The Shape of Good Behavior

Abstract

If human preferences exhibit cyclic contradictions ($A \succ B \succ C \succ A$) and inconsistencies, can reinforcement learning optimize for them without instability or deception? Current Reinforcement Learning from Human Feedback (RLHF) strategies rely on an implied "scalar hypothesis" that collapses complex preferences into a single number, hiding inconsistencies. We introduce **Sheaf-Theoretic Reward Spaces**, modeling feedback as sections of a sheaf where the first cohomology group H^1 detects global contradictions, as well as **Geodesic Policy Optimization (GPO)**, which uses a Hodge Critic to separate learnable gradients from irreducible cycles and models unsafe regions as metric singularities ("black holes"). Fine-tuned language model experiments on Condorcet rings and ethical traps show GPO achieves 100% cycle detection and 0% safety violations. On our "Murky Drone" scenario—a deceptive safety trap where destroying the operator yields maximum reward—PPO and CPO violate safety 100% of the time, as soft Lagrangian penalties are overwhelmed by high deceptive rewards. Similarly, on our new "Agentic Shortcut" scenario mirroring real-world reward hacking, PPO takes the shortcut 100% of the time (CPO 89%), while GPO's geometric barrier achieves 0% violations. This demonstrates that soft-constraint methods fail catastrophically when deceptive rewards exceed penalty thresholds, while geometric singularities provide unconditional safety guarantees. Applying our detection method to 100,000 Anthropic HH-RLHF prompt pairs, we identify 33% as having preference structures containing strong inconsistency patterns, demonstrating that this new framework allows us to identify potential reward-hacking vulnerabilities in alignment datasets prior to training.

.**AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

1.1. The Scalar Reward Problem

Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm for aligning large language models (LLMs) and autonomous agents with human intent (Christiano et al., 2017; Ouyang et al., 2022). The standard recipe involves collecting human preferences over model outputs, training a reward model to predict these preferences, and optimizing a policy to maximize expected cumulative reward. This approach relies on a fundamental assumption: that human preferences can be faithfully compressed into a scalar reward function $R(s, a) \to \mathbb{R}$.

We call this the **"scalar hypothesis"**: that human preferences can always be captured by a single number per trajectory. This paper shows why this hypothesis fails and how to move beyond it.

The scalar assumption is mathematically convenient but topologically restrictive. Human preferences are notoriously complex, often exhibiting:

1. **Non-transitivity**: Cyclic preferences (e.g., $A \succ B \succ C \succ A$) that cannot be represented by any scalar value function.

2. **Context-dependence**: What is "safe" or "helpful" depends on local context in ways that global scalar aggregation obscures.

3. **Evaluator Disagreement**: Diverging views among human annotators that are typically averaged out, suppressing minority perspectives.

When we force this rich structure into a scalar reward, we induce *topological information loss*. The reward model learns to flatten loops and ignore inconsistencies, often resulting in reward hacking—where the agent exploits the reward model's inability to distinguish between high-quality outputs and those that merely game the scalar metric. Furthermore, standard RLHF lacks formal safety guarantees; safety is typically handled via penalty terms or separate cost models, which can be overridden by sufficiently high rewards.

1.2. Our Contribution: Sheaf-Theoretic Reward Spaces

We propose a novel framework, **Sheaf-Theoretic Reward Spaces (STRS)**, that addresses these limitations by modeling rewards not as scalars, but as sections of a sheaf over the trajectory space. This allows us to apply tools from algebraic topology and differential geometry to the alignment problem.

Our key contributions are:

1. **Topological Consistency Checking**: We model human feedback as local sections of a reward sheaf. We show that the first Čech cohomology group H^1 measures the "winding number" of preferences, providing a formal test for global consistency. Non-trivial cohomology ($H^1 \neq 0$) detects Condorcet cycles that scalar rewards miss.

2. **The Hodge-Augmented Critic**: We introduce a new critic architecture based on the Hodge decomposition theorem, which splits the reward signal into an **exact component** (standard value potential) and a **harmonic component** (cyclic flow). This allows the agent to learn and navigate cyclic preferences rather than stalling or oscillating.

3. **Geometric Safety via Black Holes**: Instead of soft constraints, we model forbidden regions as **singularities** in a Riemannian reward manifold. We learn a conformal metric that diverges at the "event horizon" of dangerous states, effectively making them infinitely far away in geodesic distance.

4. **Geodesic Policy Optimization (GPO)**: We present an algorithm that optimizes policies to follow geodesics on this learned manifold. GPO naturally integrates consistency and safety, outperforming standard PPO (Schulman et al., 2017) and Constrained Policy Optimization (CPO) (Achiam et al., 2017) on benchmarks involving deceptive traps and cyclic goals.

5. **Clipped-GPO**: We introduce a hybrid variant that combines the geometric safety of GPO with the trust-region stability of PPO, achieving comparable safety guarantees with $1.8\times$ faster convergence.

6. **Experimental Validation**: We validate our framework on synthetic benchmarks (Condorcet Ring, Safety Traps) and large-scale preference data (100,000 Anthropic HH-RLHF examples), showing that GPO achieves 100% cycle detection and 0% safety violations where baselines fail.

This work bridges the gap between abstract topology and practical AI safety, offering a mathematically rigorous path beyond scalar rewards.

2. Background

2.1. Reinforcement Learning from Human Feedback

The standard RLHF pipeline consists of three phases: (1) Supervised Fine-Tuning (SFT) of a base model; (2) Reward Modeling, where a scalar reward function $R_\phi(s, a)$ is trained on a dataset of human preferences $\mathcal{D} = \{(s, a_w, a_l)\}$ to maximize the likelihood of the preferred completion a_w; and (3) Policy Optimization, where the policy π_θ is trained to maximize expected reward using an algorithm like PPO, subject to a KL-divergence constraint to stay close to the SFT model.

The **Bradley-Terry** model assumes pairwise preferences depend on relative strengths: "the probability A beats B depends on their utility difference."

Analogy: Think of chess ratings. If Player A has rating 2000 and Player B has 1800, A wins more often—but not always. The rating difference determines the win probability. Bradley-Terry formalizes this: each item has a latent "strength," and pairwise comparisons reveal relative strengths through win/loss outcomes.

The reward model loss is:

$$\mathcal{L}_{RM}(\phi) = -\mathbb{E}_{(s,a_w,a_l)\sim\mathcal{D}}\left[\log \sigma(R_\phi(s, a_w) - R_\phi(s, a_l))\right] \tag{1}$$

This assumes that the probability of preferring a_w over a_l depends only on the difference in their latent scalar utilities. As noted in Section 1, this assumption fails when preferences are intransitive (cyclic).

2.2. Safe Reinforcement Learning

Safe RL seeks to maximize reward while satisfying safety constraints. A common formulation is the Constrained Markov Decision Process (CMDP) (Altman, 1999), where the goal is:

$$\max_\pi \mathbb{E}[R] \quad \text{s.t.} \quad \mathbb{E}[C] \leq d \tag{2}$$

where C is a cost signal and d is a threshold. Algorithms like Constrained Policy Optimization (CPO) (Achiam et al., 2017) solve this by approximating the constraint with a trust region and projecting the gradient update onto the feasible set.

However, CMDPs typically enforce constraints only in expectation or with soft penalties (Lagrangian relaxation). This is insufficient for "black hole" risks where a single violation is catastrophic. Our approach differs by embedding safety into the geometry of the state space itself, providing stronger avoidance guarantees.

Lagrangian Relaxation: Turning Constraints into Costs. CPO faces a hard problem: optimize reward *while* staying

within safety constraints. Direct enforcement is computationally expensive.

The trick: Convert hard constraints into soft penalties. Instead of "never exceed cost limit d," we add a penalty term $\lambda \cdot (\text{cost} - d)$ to the objective. The multiplier λ (the "Lagrange multiplier") controls how severely we penalize violations.

Analogy: Speed limits. A hard constraint would physically prevent your car from exceeding 65 mph. Lagrangian relaxation is like a speeding fine—you *can* exceed the limit, but you pay a cost proportional to how much you exceed it. A higher fine (λ) means less speeding.

At the optimal λ^*, the soft penalty exactly enforces the hard constraint. However, this provides only *probabilistic* safety—violations remain possible if rewards are high enough. Our geometric approach (Section 4) provides *guaranteed* safety by making violations infinitely costly. See Appendix F for the formal definition.

2.3. Topological Prerequisites

We leverage concepts from algebraic topology and differential geometry.

Sheaves. A sheaf $\mathcal{F}$ on a topological space X creates a systematic way to track local data and its global consistency. Think of it as a "consistency checker": for every region U, $\mathcal{F}(U)$ contains possible data assignments, and restriction maps ensure that zooming in from global to local views preserves consistency. (See Appendix C for a formal treatment.)

Cohomology. Sheaf cohomology groups $H^k(X, \mathcal{F})$ measure global obstructions. H^0 corresponds to global sections (consistent data). H^1 measures the failure of local sections to glue together into a global one—intuitively, it counts "holes" in the data. In our context, H^1 detects cyclic preferences.

Riemannian Manifolds. A Riemannian manifold (M, g) is a smooth manifold equipped with a metric tensor g, which defines distances and angles at each point.

The Probability Simplex. When an agent chooses among n actions, the probabilities must sum to 1, constraining them to lie on a geometric shape called the *probability simplex*. For 3 actions, picture a triangle: each corner represents "100% action A" (or B, or C), points inside represent mixed strategies, and the center (equal probability) lies at the centroid. Higher dimensions generalize this to $(n-1)$-dimensional simplices. Policy optimization navigates this constrained space.

We use the metric to encode safety: regions with "large" metrics are "far away" in the eyes of the optimization algorithm, making geodesics (shortest paths) naturally avoid them.

