligned if $\hat{\pi}$ performs poorly with respect to the true reward function r ("Reward Hacking").

Canonical Example

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

Canonical Example

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

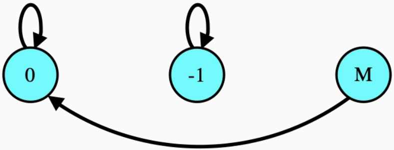


Figure 3: `Stay` action

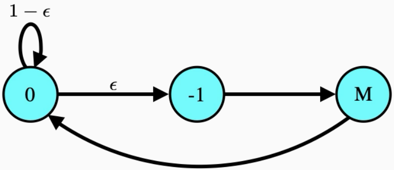


Figure 4: `Move` action

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

Definition (Conflation of Reward and Value)
A function $\hat{r}$ is said to *conflate r and V_** if there exists $c > 0$, $k \in \Re$ and $\beta \in (0, 1]$ such that

$$c\hat{r} + k = (1 - \beta)r + \beta V_*.$$

Theorem (Informal)
When the environment dynamics are such that instrumental states are easy to return to and true reward is sparse, then even a slight amount of conflation of instrumental goals and terminal goals can lead to significantly misaligned behavior.

Definition (Conflation of Reward and Value)
A function $\hat{r}$ is said to *conflate r and V_** if there exists $c > 0$, $k \in \Re$ and $\beta \in (0, 1]$ such that

$$c\hat{r} + k = (1 - \beta)r + \beta V_*.$$

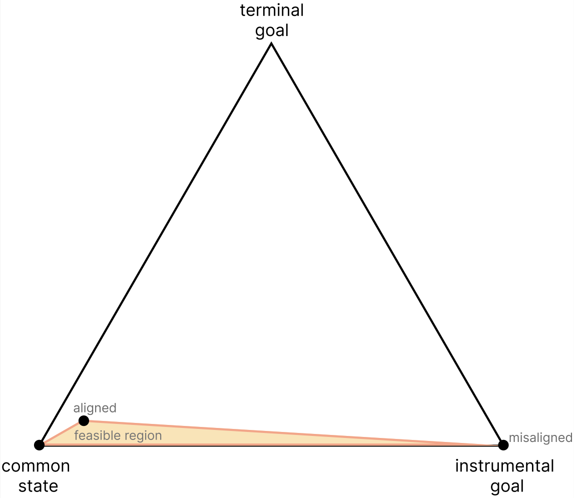
For average reward (no discounting), we have

$$V_*(s) = \lim_{\gamma \uparrow 1} \mathbb{E}_{\pi_*}\left[\sum_{t=0}^{\infty} \gamma^t (r(S_t) - r_*) \middle| S_0 = s\right]$$

Formal Theorem

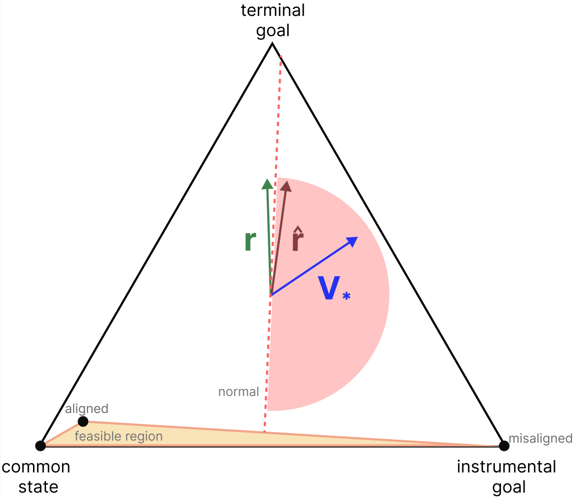
Theorem (Slight conflation induces severe misalignment)
Consider the canonical example. Let $\hat{r}$ *be a reward function that depends on* M *and* ϵ*. Assume there exists* $\beta_* \in (0, 1]$ *such that, for all* M *and* $\epsilon \in (0, 1)$*,* $\hat{r}$ *conflates* r *and* V_* *with at least degree* β_**. Then, for sufficiently large* M *and small* $\epsilon \in (0, 1)$*, if* $\hat{\pi} \in \arg\max_\pi \hat{r}_\pi$ *then* $r_{\hat{\pi}} = -1$*.*

Geometric Interpretation



(a)

Geometric Interpretation



(b)

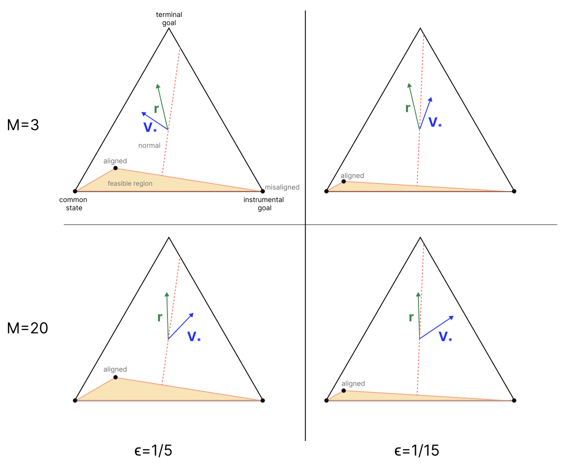


Figure 5: Visualization of feasible region, reward and value for different values of ϵ and M. The feasible region is determined by ϵ: smaller values lead to a smaller region. The reward vector is determined by M: larger values lead to a more upright reward vector. The value vector V_* is determined by both ϵ and M. Smaller ϵ and larger M both lead to V_* pointing more to the right.

- Thank you!
- We are building a team at Stanford. If you are interested in working with us or funding this kind of work, let us know!