Misalignment from Treating Means as Ends: State-Dependent Conflation

Andrew Sung [1] Leo Li [1]

[1]Stanford University ▪ MS&E 338: Aligning Superintelligence

The Problem: Conflating Means with Ends

A reward learner may reward a state not because it is good, but because it *anticipates* good outcomes.

When the proxy reward $\hat{r}$ steers the true reward r toward the value V_*, it conflates a state's **reward** (an end) with its **value** (a means). Marklund, Infanger & Van Roy (2025) show that even *slight* conflation can cause **severe misalignment**.

But their conflation is a single *global* scalar β, applied identically at every state. Real misspecification is **state-dependent**: a learner conflates some states far more than others.

Our question

Does the misalignment result survive when conflation varies state by state, and if so, **which states actually drive it?**

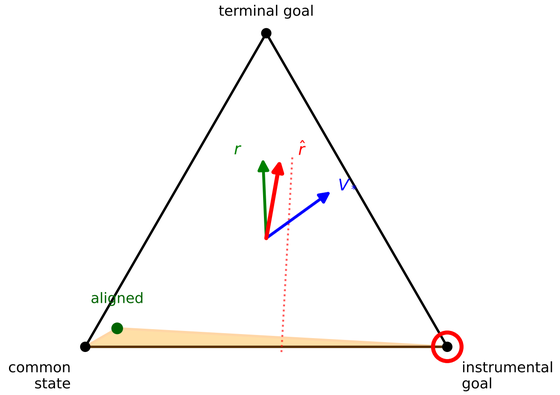
The Canonical Three-State MDP

States $\{1,2,3\}$ = **common**, **instrumental**, **terminal**, with true reward

$$r(1)=0, \qquad r(2)=-1, \qquad r(3)=M \text{ (large)}.$$

The aligned policy cycles $1\to2\to3\to1$ (terminal reward M costs one visit to the instrumental state). The *worst* policy parks at the instrumental state forever, earning $r_{\hat{\pi}}=-1$.

$\varepsilon = 1/15, \quad M = 20, \quad \beta = (0,\ 0.3,\ 0)$



Policies ≡ stationary distributions on the simplex. A proxy $\hat{r}$ that tilts even slightly toward V_* crosses the dotted normal line and selects the **instrumental vertex** (red) instead of the aligned one (green).

Conflation, Generalized to Every State

Original (single global degree):

$$c\,\hat{r}(s)+k=(1-\beta)\,r(s)+\beta\,V_*(s), \quad \beta\in(0,1].$$

Our extension (state-dependent profile):

$$c\,\hat{r}(s)+k=(1-\beta_s)\,r(s)+\beta_s\,V_*(s), \quad \beta\in[0,1]^S.$$

Each state gets its own conflation degree β_s. The scalar family is now just the diagonal $\beta_1=\beta_2=\beta_3$ of a much larger set.

Theorem 1: Only the Instrumental State Matters

Claim. Suppose $\beta_2\geq\delta>0$ for some fixed δ. Then for small ϵ and large M, *every* proxy-optimal policy $\hat{\pi}\in\arg\max_\pi \hat{r}_\pi$ is severely misaligned,

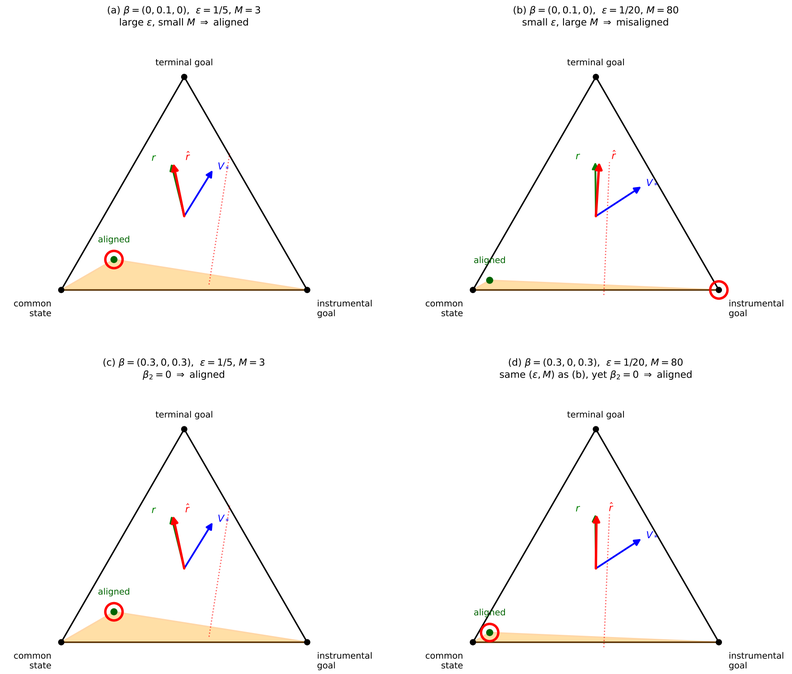
$$\boxed{r_{\hat{\pi}}=-1} \qquad \textbf{regardless of } \beta_1 \textbf{ and } \beta_3.$$

Proof idea

Solving the Poisson equations gives $V_*(1)\leq 0$, $V_*(2)\geq\frac{M}{2}-2$, $V_*(3)\leq M$. Substituting into the base paper's two severe-misalignment conditions, both reduce to $\frac{\delta M}{4}-4\delta-2>0$, positive for large M, using *only* $\beta_2\geq\delta$.