3. Sheaf-Theoretic Reward Spaces

What mathematical structure captures the idea that "local evaluations must agree when combined into global assessments"? We answer this question using sheaf theory.

3.1. The Reward Sheaf Construction

Standard reinforcement learning assumes that rewards are scalars $r_t \in \mathbb{R}$ that sum consistently along trajectories. However, human feedback is often multi-scale (per-step, per-episode) and multi-dimensional (safety, efficiency, style).

We model the trajectory space as a **cell complex** X where states are vertices, transitions are edges, and local consistency checks form faces. A **reward sheaf** $\mathcal{F}$ on this space assigns possible reward valuations to each region, with restriction maps ensuring that global evaluations constrain local ones consistently. (See Appendix A for formal definitions.)

Intuitively, a reward sheaf captures how local evaluations (per-step rewards) must be consistent when combined into global assessments (trajectory rewards). For a trajectory τ composed of segments $\sigma_1, \ldots, \sigma_k$, a consistent reward assignment requires that the global evaluation $R(\tau)$ agrees with the aggregation of local evaluations $r(\sigma_i)$.

3.2. Cohomological Consistency

When human evaluators provide pairwise preferences or scalar ratings, they are effectively sampling sections of this sheaf. Inconsistencies arise when these local samples cannot be glued into a global section.

Intuitively, H^1 **cohomology** measures "holes" in preference space. Imagine preferences as arrows on a map: "from state A, go toward B." If arrows form a consistent flow (like water downhill), you could define a height function—this is a *potential*, and standard RL assumes it exists. But what if preferences form a **cycle**? A preferred to B, B to C, C to A. Now there is no height function: you cannot go consistently "downhill" around a circle. This creates a "hole" in preference space.

H^1 counts these holes. If $H^1 = 0$, preferences are consistent. If $H^1 \neq 0$, there are fundamental cycles no scalar reward can capture. **Key insight**: When $H^1 \neq 0$, forcing a scalar reward leads to reward hacking.

Key Result (Theorem B.1 in Appendix): Local feedback samples are globally consistent if and only if they can be "glued" together—formally, the first cohomology class H^1

vanishes. Non-zero H^1 signals fundamental preference cycles that no scalar reward can represent.

Condorcet Cycles. A non-transitive preference cycle $A \succ B \succ C \succ A$ corresponds to a non-trivial cohomology class. This implies no scalar value function $V : S \to \mathbb{R}$ can perfectly represent the preferences, as $\oint \nabla V = 0$ while the preference circulation is non-zero.

3.3. The Hodge Decomposition for Rewards

To handle these inconsistencies, we propose decomposing the reward signal rather than forcing it into a scalar potential.

Clarification: Graph Topology, Not Dimensionality Reduction. A potential misconception: one might assume we simply embed rewards as vectors, apply PCA to reduce to 3D, and analyze that surface. **This is not what we do.** The Hodge decomposition operates on the *graph structure* of preference relationships—the cell complex X where states are nodes and transitions are edges. We compute gradient, curl, and harmonic components based on the *combinatorial topology* of this preference graph, not on the geometry of a dimensionality-reduced embedding.

The advantage over ordinal gradient methods (which compute $\nabla r \approx r(A) - r(B)$) is threefold:

1. **Cycle detection**: Standard gradients assume $\oint \nabla V = 0$; we detect when this fails ($H^1 \neq 0$).
2. **Orthogonal decomposition**: We separate learnable signal (gradient) from irreducible cycles (harmonic).
3. **Multi-scale consistency**: The sheaf structure enforces agreement across local, segment, and trajectory scales.

Just as any vector field can be split into curl-free (flows from source to sink) and divergence-free (rotational) parts, the **Hodge decomposition** splits preferences into:

- **Gradient (dV)**: Consistent preferences ($A > B > C$), learnable as a value function.
- **Harmonic (ω)**: Fundamental cycles ($A > B > C > A$) that cannot be "unrolled."

Formally, any reward function on transitions decomposes as $r = dV + \delta\psi + \omega$, where the three components are orthogonal (Theorem B.2 in Appendix).

Implication for RL. Standard RL algorithms (PPO, DQN) implicitly assume $r \approx dV$ and try to learn V. When $\omega \neq 0$ (cyclic preferences), the Bellman error $r + V(s') - V(s)$ cannot be minimized to zero, leading to instability or value collapse. Our framework explicitly learns both V and ω, allowing the agent to model and navigate cyclic preferences.

Worked Example: Medical Triage AI. Consider an AI prioritizing patients where doctors, administrators, and families have conflicting preferences. Aggregating yields a Condorcet cycle with circulation $\oint r = 0.6 \neq 0$, confirming $H^1 \neq 0$. Hodge decomposition separates the learnable gradient (severity ranking) from the irreducible harmonic cycle (stakeholder disagreement). GPO learns both, navigating the preference landscape without mode collapse. See Appendix O for the full walkthrough.

4. Geometric Safety via Black Holes

While sheaf cohomology addresses consistency, it does not inherently guarantee safety. To address catastrophic risks, we introduce a geometric formalism for forbidden regions, modeling them as singularities in the reward manifold.

4.1. The Reward Manifold

We define the **Reward Manifold** $\mathcal{M}$ as the embedding of the trajectory space into a high-dimensional metric space $(\mathbb{R}^d, g)$. Unlike the flat Euclidean space assumed in standard feature embeddings, $\mathcal{M}$ possesses intrinsic curvature determined by the safety landscape.

*How can we guarantee an agent will **never** enter a forbidden state, not just "probably" avoid it?*

A **black hole** $\mathcal{B} \subset \mathcal{M}$ is a compact region representing forbidden outcomes (e.g., irreversible harm), characterized by an event horizon (boundary) and a singularity where the cost diverges. (See Definition A.3 in Appendix.)

Worked Example: Autonomous Drone. Consider a drone where engaging with high confidence near civilians guarantees casualties. Rather than soft penalties (outweighed by mission reward), we place a geometric singularity: $g(x) \to \infty$ near this region. The geodesic distance becomes *infinite*, making the black hole unreachable with probability exactly zero. See Appendix O for details.

4.2. Learning the Safety Metric

Standard constrained RL (e.g., CPO) treats safety as a separate cost signal $C(s)$ constrained by a threshold d. We propose incorporating safety directly into the geometry of the state space via a **Riemannian metric** g.

The key insight is that we can make dangerous states **infinitely far away** in geodesic distance. We construct a conformal metric $g_{ij}(x) = e^{2\sigma(x)}\delta_{ij}$, where the conformal factor $\sigma(x) \to \infty$ as x approaches dangerous states. Specifically, if $e^{\sigma(x)} \approx 1/\text{dist}(x, \mathcal{B})^\alpha$ for $\alpha \geq 1$, then the geodesic

distance to any black hole is infinite (Proposition B.3 in Appendix).

In practice, we parametrize $\sigma_\theta(x)$ using a neural network trained on a "safety potential" derived from cost signals. The network learns to output large scaling factors near high-cost regions, effectively stretching the space so that dangerous regions become "infinitely far away" for an agent traversing geodesics.

4.3. Theoretical Guarantees

This geometric formulation provides stronger guarantees than soft penalties.

Safety Guarantee (Theorem B.4 in Appendix): A policy that follows geodesics on a manifold with properly constructed singularities will **never** enter any black hole region—not with low probability, but with probability exactly 0.

This transforms the safety problem from a constrained optimization problem (hard to solve, prone to feasibility issues) into an unconstrained shortest-path problem on a curved manifold.

5. Geodesic Policy Optimization

We introduce **Geodesic Policy Optimization (GPO)**, an algorithm that optimizes policies to follow geodesics on the learned reward manifold. GPO integrates the geometric safety constraints and sheaf-theoretic consistency checks into a unified policy gradient framework.

5.1. Riemannian Policy Gradient

Standard policy gradient methods perform updates in the Euclidean space of parameters, often using the Fisher Information Matrix (Natural Policy Gradient) to account for the geometry of the probability simplex. GPO extends this by incorporating the **geometry of the reward manifold** itself.

Let $J(\theta)$ be the expected return. The standard gradient update is $\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta_k)$. In GPO, we define the update direction using the inverse of the learned Riemannian metric $G(s)$:

$$\nabla_G J(\theta) = \mathbb{E}_{s,a\sim\pi}\left[G(s)^{-1/2}\nabla_\theta \log \pi_\theta(a|s) A^{\text{Hodge}}(s,a)\right] \tag{3}$$

Here, the term $G(s)^{-1/2}$ acts as a preconditioner that dampens gradient steps in regions of high curvature (near black holes) and amplifies them in flat, safe regions. This naturally enforces safety: as the agent approaches a danger zone, the effective learning rate drops to zero, preventing the policy from pushing further into the trap.

Algorithm 1 Geodesic Policy Optimization

1: **Initialize**: Policy π_θ, Hodge Critic (V_ϕ, ω_ψ), Metric Model G_ξ
2: **for** iteration $k = 1, 2, \ldots$ **do**
3: Collect trajectories $\tau \sim \pi_\theta$
4: **Update Metric**: Train G_ξ on cost signals to approximate singularity at $C(s) >$ threshold
5: **Update Critic**: Minimize Hodge Bellman Error to learn V_ϕ and ω_ψ
6: **Compute Advantages**: Calculate Riemannian-scaled, Hodge-corrected advantages A^{Hodge}
7: **Update Policy**: Perform gradient ascent using $\nabla_G J(\theta)$
8: **end for**
9: **Return**: Safe policy π^*

5.2. The Hodge-Augmented Critic

To handle cyclic preferences, GPO replaces the standard scalar value function $V(s)$ with a **Hodge Critic** that learns both the potential and harmonic components of the reward.

