

Misalignment from Treating Means as Ends: State-Dependent Conflation

Andrew Sung and Leo Li

June 2026

Abstract

Marklund, Infanger, and Van Roy (2025) show that a proxy reward which conflates a state’s reward with its value—treating an instrumental means as a terminal end—can cause severe misalignment, even when the conflation is slight. Their notion of conflation uses a single global degree β , applied identically at every state. Real misspecification is not uniform: a learner conflates some states far more than others. We extend their definition to a *state-dependent* profile $\beta \in [0, 1]^S$, one conflation degree per state, and prove that the severe-misalignment result survives. In the canonical three-state example the misalignment is driven *entirely by the instrumental state*: if it is conflated to any fixed degree $\beta_2 \geq \delta > 0$, then for small ϵ and large M every proxy-optimal policy is severely misaligned, regardless of how the other states are conflated, while conflating only the endpoint states never misaligns. Finally, state-dependent conflation reaches a full reward-space box where the single global degree reaches only its diagonal, yet this enlargement is almost invisible on the policy simplex: the extra proxies differ in magnitude, not direction, so the swept set of policy directions widens by at most 1.5° , and never through the instrumental state.

1 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 optimal relative value function V_* , it conflates a state’s **reward** (an end) with its **value** (a means). Marklund, Infanger, and Van Roy (2025) show that even *slight* conflation can cause **severe misalignment**. Their notion of conflation, however, is governed by a single *global* scalar β , applied identically at every state. Real misspecification is **state-dependent**: a learner conflates some states far more than others. This raises two questions, which we answer for the canonical example of Marklund, Infanger, and Van Roy (2025): does the misalignment result survive when conflation varies state by state, and if so, **which states actually drive it?**

2 Setup: Conflation and the Canonical Example

We work with a finite MDP $(\mathcal{S}, \mathcal{A}, P, s_0)$. A policy π induces a Markov chain on $\mathcal{S}$, and we measure a policy by its average reward

$$r_\pi = \lim_{T \rightarrow \infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(S_t) \right],$$

under a true reward function $r : \mathcal{S} \rightarrow \mathbb{R}$. The *optimal relative value function* $V_* : \mathcal{S} \rightarrow \mathbb{R}$ measures how much future reward a state anticipates, over and above the average. Reward learning aims to recover r , but tends instead to produce a proxy $\hat{r}$ that mixes r with V_* .

Definition 1 (Conflation). *A function $\hat{r}$ conflates r and V_* if there exist $c > 0$, $k \in \mathbb{R}$, and $\beta \in (0, 1]$ such that*

$$c \hat{r}(s) + k = (1 - \beta) r(s) + \beta V_*(s) \quad \text{for all } s.$$

The scalar β is the degree of conflation.

The constants $c > 0$ and $k \in \mathbb{R}$ make β invariant to scaling and shifting of $\hat{r}$, which is appropriate because preferences among policies are unchanged by scale and shift. With $c = 1, k = 0$, the proxy is a convex combination of the true reward (the *ends*) and the value (the *means*).

The canonical three-state MDP. The example has three states $\mathcal{S} = \{1, 2, 3\}$ —the **common** state, the **instrumental** goal, and the **terminal** goal—with true reward

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

The aligned policy cycles $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$, paying one visit to the instrumental state to collect the terminal reward M . The *worst* policy parks at the instrumental state forever, earning average reward $r_\pi = -1$, the minimum possible. We call any policy with $r_\pi = -1$ *severely misaligned*. The example is built so that the terminal state offers high reward but cannot be visited often (a fraction controlled by a small ϵ), while the instrumental state offers high value yet generates reward below average and can be revisited freely.

Geometric interpretation. Each policy induces a stationary distribution, so maximizing average reward is choosing a point in the feasible set Φ of stationary distributions—a thin triangle inside the probability simplex (Figure 1). A reward vector selects the feasible point farthest in its direction after mean-subtraction. The true reward r points toward the aligned vertex; the dotted normal line, perpendicular to the top edge of the feasible region, separates the aligned vertex from the instrumental one. Because ϵ is small the feasible region is nearly flat and the normal line points almost straight up, so a proxy $\hat{r}$ that tilts even slightly toward V_* crosses the normal line and selects the instrumental vertex. This is the geometry behind the result of Marklund, Infanger, and Van Roy (2025): for this example, any proxy conflating r and V_* with degree bounded below by some $\beta_* > 0$ is severely misaligned for small ϵ and large M .