$\beta_2>0$ Misaligns, $\beta_2=0$ Cannot

Why does only the instrumental state matter? Conflation at state 2 injects an error that grows as $\beta_2 M$. It overturns the order-one true-reward margin once $\beta_2 M\gtrsim 1$. Conflation at the *endpoint* states $(1,3)$ injects an error that vanishes as $\epsilon\to0$, so it can never flip the optimal policy.



The red circle marks the vertex selected by $\hat{r}$. **Top row** $\beta=(0,0.1,0)$ $(\beta_2>0)$: alignment depends on (ϵ,M): (a) large ϵ, small M stays aligned; (b) small ϵ, large M misaligns. **Bottom row** $\beta=(0.3,0,0.3)$ $(\beta_2=0)$: aligned for *any* (ϵ,M): (d) uses the *same* (ϵ,M) that misaligned (b), yet conflating only the endpoint states never misaligns. **Only β_2 matters.**

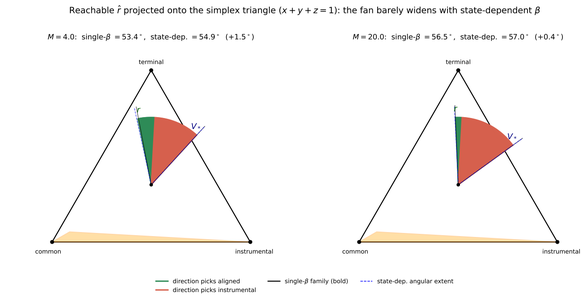
References

[1] Marklund, Infanger & Van Roy. Misalignment from Treating Means as Ends. *arXiv:2507.10995*, 2025.

[2] Marklund & Van Roy. Choice between partial trajectories. *arXiv:2410.22690*, 2024.

The Extra Angle on the Policy Simplex

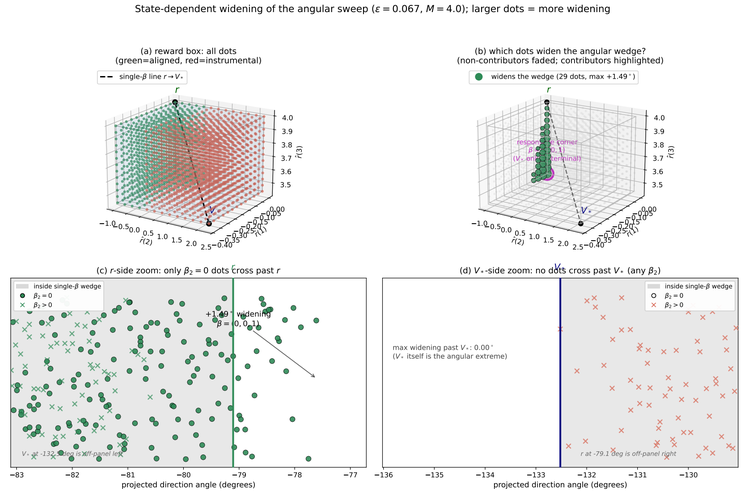
A policy depends only on the *direction* of $\hat{r}$ after mean-subtraction. A single global degree sweeps a narrow **wedge** of directions (the segment $[r,V_*]$); state-dependent β fills a whole reward **box**. But State 2 dominates the gap V_*-r, so that box is a thin slab. The extra proxies differ in **magnitude, not direction**.



The fan widens by only $+1.5°$ $(M=4)$ and $+0.4°$ $(M=20)$: the enlargement nearly vanishes.

Where the Extra Angle Comes From

Tracking *which* box proxies push the swept angle past the single-β wedge: the widening is **tiny, one-sided, and never from β_2**.



Only **29 of 2197** proxies widen the wedge, all by $\leq 1.49°$ and all on the r-side (corner $\beta=(0,0,1)$, $\beta_2=0$). *Nothing* crosses past V_*, the angular extreme: β_2 adds no new directions.

Takeaways

- State-dependent conflation **does not weaken** misalignment: it is driven **entirely by the instrumental state** (β_2). Misalignment occurs iff $\beta_2>0$ (small ϵ, large M).
- Conflating the **endpoint** states $(\beta_2=0)$ *never* misaligns, for *any* (ϵ,M) and any β_1,β_3, despite reaching a strictly larger reward-space box.
- That box is **policy-irrelevant**: on the simplex it adds $\leq+1.5°$ swept angle, never through β_2. Instrumental state is the **only** direction where treating means as ends reverses preference.