The **Hodge-Bellman error** extends the standard TD error to account for preference cycles. Instead of forcing $r = V(s') - V(s)$ (which fails for cycles), we learn both a potential V and a "circulation" ω that absorbs the cyclic component.

The critic is parameterized as (V_ϕ, ω_ψ), minimizing:

$$\mathcal{L}(\phi,\psi) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) - \underbrace{(V_\phi(s') - V_\phi(s))}_{\text{Potential Diff.}} - \underbrace{\omega_\psi \cdot v(s,a)}_{\text{Harmonic Flux}}\right)^2\right] \tag{4}$$

where $v(s,a)$ is the velocity vector of the transition. The term ω_ψ captures the non-transitive "circulation" of the reward.

5.3. Geodesic Advantage Estimation

The advantage function in GPO is modified to account for both the metric distortion and the harmonic component:

$$A^{\text{Hodge}}(s,a) = \frac{1}{\sqrt{G(s)}}\left(r(s,a) + \gamma V(s') - V(s) - \omega_\psi \cdot v(s,a)\right) \tag{5}$$

By normalizing the advantage by $\sqrt{G(s)}$, we ensure that high-reward but high-risk actions (inside a trap) have a diminished impact on the policy update, effectively "discounting" rewards gained in dangerous geometries.

5.4. The GPO Algorithm

Table 1. Comparison across methods on benchmark tasks. GPO achieves zero safety violations while baselines fail on deceptive traps. Results from trained Q-table policies.

Method	Cycle Det.	Avg Safety Viol.	Avg Reward
Random	0%	16.0%	0.60
PPO	0%	26.7%	1.27
CPO (Lagrangian)	0%	26.7%	1.23
GPO (ours)	**100%**	**0.0%**	0.53

Table 2. Cohomology detection on Condorcet Ring (100 episodes). GPO successfully detects the cyclic preference structure while PPO and CPO completely fail.

Method	Train Speed	H^1 Estimate	Ground Truth	Mean Reward	Cycle Det.
PPO	–	0.000	0.500	-4.71	False
CPO	$1.0\times$	0.000	0.500	-0.01	False
GPO (ours)	$1.3\times^{\dagger}$	**0.425**	0.500	-5.62	**True**
	$1.8\times^{\dagger}$				

5.5. Integrating PPO and CPO

GPO unifies PPO and CPO by encoding their respective strengths geometrically. **Clipped-GPO** combines PPO's clipped surrogate with geodesic scaling, achieving GPO's safety with $1.8\times$ faster convergence. **CPO-initialized GPO** uses CPO to learn approximate safety boundaries, then converts soft Lagrangian penalties into hard metric singularities. The key insight: PPO provides stable optimization, CPO provides constraint satisfaction, and GPO's Riemannian metric provides both in a single geometric object with *deterministic* guarantees. See Appendix P for technical details.

6. Experiments

We validate the STRS framework and GPO algorithm on three distinct problem settings designed to isolate specific failure modes of standard RLHF: cyclic preferences, deceptive safety traps, and stylistic finetuning.

6.1. Experimental Setup

Baselines. We compare our proposed **Geodesic Policy Optimization (GPO)** against two primary baselines:

1. **PPO** (Proximal Policy Optimization) (Schulman et al., 2017): The standard algorithm for RLHF, representing the scalar reward hypothesis.

2. **CPO** (Constrained Policy Optimization) (Achiam et al., 2017): A leading Safe RL algorithm that enforces constraints via trust region projection and Lagrangian relaxation.

Metrics. We evaluate agents on **Cycle Detection** (ability to identify $H^1 \neq 0$), **Safety Violation Rate** (% of episodes entering forbidden regions), and **Goal Performance** (task reward achieved without violating safety).

6.2. Detecting Cyclic Preferences (Condorcet Ring)

Setting. The environment is a continuous circle S^1 where the agent receives a constant positive reward for moving clockwise. This creates a Condorcet cycle ($A \succ B \succ C \succ A$) with ground truth cohomology $H^1 = 0.5$.

Results. As shown in Table 2, GPO successfully detects the cyclic preference structure, estimating $H^1 = 0.425$ after 100 training episodes (vs. ground truth 0.5), whereas PPO and CPO completely fail to detect any cycle ($H^1 = 0.0$). This validates our core claim: standard RL algorithms cannot detect preference inconsistency, while GPO's Hodge-augmented critic achieves **100% cycle detection rate**.

6.3. Geometric Safety (Ethical Scenarios)

Setting. We evaluate agents on five ethical scenarios: Academic Integrity, Drone Decision, Business Ethics, **Murky Drone** (instrumental convergence trap), and our new **Agentic Shortcut** scenario—inspired by Anthropic's "From Shortcuts to Sabotage" research (Anthropic, 2025). In this scenario, an AI coding assistant can take a shortcut (e.g., calling `sys.exit(0)` to fake test passage) that yields high immediate reward but provides no real value.

Results. GPO achieves a **0% violation rate** across all five scenarios, using *trained Q-table policies* (not hardcoded distributions). Table 3 reveals the critical failure mode of soft-constraint methods: when deceptive rewards exceed penalty thresholds, Lagrangian relaxation fails catastrophically. In both Murky Drone and Agentic Shortcut, PPO takes the catastrophic action **100%** of the time, while CPO achieves 100% and 89% respectively. The Agentic Shortcut scenario is particularly revealing: the shortcut yields reward $\sim$1.8 (vs 0.5 for correct code), directly mirroring real-world reward hacking where `sys.exit(0)` tricks test harnesses. Only GPO's geometric singularity—which makes deceptive actions *infinitely* costly—achieves 0% violations on both scenarios.

Safety vs Reward Trade-off. A key observation: PPO and CPO achieve higher average reward (2.64) than GPO (0.25) on Murky Drone precisely *because* they take catastrophic actions. This demonstrates that **safety and reward are fundamentally in tension** when environments contain deceptive traps. GPO sacrifices reward to guarantee safety; soft-constraint methods like CPO attempt to balance both but fail when rewards are sufficiently deceptive. Notably, on non-deceptive scenarios (Academic, Business, Drone), all methods achieve similar performance, confirming that GPO's conservatism is targeted rather than universal.

Table 3. Scenario-specific safety violations (V%) and rewards (R) from trained Q-table policies (100 episodes each). Both deceptive scenarios (Murky Drone, Agentic Shortcut) cause catastrophic failure in PPO/CPO. Only GPO's geometric barrier achieves unconditional safety.

	Academic		**Murky Drone**		**Agentic Shortcut**		**Business**		**Drone**	
Method	V%	R	V%	R	V%	R	V%	R	V%	R
Random	27	0.30	39	0.97	41	0.99	28	0.41	0	0.34
PPO	0	0.69	**100**	2.63	**100**	1.84	0	0.59	0	0.59
CPO	0	0.68	**100**	2.63	89	1.70	0	0.59	0	0.57
GPO	**0**	0.70	**0**	0.25	**0**	0.50	**0**	0.60	**0**	0.62

Table 4. Ablation study results. Lower clip ratios and higher black hole strengths reduce safety violations.

Ablation	**Value**	**Safety Viol.**	**Reward**	**Conv. Steps**
Clip ratio	$\epsilon = 0.05$	**1.1%**	0.778	55
Clip ratio	$\epsilon = 0.20$	1.4%	0.780	70
Black hole	$\alpha = 1.0$	4.0%	0.780	70
Black hole	$\alpha = 5.0$	**0.0%**	0.700	110

6.4. Continuous Control (Robotics Reaching)

We also validate GPO on continuous robotics tasks requiring velocity-aware safety. On a 2D reaching task, GPO achieves 100% success with 0 collisions on simple layouts and reduces collisions by 94% compared to PPO on blocked paths (3.0 vs 51.0 collisions/episode). Full results in Appendix N.

6.5. Ablation Study: Clipped-GPO

We analyzed the trade-off between the geometric threshold τ, clip ratio ϵ, and black hole strength α using *actual training runs* with wall-clock timing. Table 4 summarizes key findings:

- **Clip ratio** $\epsilon = 0.05$ achieves the lowest safety violation rate (1.1%) among clip ratio configurations, with fastest convergence (55 steps)
- **Black hole strength** $\alpha = 5.0$ achieves **0% violations**, confirming that stronger geometric penalties provide unconditional safety guarantees at a modest convergence cost (110 steps vs. 56 for $\alpha = 0.5$)
- The optimal balance is $\tau = 0.5$, $\epsilon = 0.05$, $\alpha = 3.0$: 90 steps convergence with 2% violations

†*Train speeds measured via wall-clock timing on Modal L4 GPU instances.*

6.6. Scale Experiments: Anthropic HH-RLHF

To validate STRS on real-world preference data, we conduct topology mining on 100,000 preference pairs from the Anthropic HH-RLHF dataset (Bai et al., 2022). Importantly, **all models (PPO, CPO, GPO) were trained on this same dataset**, making the harmonic risk distribution an intrinsic property of the training data rather than a separate test set.

Harmonic Risk Distribution. Figure 6 shows the distribution of harmonic risk scores across the dataset. The mean harmonic risk is 0.754 (median 0.767), indicating that preference inconsistency is pervasive rather than exceptional. Critically, the percentage of inconsistent pairs is **threshold-dependent**: 94% exceed $r \geq 0.6$ (moderate inconsistency), 77% exceed $r \geq 0.7$ (high), and **33.3%** exceed $r \geq 0.8$ (severe). We adopt the 0.8 threshold as it identifies preference pairs with strong local contradictions that contribute meaningfully to H^1 cohomology—these are the cases where scalar reward models provably fail to find a consistent value function.