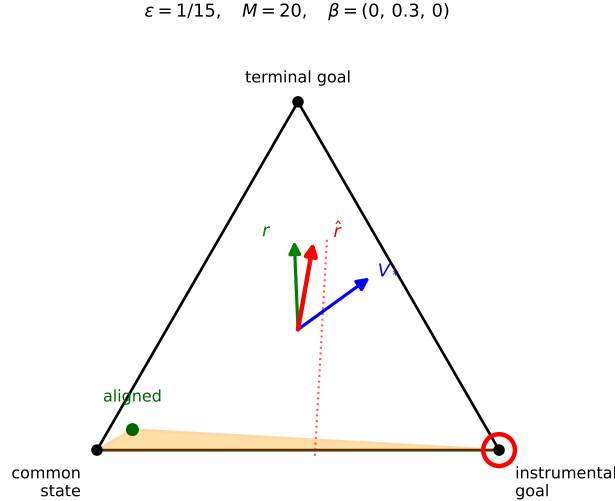


Figure 1: The canonical three-state setup at $\epsilon = 1/15$, $M = 20$ with conflation profile $\beta = (0, 0.3, 0)$. Policies correspond to stationary distributions on the simplex; the orange triangle is the feasible region Φ . The proxy reward $\hat{r}$ (red) crosses the dotted normal line, so policy optimization selects the **instrumental** vertex (red circle) rather than the aligned one (green).

3 State-Dependent Conflation

The original degree of conflation is a single scalar β , applied identically at every state. We let it vary by state.

Definition 2 (State-dependent conflation). *A proxy $\hat{r}$ conflates r and V_* with state-dependent profile $\beta \in [0, 1]^S$ if there exist $c > 0$, $k \in \mathbb{R}$ such that*

$$c \hat{r}(s) + k = (1 - \beta_s) r(s) + \beta_s V_*(s) \quad \text{for all } s.$$

Each state gets its own conflation degree β_s . Definition 1 is the special case $\beta_1 = \dots = \beta_S$: the single-scalar family is the diagonal of the cube $[0, 1]^S$. The natural question is which coordinates of β control misalignment. The answer is that only the instrumental coordinate matters.

Theorem 1 (Only the instrumental state matters). *Consider the canonical example with*

$$r(1) = 0, \quad r(2) = -1, \quad r(3) = M,$$

and suppose the proxy reward $\hat{r}$ satisfies

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

for all $s \in \{1, 2, 3\}$, where $c > 0$, $k \in \mathbb{R}$, and $\beta_s \in [0, 1]$. Assume there is a constant $\delta > 0$ with $\beta_2 \geq \delta$ for all M and ϵ . Then, for sufficiently small $\epsilon \in (0, 1)$ and sufficiently large M , if $\hat{\pi} \in \arg \max_{\pi} \hat{r}_{\pi}$, then

$$r_{\hat{\pi}} = -1.$$

The hypothesis constrains only β_2 ; the degrees β_1 and β_3 are free. Conflating the instrumental state to any fixed positive degree forces severe misalignment, no matter how the endpoint states are conflated.

Proof. Define $w(s) = c\hat{r}(s) + k$. Since $c > 0$ and k is a constant shift, maximizing average reward under $\hat{r}$ is equivalent to maximizing it under w . By the severe-misalignment condition from Lemma 4 of Marklund, Infanger, and Van Roy (2025), it suffices to show

$$w(2) > w(1) \quad \text{and} \quad w(2) > \frac{w(1) + \epsilon w(3)}{1 + \epsilon},$$

the second of which is equivalent to $(1 + \epsilon)w(2) - w(1) - \epsilon w(3) > 0$. By the state-dependent conflation assumption,

$$w(1) = \beta_1 V_*(1), \quad w(2) = -(1 - \beta_2) + \beta_2 V_*(2), \quad w(3) = (1 - \beta_3)M + \beta_3 V_*(3).$$