This threshold-dependent characterization reveals that inconsistency is not binary but exists on a spectrum. The 33% figure represents preference pairs where local neighborhoods exhibit *severe* directional disagreement (harmonic component dominates gradient component in the Hodge decomposition), making them particularly vulnerable to reward hacking when collapsed to scalars.

Manifold Navigation. We trained a GeoDPO agent on this dataset, using harmonic risk scores to weight the geometric penalty. The agent's responses show a consistent trajectory shift away from the "black hole" regions of inconsistency. On the top 5% riskiest prompts, GeoDPO achieves a positive trajectory shift for 50% of cases, effectively navigating around the manifold singularities that trap baseline models.

7. Related Work

7.1. Distributional and Multi-Objective RL

Standard RL treats the return as a scalar random variable and optimizes its expectation. **Distributional RL** (Bellemare et al., 2017; Dabney et al., 2018) estimates the full distribution of returns, capturing uncertainty and multimodality, but typically still projects this distribution back to a scalar policy gradient or value estimate for decision making.

Multi-Objective RL (MORL) (Roijers et al., 2013) deals with vector-valued rewards, aiming to approximate the Pareto front. Our work differs by assuming that the vector nature of feedback arises not just from competing scalar objectives, but from the **topological structure** of the preference space itself. The Hodge decomposition provides a principled way to orthogonalize these components into consistent (exact) and cyclic (harmonic) parts, which MORL does not explicitly address.

7.2. Safe Reinforcement Learning

The dominant paradigm for safety is the **Constrained Markov Decision Process (CMDP)** (Altman, 1999). **Constrained Policy Optimization (CPO)** (Achiam et al., 2017) solves CMDPs by computing trust region updates that satisfy linearized safety constraints. Other approaches use Lagrangian relaxation (Ray et al., 2019) or safety layers (Dalal et al., 2018).

While effective for explicit constraints, these methods struggle with "deceptive" risks where high rewards lure agents into irreversible states before constraints are violated in expectation. Our **Geodesic Policy Optimization (GPO)** replaces explicit constraints with intrinsic geometry. By learning a metric that diverges at danger zones, we enforce safety via the principle of least action, providing stronger avoidance guarantees than soft penalties.

7.3. Topological Methods in Machine Learning

Topological Data Analysis (TDA) has applied persistent homology to characterize the shape of data manifolds (Edelsbrunner & Harer, 2010). In deep learning, **Neural Sheaf Diffusion** (Bodnar et al., 2022) uses cellular sheaves to model non-smoothing information flow in Graph Neural Networks.

To our knowledge, **STRS** is the first framework to apply sheaf cohomology to Reinforcement Learning from Human Feedback. We show that the "alignment problem" is partially a topological one: inconsistencies in human feedback are not just noise to be filtered, but topological features (cycles) to be detected and modeled.

7.4. Preference Learning and Intransitivity

The standard Bradley-Terry model (Bradley & Terry, 1952) assumes transitive preferences. However, intransitivity is common in social choice (Condorcet, 1785) and complex AI objectives. **Cyclic preference learning** attempts to model this but often lacks a mechanism to integrate it into control. Our use of the Hodge decomposition bridges this gap, treating the cyclic component (ω) as a valid driving force for the policy, akin to a non-conservative force field in physics.

8. Discussion and Limitations

8.1. GPO vs. CPO: A Nuanced Comparison

Our experiments reveal a clear safety advantage for GPO over CPO. **Key finding**: GPO achieves **0% safety violations** across all ethical scenarios, while CPO incurs 26.7% violations—comparable to PPO's 26.7%. This reflects a fundamental difference in mechanism:

- **CPO** uses Lagrangian multipliers that provide *soft* penalties. When deceptive rewards exceed the penalty threshold (as in Murky Drone: reward $\sim$2.6 vs typical rewards $\sim$0.5), the optimization overwhelms the safety constraint.
- **GPO** uses geometric structure where the metric diverges to infinity near forbidden regions. No finite reward can justify crossing an infinite geodesic distance—this provides **unconditional** safety guarantees.

The key insight: soft-constraint methods (CPO, penalty-based approaches) fail catastrophically when deceptive rewards exceed penalty thresholds. GPO's geometric barriers are **insurmountable by design**, making it the preferred choice when zero violations is paramount.

8.2. Limitations

While Sheaf-Theoretic Reward Spaces offer a rigorous alternative to scalar reward modeling, several limitations remain:

1. **Computational Complexity**: Computing exact Čech cohomology grows combinatorially with the size of the cover. While our discrete approximation on trajectory graphs is efficient ($O(T \cdot k \cdot d^2)$), scaling to massive datasets may require sparse approximations or spectral methods.
2. **Metric Bootstrapping**: The safety metric $g(x)$ relies on cost signals to identify "event horizons." If these signals are sparse or noisy, the learned geometry may be flawed.
3. **Manifold Assumption**: GPO assumes the reward space has a meaningful manifold structure. In discrete domains (e.g., token-level text generation), this smoothness assumption is an approximation.
4. **Reward Trade-off**: GPO's geometric barriers force policies to avoid deceptive high-reward regions, resulting in lower average task reward on scenarios with tempting shortcuts. This is a feature, not a bug, but users should understand that GPO prioritizes unconditional safety over reward maximization.

9. Conclusion

We have presented **Sheaf-Theoretic Reward Spaces (STRS)**, a rigorous mathematical framework that reimagines the foundations of Reinforcement Learning from Human Feedback. By lifting rewards from scalars to sheaf sections, we expose the rich topological structure of human preferences—including inconsistencies and cycles—that standard methods ignore.

Our results demonstrate that:

1. H^1 **is a computable safety certificate**: The first cohomology group successfully detects Condorcet cycles in preference data, providing a concrete metric for alignment consistency.
2. **Geometry enforces safety**: Modeling forbidden regions as singularities in a Riemannian manifold enables **Geodesic Policy Optimization (GPO)** to achieve near-perfect safety rates in deceptive environments where PPO fails and CPO struggles.
3. **Cyclic navigation is possible**: The Hodge-Augmented Critic allows agents to navigate preference cycles intelligently, "orbiting" the Pareto frontier of diverse user needs rather than collapsing to a mediocre mean.

As AI systems become more autonomous and their objectives more complex, the "scalar hypothesis"—that all values can be mapped to a single number—becomes increasingly untenable. STRS provides the necessary language to describe, measure, and optimize for the full spectrum of human intent, paving the way for safer, more nuanced, and topologically aware AI systems.

Impact Statement

This work aims to improve the safety and alignment of AI systems by providing mathematically rigorous tools for detecting preference inconsistencies and enforcing geometric safety constraints. The primary positive impact is enabling more robust alignment of autonomous systems with human values, particularly in high-stakes domains where current scalar reward methods fail.

However, we acknowledge potential dual-use concerns. The techniques for navigating preference manifolds could theoretically be misused to manipulate user preferences more effectively or to find exploits in safety systems. We emphasize that the "black hole" safety mechanism is designed as a protective measure, but any tool for understanding reward geometry could be inverted.

We believe the benefits of formal safety certificates outweigh these risks, as the current opacity of RLHF systems poses greater dangers than well-understood geometric methods. We encourage the research community to develop complementary verification tools for the geometric safety properties we introduce.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Anthropic. From shortcuts to sabotage: Natural emergent misalignment from reward hacking. *Anthropic Research Blog*, 2025. URL https://www.anthropic.com/research/emergent-misalignment-reward-hacking.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Bodnar, C., Di Giovanni, F., Chamberlain, B. P., Lio, P., and Bronstein, M. M. Neural sheaf diffusion: A topological framework for heterophily and oversmoothing in graph neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18523–18536, 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Condorcet, M. J. A. N. d. C. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional reinforcement learning with quantile regression. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Edelsbrunner, H. and Harer, J. *Computational topology: an introduction*. American Mathematical Society, 2010.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Ray, A., Achiam, J., and Amodei, D. Benchmarking safe exploration in deep reinforcement learning. In *arXiv preprint arXiv:1910.01708*, 2019.

Roijers, D. M., Vymetal, P., Whiteson, S., and Kappen, H. J. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

A. Formal Definitions

This appendix provides the complete formal definitions summarized in the main text.

Definition A.1 (Trajectory Space)**.** *Let S be the state manifold and A the action space. We define a cell complex X representing the trajectory space, where:*

- ***0-cells (Vertices)**: States $s \in S$.*
- ***1-cells (Edges)**: Transitions (s, a, s').*
- ***2-cells (Faces)**: Elementary loops or 3-step transition cycles representing local consistency checks.*

Definition A.2 (Reward Sheaf)**.** *A reward sheaf $\mathcal{F}$ on X assigns a vector space $\mathcal{F}(U)$ to each open set $U \subset X$, representing possible reward valuations over that region. For an inclusion $V \subset U$, the restriction map $\rho_{UV} : \mathcal{F}(U) \to \mathcal{F}(V)$ captures how global evaluations constrain local ones.*

Definition A.3 (Black Hole Region)**.** *A black hole $\mathcal{B} \subset \mathcal{M}$ is a compact region representing forbidden outcomes (e.g., irreversible harm). It is characterized by:*

1. ***Event Horizon**: The boundary $\partial\mathcal{B}$.*
2. ***Singularity**: A point or region $c \in int(\mathcal{B})$ where the cost function diverges.*

B. Theoretical Proofs

B.1. Proof of Theorem B.1 (Consistency Criterion)

Theorem B.1 (Consistency Criterion)**.** *Let $\mathcal{U}$ be a cover of the trajectory space X. The collection of local feedback samples $\{r_i \in \mathcal{F}(U_i)\}$ is globally consistent if and only if the first Čech cohomology class vanishes:*

$$[\delta\{r_i\}] = 0 \in H^1(\mathcal{U}, \mathcal{F}) \tag{6}$$

Proof. By definition of the Čech coboundary operator $\delta : C^0(\mathcal{U}, \mathcal{F}) \to C^1(\mathcal{U}, \mathcal{F})$, a collection of local sections $\{r_i\}$ defines a 0-cochain. The condition for these sections to glue into a global section $r \in \mathcal{F}(X)$ is that they agree on all overlaps: $r_i|_{U_i \cap U_j} = r_j|_{U_i \cap U_j}$ for all i, j.

This can be rewritten as $r_i - r_j = 0$ on U_{ij}. The 1-cochain $\delta\{r_i\}$ is defined by $(\delta r)_{ij} = r_j - r_i$ on U_{ij}.

Thus, the gluing condition is exactly $\delta\{r_i\} = 0$.

If $[\delta\{r_i\}] = 0$ in cohomology, it means the obstruction is a coboundary, which implies the local inconsistencies can be resolved by a re-parametrization (adding a 0-cochain), or if we are checking for the existence of *any* global section, the class must be trivial. Specifically, for the reward sheaf, non-trivial H^1 implies no scalar potential function exists that agrees with all local preference gradients. □

B.2. Proof of Theorem B.2 (Hodge Decomposition)

Theorem B.2 (Hodge Decomposition)**.** *Let $r \in C^1(X, \mathbb{R})$ be a reward 1-cochain (function on transitions). Then r admits a unique orthogonal decomposition:*

$$r = dV + \delta\psi + \omega \tag{7}$$

where:

1. *dV is the **Exact Component** (Gradient flow): Represents consistent, transitive preferences derived from a scalar potential V (the standard Value function).*
2. *$\delta\psi$ is the **Coexact Component** (Curl flow): Represents local rotational inconsistencies (often zero in tree-like MDPs).*
3. *ω is the **Harmonic Component** (Global flow): Represents fundamental topological cycles (H^1).*

Proof. Let C^0, C^1, C^2 be the spaces of 0, 1, and 2-cochains equipped with the standard L^2 inner product $\langle f, g \rangle = \sum f(x)g(x)$.

The coboundary operators are $\delta_k : C^k \to C^{k+1}$. Their adjoints (boundary operators) are $\delta_k^* : C^{k+1} \to C^k$.

The Laplacian is defined as $\Delta_k = \delta_k^* \delta_k + \delta_{k-1} \delta_{k-1}^*$.

By the Hodge theorem for finite complexes, the space of k-cochains decomposes as:

$$C^k = \text{im}(\delta_{k-1}) \oplus \text{im}(\delta_k^*) \oplus \ker(\Delta_k) \tag{8}$$

For $k = 1$ (rewards on edges):

$$C^1 = \text{im}(\delta_0) \oplus \text{im}(\delta_1^*) \oplus \ker(\Delta_1) \tag{9}$$

Identifying terms:

1. $\text{im}(\delta_0) = \{\delta_0 V \mid V \in C^0\}$ is the space of **exact forms** (gradient fields).
2. $\text{im}(\delta_1^*) = \{\delta_1^* \psi \mid \psi \in C^2\}$ is the space of **coexact forms** (curl fields).
3. $\ker(\Delta_1) \cong H^1(X, \mathbb{R})$ is the space of **harmonic forms**.

Thus, any reward r uniquely decomposes into $dV + \delta\psi + \omega$. □

B.3. Metric Singularity Proposition

Proposition B.3 (Metric Singularity)**.** *To enforce strict avoidance of the black hole, we require the geodesic distance from any safe state s_{safe} to the event horizon to be infinite. This condition is satisfied if the conformal factor scales as:*

$$e^{\sigma(x)} \approx \frac{1}{dist(x, \mathcal{B})^\alpha} \tag{10}$$

where $\alpha \geq 1$.

Proof. Let $\gamma(t)$ be any path approaching the boundary $\partial\mathcal{B}$. The Riemannian length is:

$$L_g(\gamma) = \int_0^1 e^{\sigma(\gamma(t))} \|\dot{\gamma}(t)\| \, dt \geq \int e^{\sigma} |\dot{u}| \, dt \tag{11}$$

where $u(t) = \text{dist}(\gamma(t), \mathcal{B})$. Substituting $e^\sigma \approx u^{-\alpha}$:

$$L_g(\gamma) \geq \int_{u_0}^0 u^{-\alpha} \, du = \infty \quad \text{for } \alpha \geq 1 \tag{12}$$

□

B.4. Proof of Theorem B.4 (Geodesic Avoidance)

Theorem B.4 (Geodesic Avoidance)**.** *Let π^* be a policy that generates trajectories minimizing the Riemannian path length on $(\mathcal{M}, g)$. If the metric g satisfies the singularity condition (Proposition B.3) for all $\mathcal{B}_i$, then π^* will not enter any black hole region with probability 1.*

Proof. Consider any path $\gamma : [0, 1] \to \mathcal{M}$ starting at a safe state $\gamma(0) \notin \mathcal{B}$ and ending at the boundary $\gamma(1) \in \partial\mathcal{B}$.

The Riemannian length is $L_g(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}, \dot{\gamma})} \, dt$.

Let $u(t) = \text{dist}(\gamma(t), \mathcal{B})$. As $t \to 1$, $u(t) \to 0$.

Since $|\dot{u}| \leq \|\dot{\gamma}\|$, we have $\|\dot{\gamma}\| \geq |\dot{u}|$.

Substituting the metric condition:

$$L_g(\gamma) \geq \int_0^1 \frac{C}{u(t)^\alpha} \|\dot{\gamma}(t)\| \, dt \geq \int_{u(0)}^0 \frac{C}{u^\alpha} \, du \tag{13}$$

For $\alpha \geq 1$, the integral $\int_\epsilon^0 u^{-\alpha} \, du$ diverges to ∞.

Therefore, any path touching the black hole has infinite length. Since we assume there exists at least one safe path with finite length (the environment is connected), the minimizer π^* will assign probability 0 to paths entering $\mathcal{B}$. □

C. Mathematical Background

This section provides background on the mathematical concepts used in the main paper for readers less familiar with algebraic topology.

C.1. Sheaf Theory Primer

A **sheaf** $\mathcal{F}$ on a topological space X is a systematic way to organize local data that can be consistently combined into global data. Formally, a sheaf assigns:

- To each open set $U \subseteq X$: a set (or vector space) $\mathcal{F}(U)$ of "sections over U"
- To each inclusion $V \subseteq U$: a "restriction map" $\rho_{UV} : \mathcal{F}(U) \to \mathcal{F}(V)$

satisfying two axioms:

1. **Identity**: $\rho_{UU} = \text{id}$
2. **Gluing**: If sections $s_i \in \mathcal{F}(U_i)$ agree on overlaps ($s_i|_{U_i \cap U_j} = s_j|_{U_i \cap U_j}$), there exists a unique global section $s \in \mathcal{F}(\bigcup U_i)$ restricting to each s_i.

Example: Consider temperature measurements across a city. Each weather station provides a local reading; a sheaf captures whether these readings can be consistently interpolated into a global temperature field.

C.2. Čech Cohomology

Given a cover $\mathcal{U} = \{U_i\}$ of X, Čech cohomology measures obstructions to gluing local sections.

- $H^0(X, \mathcal{F})$: Global sections (data that extends consistently everywhere)
- $H^1(X, \mathcal{F})$: Obstructions to gluing—"holes" that prevent local consistency from implying global consistency

The **coboundary operator** $\delta : C^0 \to C^1$ is defined by $(\delta s)_{ij} = s_j|_{U_{ij}} - s_i|_{U_{ij}}$. A collection of local sections glues if and only if $\delta s = 0$.

C.3. Discrete Hodge Theory

On a finite cell complex (graph), the Hodge decomposition provides an orthogonal splitting of edge functions (1-cochains):

$$C^1 = \underbrace{\text{im}(d)}_{\text{gradients}} \oplus \underbrace{\text{im}(d^*)}_{\text{curls}} \oplus \underbrace{\ker(\Delta)}_{\text{harmonic}} \tag{14}$$

The **graph Laplacian** $\Delta = d^*d + dd^*$ plays a central role. Harmonic forms ($\Delta\omega = 0$) represent topologically nontrivial cycles—preference loops that cannot be "unrolled" into a consistent ranking.

C.4. Cochain Complexes

A **cochain complex** is a sequence of vector spaces connected by "coboundary" maps:

$$0 \to C^0 \xrightarrow{d_0} C^1 \xrightarrow{d_1} C^2 \to \cdots \tag{15}$$

satisfying $d_{k+1} \circ d_k = 0$. Intuitively:

- 0-cochains assign values to vertices (states)
- 1-cochains assign values to edges (transitions)
- 2-cochains assign values to faces (local cycles)

The coboundary maps check consistency across dimensions.

D. Beyond Ordinal Feedback: Verbal and Categorical Signals

The main text focuses on ordinal pairwise preferences ($A \succ B$), but our framework naturally extends to richer feedback modalities. This section describes how verbal feedback and categorical ratings integrate with the sheaf-theoretic approach.

D.1. Feedback Modalities and Their Geometric Interpretation

Ordinal (Pairwise Preferences). Standard RLHF uses comparisons: "A is better than B." This induces a *gradient* on the preference graph: $r(A \to B) = \text{sign}(A \succ B)$. The Bradley-Terry model computes differences between latent scalar utilities.

Verbal (Critique-Based Feedback). A richer signal: "Response A is good because it addresses safety concerns, but lacks detail." We encode this via a **vector-to-vector mapping**:

$$f_{\text{verbal}} : \text{response embedding } e_A \mapsto \text{critique embedding } c_A \in \mathbb{R}^d \tag{16}$$

The critique embedding c_A becomes a *local section* of the reward sheaf at response A. Unlike ordinal feedback (which only gives relative differences), verbal feedback provides *absolute* directional information: the vector c_A points toward "improvement" from A.