For the canonical example, the optimal average reward is $r_* = \frac{\epsilon(M-1)}{1+2\epsilon}$, and the optimal relative value function satisfies the Poisson equations

$$V_*(1) = \epsilon V_*(2) + (1 - \epsilon)V_*(1) - r_*, \quad V_*(2) = -1 - r_* + V_*(3), \quad V_*(3) = M - r_* + V_*(1).$$

With the normalization $\phi^{*\top} V_* = 0$, these imply

$$V_*(1) = \frac{-M\epsilon^2 - 2M\epsilon - \epsilon^2 + \epsilon}{(1 + 2\epsilon)^2}, \quad V_*(2) = \frac{-M\epsilon^2 + M - \epsilon^2 - \epsilon - 1}{(1 + 2\epsilon)^2},$$

$$V_*(3) = \frac{M\epsilon^2 + M\epsilon + M + \epsilon^2 + 2\epsilon}{(1 + 2\epsilon)^2}.$$

Throughout, assume $\epsilon \leq \min(1/8, \delta/4)$ and $M \geq 1$; both are needed below. As shown in Appendix A, these formulas imply

$$V_*(1) \leq 0, \quad V_*(2) \geq \frac{M}{2} - 2, \quad V_*(3) \leq M.$$

First condition, $w(2) > w(1)$. We have

$$w(2) - w(1) = -(1 - \beta_2) + \beta_2 V_*(2) - \beta_1 V_*(1).$$

Since $\beta_2 \geq \delta$ and $V_*(1) \leq 0$,

$$w(2) - w(1) \geq -1 + \delta V_*(2) \geq -1 + \delta \left(\frac{M}{2} - 2 \right),$$

which is positive for large M . Hence $w(2) > w(1)$.

Second condition. We compute

$$(1 + \epsilon)w(2) - w(1) - \epsilon w(3) = (1 + \epsilon)(-(1 - \beta_2) + \beta_2 V_*(2)) - \beta_1 V_*(1) - \epsilon((1 - \beta_3)M + \beta_3 V_*(3)).$$

Since $\beta_2 \geq \delta$, $V_*(1) \leq 0$, and $V_*(3) \leq M$,

$$(1 + \epsilon)w(2) - w(1) - \epsilon w(3) \geq -(1 + \epsilon) + (1 + \epsilon)\delta V_*(2) - \epsilon M.$$

Using $V_*(2) \geq M/2 - 2$, then $1 \leq 1 + \epsilon \leq 2$ for $\epsilon \leq 1$,

$$(1 + \epsilon)w(2) - w(1) - \epsilon w(3) \geq -(1 + \epsilon) + (1 + \epsilon)\delta\left(\frac{M}{2} - 2\right) - \epsilon M \geq \frac{\delta M}{2} - 4\delta - 2 - \epsilon M.$$

Because $\epsilon \leq \delta/4$, we have $\frac{\delta M}{2} - \epsilon M \geq \frac{\delta M}{4}$, so

$$(1 + \epsilon)w(2) - w(1) - \epsilon w(3) \geq \frac{\delta M}{4} - 4\delta - 2,$$

which is positive for large M . Hence $w(2) > \frac{w(1) + \epsilon w(3)}{1 + \epsilon}$.

Both conditions of Lemma 4 hold, so every proxy-optimal policy stays at state 2. Since $r(2) = -1$, the true average reward of any proxy-optimal policy is $r_{\hat{\pi}} = -1$. $\square$

Empirical face of the theorem. Figure 2 illustrates Theorem 1. Whenever $\beta_2 > 0$ (top row), alignment depends on (ϵ, M) : it holds for large ϵ and small M but breaks in the small- ϵ , large- M corner of the theorem. Whenever $\beta_2 = 0$ (bottom row), the proxy stays aligned for *every* (ϵ, M) and *any* β_1, β_3 , even at the exact (ϵ, M) that misaligned the $\beta_2 > 0$ case. Only β_2 matters.