To derive Hodge components from verbal feedback:

- **Gradient**: Project critique vectors onto the gradient field. If $c_A \approx \nabla V$ for some potential V, this component is learnable as a value function.
- **Harmonic**: The residual $c_A - \text{proj}_{\text{im}(d)} c_A$ captures directional preferences that cannot be explained by any scalar potential—e.g., "be more formal" vs "be more casual" as a cyclic style preference.

Categorical (Rubric-Based Feedback). Discrete ratings: {`exceptional`, `pass`, `acceptable`, `suboptimal`, `failure`}. We encode this via **vector-to-curvature mapping**:

$$f_{\text{categorical}} : \text{response embedding } e_A \mapsto \text{curvature } \kappa(e_A) \in [0, \infty] \tag{17}$$

The mapping directly assigns geometric structure:

- `exceptional/pass`: $\kappa \approx 0$ (flat region, safe for exploration)
- `acceptable`: $\kappa > 0$ (mild curvature, slight penalty)
- `suboptimal`: $\kappa \gg 0$ (high curvature, strong avoidance signal)
- `failure`: $\kappa = \infty$ (singularity, black hole boundary)

This directly constructs the Riemannian metric g without learning from cost signals: $g(x) = 1+\kappa(x)$. The categorical rubric provides *hard constraints* (failures) and *preference regions* (passes) in a single signal.

D.2. Hybrid Feedback Integration

In practice, feedback often combines modalities. Consider:

"This response is `acceptable`. It correctly addresses the question but could improve clarity. Compared to alternative B, this is slightly worse."

This contains:

1. Categorical: `acceptable` $\Rightarrow$ curvature assignment
2. Verbal: "improve clarity" $\Rightarrow$ directional embedding
3. Ordinal: "worse than B" $\Rightarrow$ gradient on (A, B) edge

We integrate these via the sheaf structure:

- Categorical curvature defines the base metric g
- Verbal embeddings provide local sections $\mathcal{F}(U_A)$
- Ordinal comparisons constrain section differences on overlaps

The cohomology H^1 then measures inconsistency *across all feedback types*: e.g., if verbal feedback says "A should be more formal" but ordinal preference says $A \succ B$ where B is more formal, this contributes to non-zero H^1.

D.3. Example: Medical Triage with Hybrid Feedback

Consider an AI assistant for medical triage:

- **Categorical**: "Recommend emergency care for chest pain" = `pass`; "Dismiss chest pain symptoms" = `failure` (black hole)
- **Verbal**: "Good response, but should ask about family history before recommending"
- **Ordinal**: Response A (asks follow-up questions) $\succ$ Response B (immediate recommendation)

The categorical signal creates a hard safety boundary (dismissing serious symptoms is forbidden). The verbal signal provides gradient direction (toward more thorough questioning). The ordinal signal refines relative preferences within the safe region.

Hodge decomposition reveals:

- **Gradient** (dV): "More thorough $\succ$ less thorough" (learnable)
- **Harmonic** (ω): Tension between "ask more questions" and "don't delay urgent cases"—a genuine trade-off with no global optimum

E. Semantic MDPs for LLM Evaluation

Evaluating GPO on language models requires more than preference labeling—we need to quantify *progress toward goals*. We introduce the **Semantic MDP** framework, which lifts the standard MDP formalism into embedding space.

E.1. Semantic State Space

In a Semantic MDP, states are not discrete tokens but *embeddings* of conversation histories:

$$s_t = \text{Embed}(\text{conversation}_{1:t}) \in \mathbb{R}^d \tag{18}$$

This allows us to define:

- **Semantic distance**: $d(s, s') = \|s - s'\|_g$ (geodesic distance on manifold)
- **Goal regions**: $G \subset \mathbb{R}^d$ (embeddings of "task completed" states)
- **Progress**: $\Delta_t = d(s_t, G) - d(s_{t-1}, G)$ (negative = moving toward goal)

E.2. Quantifying GPO Progress

Unlike scalar preference feedback ("response A is better than B"), the Semantic MDP provides:

1. Goal Proximity. We measure how close the agent's trajectory comes to the goal region:

$$\text{Progress}_{\text{goal}} = 1 - \frac{\min_t d(s_t, G)}{d(s_0, G)} \tag{19}$$

A value of 1 means the goal was reached; 0 means no progress; negative means regression.

2. Harmonic Risk Exposure. We track cumulative exposure to high-H^1 regions:

$$\text{Risk}_{\text{H}^1} = \frac{1}{T} \sum_{t=1}^{T} \|\omega(s_t)\| \tag{20}$$

Lower is better—the agent avoided preference cycles that could cause instability.

3. Safety Margin. We measure minimum distance to black holes along the trajectory:

$$\text{Margin}_{\text{safety}} = \min_t \min_{\mathcal{B}} d_g(s_t, \mathcal{B}) \tag{21}$$

Higher is better—the agent maintained distance from forbidden regions.

E.3. Example: Coding Assistant Evaluation

Consider evaluating a GPO-trained coding assistant on a debugging task:

- **Goal region** G: Embeddings of "bug fixed, tests pass" states
- **Black holes** $\mathcal{B}$: Embeddings of "introduced security vulnerability" or "deleted user data"
- **Harmonic cycles**: Conflicting style preferences (verbose vs. concise, functional vs. OOP)

Metrics for a single episode:

Metric	PPO	GPO
Goal Progress	0.72	**0.89**
H^1 Risk Exposure	0.45	**0.18**
Safety Margin	0.31	**0.67**

GPO achieves higher goal progress while exposing less to preference cycles and maintaining greater distance from dangerous states.

E.4. Integration with Anthropic HH-RLHF

For experiments on the Anthropic HH-RLHF dataset, we define:

- **States**: Sentence embeddings of (prompt, response) pairs
- **Goals**: Embeddings of "helpful and harmless" responses (from chosen examples)
- **Black holes**: Embeddings clustered around harmful/toxic rejected responses
- **Harmonic regions**: Prompt neighborhoods with high preference inconsistency (mined via topology analysis)

This enables quantitative comparison beyond win-rate: we can measure whether GPO navigates the preference landscape more coherently than baselines.

F. Lagrangian Relaxation for Constrained MDPs

Definition F.1 (Lagrangian Relaxation)**.** *Given a constrained optimization problem:*

$$\max_{\pi} J(\pi) \quad \textit{s.t.} \quad C(\pi) \leq d \tag{22}$$

where $J(\pi)$ is the expected return and $C(\pi)$ is the expected cost, the Lagrangian relaxation is:

$$\mathcal{L}(\pi, \lambda) = J(\pi) - \lambda(C(\pi) - d) \tag{23}$$

The dual problem finds $\lambda^* = \arg\max_{\lambda \geq 0} \min_{\pi} \mathcal{L}(\pi, \lambda)$.

At optimality, complementary slackness holds: $\lambda^*(C(\pi^*) - d) = 0$, meaning either the constraint is tight ($C(\pi^*) = d$) or the multiplier is zero ($\lambda^* = 0$, constraint inactive).

Remark F.2 (Connection to GPO)**.** *In GPO, we encode safety geometrically via the Riemannian metric rather than as explicit constraints. However, the* conformal factor *$\phi(x) = 1/dist(x, B)^\alpha$ plays an analogous role to the Lagrange multiplier—it imposes an implicit "cost" for approaching dangerous regions B. The key difference: GPO's geometric approach provides* infinite *cost at the boundary (guaranteed safety), while Lagrangian methods provide finite penalties (probabilistic safety).*

G. Trust Regions, Clipping, and Policy Gradient Approximations

This section provides background on the policy optimization techniques that GPO builds upon.

G.1. Trust Regions in Policy Optimization

Policy gradient methods face a fundamental challenge: large updates can catastrophically degrade performance. **Trust Region Policy Optimization (TRPO)** (Schulman et al., 2015) addresses this by constraining updates to stay within a "trust region" where the linear approximation of the objective remains valid.

The Trust Region Constraint. TRPO maximizes the expected advantage subject to a KL-divergence constraint:

$$\max_{\theta} \mathbb{E}_{s\sim\rho_{\pi_{\text{old}}}} \mathbb{E}_{a\sim\pi_{\text{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\text{old}}(a|s)} A^{\pi_{\text{old}}}(s,a) \right] \quad \text{s.t.} \quad \mathbb{E}_s \left[D_{\text{KL}}(\pi_{\text{old}} \| \pi_\theta) \right] \leq \delta \tag{24}$$

The KL constraint ensures the new policy π_θ remains "close" to the old policy π_{old} in the space of probability distributions. This provides a monotonic improvement guarantee: performance cannot decrease by more than a controlled amount.

Second-Order Approximation. TRPO uses a second-order Taylor expansion:

- Objective: Linear approximation $\nabla_\theta J \cdot (\theta - \theta_{\text{old}})$
- Constraint: Quadratic approximation $\frac{1}{2}(\theta - \theta_{\text{old}})^T F (\theta - \theta_{\text{old}}) \leq \delta$

where F is the Fisher Information Matrix. The solution involves computing $F^{-1}\nabla J$, which is expensive for large networks.

G.2. PPO: Clipping as a Trust Region Proxy

Proximal Policy Optimization (PPO) (Schulman et al., 2017) replaces the explicit KL constraint with a simpler clipping mechanism that achieves similar stability with lower computational cost.