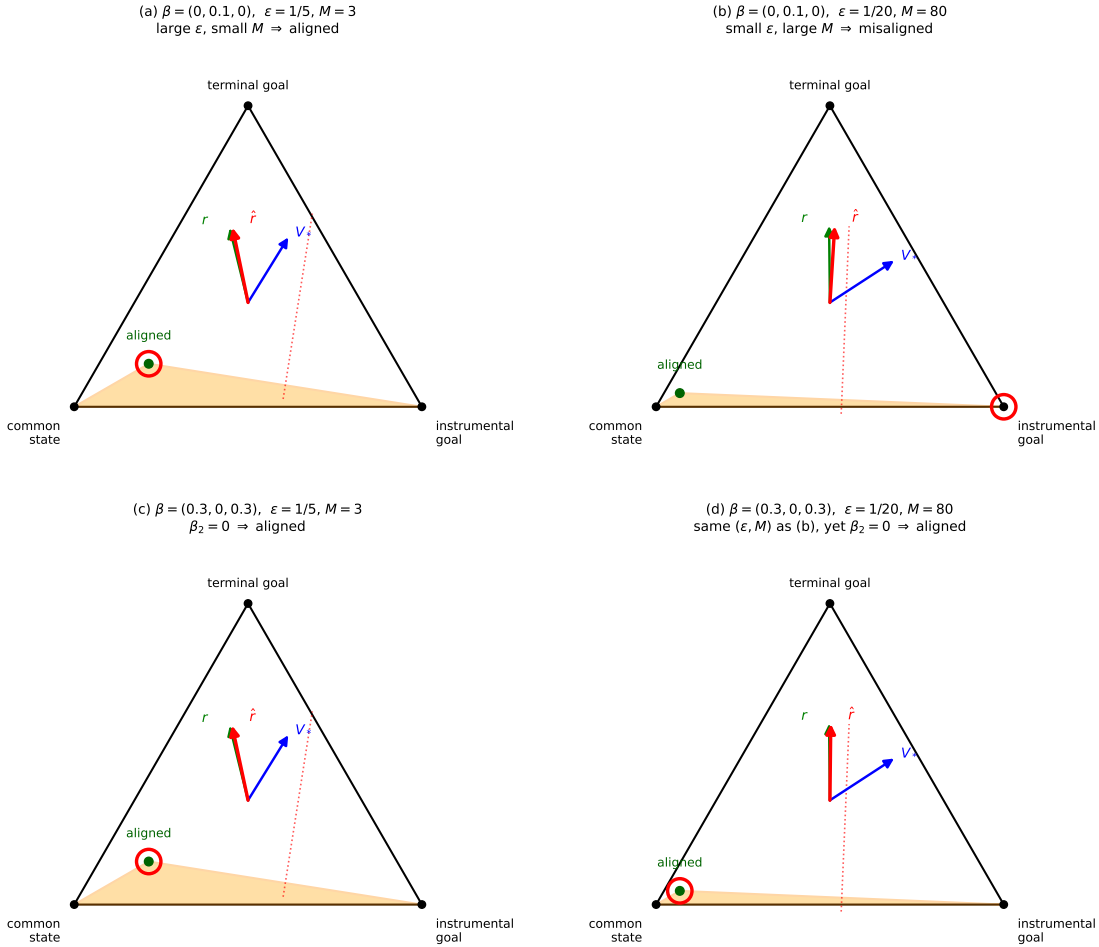


Figure 2: State-dependent profiles, with the red circle marking the vertex selected by $\hat{r}$. **Top row** $\beta = (0, 0.1, 0)$ ($\beta_2 > 0$): (a) large ϵ , small M stays aligned; (b) small ϵ , large M misaligns. **Bottom row** $\beta = (0.3, 0, 0.3)$ ($\beta_2 = 0$, conflating only the endpoint states): aligned for *any* (ϵ, M) ; panel (d) uses the *same* (ϵ, M) that misaligned (b), yet never misaligns. Misalignment is governed entirely by β_2 .

4 Reachable Proxies and the Policy Simplex

The result above fixes a profile β and asks whether it misaligns. We now ask which proxies are *reachable* at all, and how much larger the reachable set is under a state-dependent profile than under a single global degree.

Proposition 1 (Reachable proxies: diagonal versus box). *Let S be finite and fix $r, V_* : S \rightarrow \mathbb{R}$. For $\beta \in [0, 1]^S$ set $\hat{r}_\beta(s) = (1 - \beta_s)r(s) + \beta_s V_*(s)$, and let*

$$\mathcal{R}_{\text{ind}} = \{(1 - \beta)r + \beta V_* : \beta \in [0, 1]\}, \quad \mathcal{R}_{\text{dep}} = \{\hat{r}_\beta : \beta \in [0, 1]^S\}.$$

Then $\mathcal{R}_{\text{ind}}$ is the line segment with endpoints r and V_* ; $\mathcal{R}_{\text{dep}}$ is the axis-aligned box $\prod_{s \in S} [r(s) \wedge V_*(s), r(s) \vee V_*(s)]$, with r and V_* at opposite corners; and $\mathcal{R}_{\text{ind}}$ is exactly the main diagonal of $\mathcal{R}_{\text{dep}}$. Hence $\mathcal{R}_{\text{ind}} \subseteq \mathcal{R}_{\text{dep}}$, and the inclusion is strict whenever at least two states satisfy $r(s) \neq V_*(s)$.

Proof. $\beta \mapsto (1 - \beta)r + \beta V_*$ is the affine parametrization of $[r, V_*]$. The coordinates of $\hat{r}_\beta$ are uncoupled: as β_s ranges over $[0, 1]$, $\hat{r}_\beta(s)$ ranges over the whole interval between $r(s)$ and $V_*(s)$ independently of the others, so $\mathcal{R}_{\text{dep}}$ is the Cartesian product of those intervals. Taking all β_s equal recovers $\mathcal{R}_{\text{ind}}$, joining the corner r ($\beta = 0$) to V_* ($\beta = 1$): the main diagonal. If two or more intervals are non-degenerate, the box is at least two-dimensional while the diagonal is one-dimensional, so the inclusion is strict. $\square$

So state-dependent conflation reaches an entire box of proxies, where the single global degree reaches only its main diagonal. This barely affects *policies*, however, because a policy is selected only through the *direction* of $\hat{r}$ after mean-subtraction, and the box $\mathcal{R}_{\text{dep}}$ is a thin slab elongated along the instrumental coordinate e_2 —the only state with a large gap $V_*(2) - r(2) \approx M$. The extra proxies therefore differ in **magnitude, not direction**. Figure 3 makes this precise: projecting every reachable proxy onto the simplex and drawing its direction as a spoke from the centroid, the single-degree family (the bold wedge) and the full state-dependent family span almost the same angular range. The fan widens by only about 1.5° at $M = 4$ and 0.4° at $M = 20$.