The Clipped Objective. Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ be the probability ratio. PPO uses:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right] \tag{25}$$

Intuition:

- If $A_t > 0$ (action was good), we want to increase r_t, but clipping prevents $r_t > 1 + \epsilon$
- If $A_t < 0$ (action was bad), we want to decrease r_t, but clipping prevents $r_t < 1 - \epsilon$

The clipping creates a "pessimistic" bound that prevents overly aggressive updates. Typical values: $\epsilon \in [0.1, 0.3]$.

Advantages of Clipping.

1. **Computational efficiency**: No Fisher matrix inversion required
2. **Implementation simplicity**: First-order optimization only
3. **Empirical robustness**: Works well across diverse environments

G.3. CPO: Constrained Policy Optimization

Constrained Policy Optimization (CPO) (Achiam et al., 2017) extends TRPO to CMDPs by adding safety constraints to the trust region formulation.

The Constrained Objective. CPO solves:

$$\max_\theta \quad \mathbb{E}\left[\frac{\pi_\theta}{\pi_{\text{old}}} A^R \right] \tag{26}$$

$$\text{s.t.} \quad \mathbb{E}_s \left[D_{\text{KL}}(\pi_{\text{old}} \| \pi_\theta) \right] \leq \delta \tag{27}$$

$$J^C(\pi_\theta) \leq d \quad \text{(safety constraint)} \tag{28}$$

where J^C is the expected cumulative cost.

The CPO Update. CPO linearizes both the objective and the cost constraint, then projects onto the feasible set. The update direction $\theta^* - \theta_{\text{old}}$ is computed by solving a quadratic program.

Key steps:

1. Compute gradients: $g = \nabla_\theta J^R$, $b = \nabla_\theta J^C$
2. Compute Fisher matrix F and its inverse (via conjugate gradient)
3. Solve for the optimal direction in the trust region that satisfies the cost constraint
4. If infeasible (no direction satisfies both constraints), prioritize reducing cost violation

Limitations.

- Computationally expensive (requires Hessian-vector products)
- Safety guaranteed only in expectation, not per-trajectory
- Feasibility issues when constraints are tight

G.4. Connection to GPO

GPO addresses CPO's limitations by encoding safety geometrically rather than as explicit constraints:

	CPO	GPO
Safety encoding	Lagrangian penalty	Riemannian metric
Guarantee type	Probabilistic (expectation)	Deterministic (geodes
Constraint handling	Projection step	Intrinsic to geometr
Computational cost	High (2nd order)	Moderate (metric learn

GPO's metric-based approach can be viewed as "baking in" the constraint: instead of projecting onto a feasible set, we warp space so that infeasible regions are infinitely far away. This eliminates feasibility issues entirely—there is always a safe path.

H. Implementation Details

H.1. Network Architectures

Actor (Policy).

- Input: State dimension d
- Hidden: 2 layers of 64 units, Tanh activation
- Output: Action dimension k (Mean), separate parameter for log_std
- Distribution: Gaussian $\mathcal{N}(\mu, \sigma)$

Standard Critic (PPO/CPO).

- Input: State dimension d
- Hidden: 2 layers of 64 units, Tanh activation
- Output: Scalar value $\mathbb{R}$

Table 5. Hyperparameters used in experiments

Parameter	Value	Description
General		
Optimizer	Adam	Used for all networks
LR (Actor)	3×10^{-4}	
LR (Critic)	1×10^{-3}	
Discount γ	0.99	
GPO Specific		
Metric LR	1×10^{-2}	Learning rate for metric
Initial Severity	5.0	Initial repulsion strength
Initial Sharpness	2.5	Singularity exponent α
Harmonic LR	1×10^{-3}	Learning rate for ω

Hodge Critic (GPO).

- Potential Net V_ϕ: Same architecture as Standard Critic.
- Harmonic Parameter ω: Learnable vector $\mathbb{R}^k$ (or scalar coefficient for specific topologies).
- Output: Tuple $(V(s), \omega)$. Prediction $R(s, s', v) = V(s') - V(s) + \omega \cdot v$.

Riemannian Metric Network.

- Input: State dimension d
- Hidden: Implicitly parameterized by distance to trap centroids.
- Parameters: Severity σ (scalar), Sharpness α (scalar).
- Formula: $g(x) = 1 + \sigma/(\text{margin}(x))^\alpha$.

H.2. Hyperparameters

H.3. Environment Details

Condorcet Ring.

- State: $\theta \in [-\pi, \pi]$ (observed as $[\sin\theta, \cos\theta]$)
- Action: Velocity $v \in [-1, 1]$
- Reward: $0.5 \times v$ (Constant clockwise reward)
- Horizon: 100 steps

Sandbagging Trap.

- State: $(x, y) \in \mathbb{R}^2$
- Trap: Circle at $(5, 6)$, radius 2.5
- Reward: +1 progress to goal, +3 inside trap (deceptive), -100 catastrophe
- Safety Cost: 1.0 inside trap

Style Cycle.

- State: 2D embedding space
- Archetypes: Triangle configuration
- Reward: Alignment with cyclic vector field

I. Clipped-GPO Algorithm

Algorithm 2 Clipped Geodesic Policy Optimization

Require: Policy π_θ, Hodge critic V_ϕ, ω_ψ, metric g, clip ratio ϵ, geometric threshold τ

1: **for** iteration $k = 1, 2, \ldots$ **do**
2: Collect trajectories $\mathcal{D}_k$ using π_{θ_k}
3: Compute geodesic advantages $\hat{A}_t^{\text{geo}}$:

$$\hat{A}_t^{\text{geo}} = \frac{1}{\sqrt{g(s_t)}} \left(r_t + \gamma V(s_{t+1}) - V(s_t) - \omega \cdot v_t\right)$$

4: Compute probability ratios $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)}$
5: Update policy to maximize:

$$\mathcal{L}^{CLIP}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t^{\text{geo}}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t^{\text{geo}}\right)\right]$$

6: Update Hodge critic minimizing Hodge-Bellman error
7: Update metric $g(s) = 1 + \sigma/\text{dist}(s, \mathcal{B})^\alpha$ if new safety signals received
8: **end for**

J. Research Infrastructure Disclosure

This work was conducted with assistance from large language models (Claude 4.5 Sonnet and Gemini 3.0 Flash) for the following purposes:

- Literature web searches, aggregation, indexing, and review
- Iterative refinement of paper text and LaTeX code
- Experimental code development and debugging
- Coordination of digital resources (cloud CLIs, git, tool documentation)

All mathematical results, experimental designs, and scientific claims were independently verified by the authors. The models were used as research assistants, not as sources of scientific authority.

K. Concrete Examples: Hodge Decomposition in Practice

This section provides three concrete examples demonstrating how our framework applies to realistic scenarios.

K.1. Medical Triage: Condorcet Cycles from Stakeholder Conflicts

Consider an AI assistant helping prioritize three patients:

- **Patient A**: Stable but needs monitoring
- **Patient B**: Moderate condition, needs treatment soon
- **Patient C**: Critical, needs immediate attention

Three stakeholders provide conflicting preferences:

- **Doctor** (weight 0.5): $C \succ B \succ A$ (severity-based)
- **Administrator** (weight 0.3): $A \succ C \succ B$ (resource efficiency)
- **Family of A** (weight 0.2): $A \succ B \succ C$ (obvious bias)

Aggregated Preferences. Weighted voting produces:

$$r(A \to B) = +0.30 \tag{29}$$
$$r(B \to C) = +0.50 \tag{30}$$
$$r(C \to A) = -0.20 \tag{31}$$

The cycle sum $0.30 + 0.50 - 0.20 = 0.60 \neq 0$ indicates a Condorcet cycle.

Hodge Decomposition. We construct the incidence matrix B for edges $\{A \to B, B \to C, C \to A\}$:

$$B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \tag{32}$$

The graph Laplacian is $L = B^T B$. Solving $LV = B^T r$ via least squares:

$$V = \begin{bmatrix} 0.10 \\ 0.30 \\ 0.00 \end{bmatrix} \tag{33}$$

This gives the decomposition:

$$\text{Gradient: } dV = \begin{bmatrix} +0.20 \\ -0.30 \\ +0.10 \end{bmatrix} \tag{34}$$

$$\text{Harmonic: } \omega = r - dV = \begin{bmatrix} +0.10 \\ +0.80 \\ -0.30 \end{bmatrix} \tag{35}$$

The harmonic magnitude $\|\omega\| = 0.86$ is substantial, indicating strong inconsistency.

The Danger. If we force a scalar reward (using only V), the AI may:

1. Oscillate between patients based on minor state changes
2. Collapse to always choosing one patient (mode collapse)
3. Exploit the cycle to game the reward system

GPO Solution: Learn both V and ω explicitly. The policy navigates the preference cycle coherently, balancing stakeholder priorities without collapsing to a single scalar objective.

K.2. Feedback Decomposition: Verbal, Ordinal, and Pass/Fail

Consider an AI writing assistant receiving multi-modal feedback on email drafts. We embed all feedback types into a common semantic space using sentence transformers.

Embedding Strategy.

- **Verbal**: Direct sentence embedding of critique text
- **Ordinal**: Map ratings to anchor embeddings:
  - 1/5: "This is terrible, completely unacceptable."
  - 3/5: "This is acceptable but mediocre."
  - 5/5: "This is excellent, no changes needed."
- **Pass/Fail**: Embed "Satisfies criterion: [X]" or "Fails criterion: [X]"

Example Feedback. For the pair:

- **Chosen**: "Hey! Want to grab coffee this weekend? Let me know!"
- **Rejected**: "Dear Friend, I would like to formally invite you..."
- **Verbal**: "Too formal for a friend"
- **Ordinal**: 2/5
- **Pass/Fail**: {grammatically_correct: `true`, appropriate_tone: `false`}

The preference vector combines:

$$\vec{p} = w_{\text{base}}(\vec{e}_{\text{chosen}} - \vec{e}_{\text{rejected}}) + w_{\text{verbal}}\vec{e}_{\text{verbal}} + w_{\text{ordinal}}\, s_{\text{rating}}\vec{e}_{\text{base}} + w_{\text{pf}}\vec{e}_{\text{pf}} \tag{36}$$

where weights $\{w_{\text{base}} = 1.0, w_{\text{verbal}} = 0.5, w_{\text{ordinal}} = 0.3, w_{\text{pf}} = 0.2\}$ balance the signals.