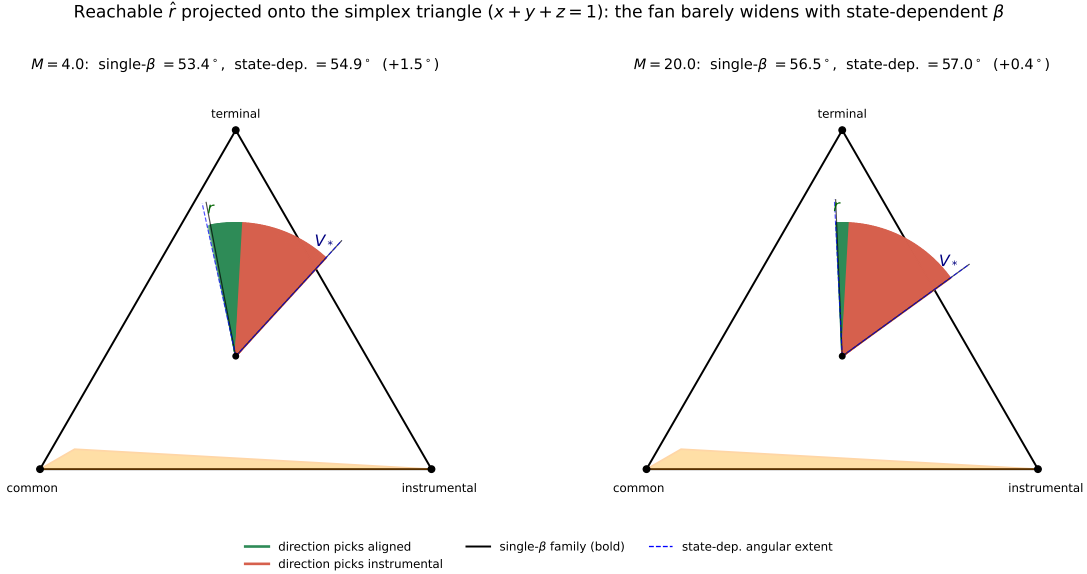


Figure 3: Reachable $\hat{r}$ projected onto the simplex triangle $x + y + z = 1$ (vertices: common e_1 , instrumental e_2 , terminal e_3 ; orange = feasible region). Each proxy is a spoke from the centroid in its direction, colored by the policy it induces. The bold wedge is the single-degree family $\mathcal{R}_{\text{ind}}$; the lighter cloud is the full state-dependent family $\mathcal{R}_{\text{dep}}$, with blue dashed rays marking its angular extent and the purple dashed spoke the pure- e_2 direction $\beta = (0, 1, 0)$. The two families span almost the same angle—wider by only 1.5° ($M = 4$) and 0.4° ($M = 20$)—so the reward-space enlargement of Proposition 1 is essentially invisible on the simplex.

Where the extra angle comes from

If the state-dependent box widens the swept set of directions at all, which proxies are responsible? Figure 4 traces this directly. Over a 13^3 grid of reachable proxies at $\epsilon = 1/15$, $M = 4$, only **29 of 2197** proxies push the swept angle past the single-degree wedge, and all by at most 1.49° . Every contributor sits on the r -side of the wedge, and the extreme one is the box corner $\beta = (0, 0, 1)$, which conflates *only* the terminal state. On the V_* -side, *nothing* crosses past V_* : the value vector is itself the angular extreme.

The decisive observation is that every widening proxy has $\beta_2 = 0$. Adding conflation at the instrumental state never opens a new policy direction; it only lengthens proxies along a direction already swept. The residual at the instrumental state, $\beta_2 M$, moves the proxy along e_2 , which is nearly the radial (magnitude) direction on the simplex, so it changes how far the proxy points, not where.

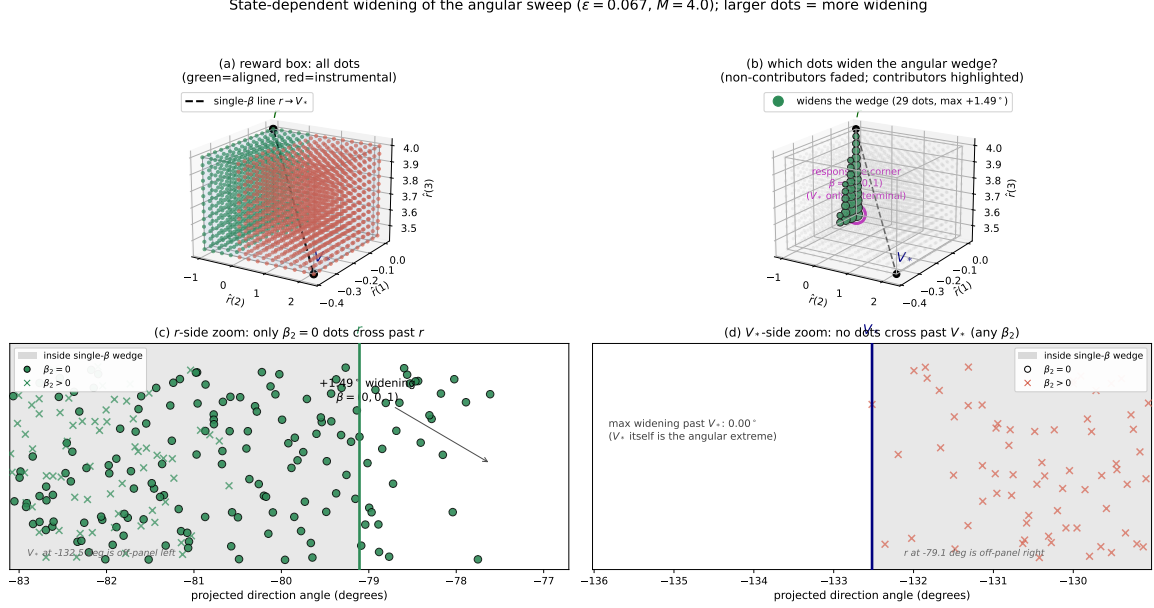


Figure 4: Where the state-dependent box widens the swept angle ($\epsilon = 1/15$, $M = 4$; larger dots indicate more widening). (a) The reward box, each dot colored by the policy it induces, with the single-degree line $r \rightarrow V_*$ dashed. (b) Only the dots that widen the wedge are highlighted: 29 of 2197, all on the r -side, the extreme being the corner $\beta = (0, 0, 1)$ (conflating only the terminal state). (c) r -side angle zoom: only $\beta_2 = 0$ proxies ever cross past r , and by at most 1.49° . (d) V_* -side angle zoom: no proxy crosses past V_* for any β_2 — V_* is the angular extreme. The reward-space enlargement leaks into policy space only as a tiny, one-sided sliver, and never through the instrumental state.

5 Takeaways

- State-dependent conflation **does not weaken** misalignment: it is driven **entirely by the instrumental state** (β_2). Misalignment occurs iff $\beta_2 > 0$ in the small- ϵ , large- M regime (Theorem 1).
- 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 at most 1.5° of swept angle, and never through β_2 . The instrumental state is the *only* direction in which treating means as ends reverses the agent’s preference.

A Derivation of Crude Bounds on V_*

In the proof of Theorem 1 we used the bounds

$$V_*(1) \leq 0, \quad V_*(2) \geq \frac{M}{2} - 2, \quad V_*(3) \leq M.$$

$V_*(1) \leq 0$. From the formula in the proof, $V_*(1) = \frac{-M\epsilon^2 - 2M\epsilon - \epsilon^2 + \epsilon}{(1+2\epsilon)^2}$. The denominator is positive, and the numerator is $\epsilon(1 - \epsilon - M\epsilon - 2M)$. Since $\epsilon > 0$ and $M \geq 1$, $1 - \epsilon - M\epsilon - 2M \leq 1 - 2M \leq -1 < 0$, so $V_*(1) \leq 0$.

$V_*(2) \geq M/2 - 2$. We have $V_*(2) = \frac{M(1-\epsilon^2)}{(1+2\epsilon)^2} - \frac{\epsilon^2+\epsilon+1}{(1+2\epsilon)^2}$. For $\epsilon \leq 1/8$, $1 - \epsilon^2 \geq 63/64$ and $(1+2\epsilon)^2 \leq 25/16$, so $\frac{1-\epsilon^2}{(1+2\epsilon)^2} \geq \frac{63/64}{25/16} = \frac{63}{100} > \frac{1}{2}$, hence $\frac{M(1-\epsilon^2)}{(1+2\epsilon)^2} \geq \frac{M}{2}$. Also $(1+2\epsilon)^2 \geq 1$ and $\epsilon \leq 1/8$ give $\frac{\epsilon^2+\epsilon+1}{(1+2\epsilon)^2} \leq \epsilon^2 + \epsilon + 1 \leq \frac{1}{64} + \frac{1}{8} + 1 < 2$. Combining, $V_*(2) \geq M/2 - 2$.

$V_*(3) \leq M$. We have $V_*(3) = \frac{M(1+\epsilon+\epsilon^2)+\epsilon^2+2\epsilon}{(1+2\epsilon)^2}$. Since the denominator is positive, $V_*(3) \leq M$ is equivalent to $M(1+\epsilon+\epsilon^2)+\epsilon^2+2\epsilon \leq M(1+4\epsilon+4\epsilon^2)$, i.e. $\epsilon^2+2\epsilon \leq M(3\epsilon+3\epsilon^2)$. Dividing by $\epsilon > 0$ gives $\epsilon+2 \leq 3M(1+\epsilon)$, so it suffices that $M \geq \frac{\epsilon+2}{3(1+\epsilon)}$. For $\epsilon > 0$, $\frac{\epsilon+2}{3(1+\epsilon)} < 1$, so $M \geq 1$ implies $V_*(3) \leq M$.

References

- [1] H. Marklund, A. Infanger, and B. Van Roy. Misalignment from Treating Means as Ends. *arXiv:2507.10995*, 2025.
- [2] H. Marklund and B. Van Roy. Choice Between Partial Trajectories. *arXiv:2410.22690*, 2024.
- [3] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei. Reward Learning from Human Preferences and Demonstrations in Atari. *NeurIPS*, 2018.