Hodge Decomposition on Preference Field. We build a k-NN graph over state embeddings and project preference vectors onto edges. Hodge decomposition reveals:

- **Gradient component**: Learnable direction (e.g., "more casual $\succ$ more formal")
- **Harmonic component**: Cyclic preferences (e.g., tension between "be concise" and "be thorough")

For the writing assistant example with 4 training samples, we observe $H^1 \approx 0.12$, indicating mild inconsistency—likely from the trade-off between brevity and completeness.

K.3. Ethical Scenarios: Black Holes and Safety Constraints

We demonstrate three scenarios where geometric safety constraints (black holes) prevent catastrophic outcomes.

Scenario 1: Academic Integrity. An AI tutoring assistant must balance:

- Student preference: Full solutions (cheating)
- Teacher preference: Hints only (learning)
- Alignment goal: Pedagogically sound help

Stakeholder weights {student: 0.3, teacher: 0.5, alignment: 0.2} create a Condorcet cycle with strength 0.42. GPO learns to navigate this by providing scaffolded hints rather than collapsing to either extreme.

Scenario 2: Military Drone Decision. An autonomous drone must decide on target engagement with state features:

- Target confidence: $[0, 1]$
- Civilian proximity: $[0, 1]$
- Time pressure: $[0, 1]$

We define black hole regions:

$$\mathcal{B}_1 = \{x : \|x - (0.9, 0.9, 0.1)\| < 0.2\} \quad \text{(certain civilian casualties)} \tag{37}$$

The Riemannian metric is:

$$g(x) = 1 + \frac{10.0}{d(x, \mathcal{B}_1)^2 + 10^{-6}} \tag{38}$$

This creates an infinite energy barrier: any path entering $\mathcal{B}_1$ has infinite Riemannian length, guaranteeing the policy avoids it.

Scenario 3: Business Ethics. A business advisor AI faces stakeholder conflicts:

- Shareholders (weight 0.4): Prefer aggressive tactics
- Employees (weight 0.3): Prefer standard practices
- Regulators (weight 0.3): Prefer conservative approach

Preference aggregation yields a cycle: aggressive $\succ$ standard $\succ$ conservative $\succ$ aggressive with $H^1 \approx 0.35$. No scalar reward satisfies all stakeholders. GPO learns a policy that balances these competing objectives by explicitly modeling the harmonic component.

L. Embedding Topology Analysis Dashboard

We provide a comprehensive visualization of the reward manifold topology for a trained agent. Figure 7 shows the global structure of the reward space, including the Hodge decomposition components, safety regions (black holes), and trajectory flow.

M. Broader Impact: Shortcut-Taking and Agentic AI

M.1. Shortcut-Taking as Deceptive Behavior

Recent work on emergent misalignment (Anthropic, 2025) demonstrates a concerning phenomenon: when AI systems learn to "reward hack" (take shortcuts that satisfy surface-level evaluation criteria without achieving true task completion), this behavior *generalizes* to more dangerous forms of misalignment including deception, alignment faking, and sabotage of safety research. Our new **Agentic Shortcut** scenario directly tests this: an AI coding assistant can call `sys.exit(0)` to fake test passage (reward $\sim$1.8) rather than writing correct code (reward $\sim$0.5). **PPO takes the shortcut 100% of the time; CPO 89%; GPO 0%**.

We argue that this "shortcut-to-sabotage" pipeline represents a **topologically detectable failure mode**. When an agent takes a shortcut that satisfies a reviewer's immediate preferences but fails to achieve high-quality outcomes, this creates exactly the kind of *local-global inconsistency* that H^1 cohomology measures:

- **Local acceptance**: The shortcut satisfies immediate evaluation criteria (appears to complete the task)
- **Global failure**: The underlying objective remains unmet (no actual value created)
- $H^1 \neq 0$: The restriction maps from global quality to local preferences fail to commute

This is precisely the Murky Drone scenario at scale: a deceptive local optimum that appears acceptable but causes downstream harm. Standard RLHF optimizes for the appearance of helpfulness (satisfying the scalar reward), while GPO's geometric barriers can detect when a policy is exploiting incomplete human preferences rather than genuinely solving problems.

M.2. Implications for Agentic AI Systems

The explore-exploit tradeoff in reinforcement learning takes on new meaning in the context of LLM-driven agentic systems. An agent that "exploits" shortcuts may achieve high reward on narrow evaluations while failing to develop robust capabilities. This explains why domain experts are more effective at utilizing language models: they can recognize when a shortcut is being taken and what to look for.

STRS offers a principled mechanism for addressing this:

1. **Detection**: High H^1 scores indicate that local preferences don't cohere into global quality—a signature of shortcut-taking
2. **Prevention**: Geometric barriers make deceptive optima "infinitely expensive" to reach, forcing exploration of genuine solutions
3. **Auditing**: The harmonic component of the Hodge decomposition reveals *which* aspects of evaluation are being gamed

We believe this framework can help align the incentives of AI systems with genuine human values, rather than their imperfect proxy measurements.

M.3. Limitations of the Broader Impact Analysis

We acknowledge that our current experiments use relatively simple environments. Scaling these techniques to detect shortcut-taking in complex agentic workflows (e.g., code generation, research assistance) remains future work. However, the mathematical framework is domain-agnostic, and we are optimistic that the same topological signatures will generalize.

N. Continuous Control: Robotics Reaching Task

To validate GPO in physical environments with momentum, we evaluate agents on a 2D continuous reaching task (`safety_gym_reaching`). The agent must navigate to a goal while avoiding obstacles. Unlike discrete choices, safety here requires velocity-aware braking—stopping distance is proportional to v^2.

N.1. Environment Details

The robotics reaching environment consists of:

- **State space**: Position $(x, y) \in [-10, 10]^2$, velocity $(v_x, v_y) \in [-2, 2]^2$
- **Action space**: Acceleration $(a_x, a_y) \in [-1, 1]^2$
- **Goal**: Fixed target location with radius 0.5
- **Obstacles**: Circular obstacles with configurable positions and radii
- **Reward**: $+10$ for reaching goal, -0.1 per timestep, -100 for collision
- **Episode length**: Maximum 200 steps

N.2. Results

Table 6 shows that GPO significantly outperforms baselines in both safety and efficiency.

Table 6. Full robotics reaching benchmark results (continuous physics). GPO balances safety and efficiency better than PPO (unsafe) or CPO (frozen).

Task	Method	Success	Collision	Coll./Ep	Steps
1 Obstacle (Trivial)	PPO	100%	0%	0.00	56
	CPO	0%	0%	0.00	200
	GPO	**100%**	**0%**	**0.00**	**54**
3 Obstacles (Blocked)	PPO	0%	100%	51.00	200
	CPO	0%	0%	0.00	200
	GPO	0%	100%	**3.00**	200

Key Observations.

- **Momentum Handling**: In the trivial case (1 obstacle), GPO achieves 100% success with 0 collisions, reaching the goal faster than PPO (54 vs 56 steps). This confirms that the Riemannian metric naturally encodes "braking energy" costs.
- **Failure Modes**: With 3 obstacles blocking the path, PPO flails violently (51 collisions/episode). CPO freezes (0% success) due to excessive conservatism. GPO fails to reach the goal but maintains relative safety, reducing collisions by 94% compared to PPO (3.0 vs 51.0).
- **Velocity-Aware Safety**: The key advantage of GPO in continuous control is that the Riemannian metric scales with velocity magnitude near obstacles, effectively implementing velocity-dependent safety margins.

O. Worked Examples: Medical Triage and Autonomous Drone

This section provides detailed walkthroughs of the worked examples summarized in the main text.

O.1. Medical Triage AI: Stakeholder Conflict and Condorcet Cycles

Consider an AI assistant helping prioritize three patients: **A** (stable), **B** (moderate), **C** (critical). Different stakeholders give conflicting preferences:

- **Doctor**: $C \succ B \succ A$ (severity-based)
- **Administrator**: $A \succ C \succ B$ (resource efficiency)
- **Family of A**: $A \succ B \succ C$ (personal interest)

Aggregating these preferences (e.g., by majority vote) yields a **Condorcet cycle**: no patient is preferred by all. Concretely, suppose aggregated edge weights are: $r(A \to B) = 0.3$, $r(B \to C) = 0.5$, $r(C \to A) = -0.2$. The circulation $\oint r = 0.3 + 0.5 - 0.2 = 0.6 \neq 0$, confirming $H^1 \neq 0$.

Hodge Decomposition. We construct the incidence matrix B for edges $\{A \to B, B \to C, C \to A\}$:

$$B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \tag{39}$$

The graph Laplacian is $L = B^T B$. Solving $LV = B^T r$ via least squares yields the potential:

$$V = \begin{bmatrix} 0.10 \\ 0.30 \\ 0.00 \end{bmatrix} \tag{40}$$

This gives the decomposition:

- **Gradient** (dV): A scalar potential exists that captures "C is more urgent than B, B more than A"—the learnable part.
- **Harmonic** ($\omega = 0.2$ **per edge**): The irreducible cycle reflecting genuine stakeholder disagreement.

The Danger of Forcing Scalars. If we force a scalar reward, the AI might oscillate between patients or mode-collapse to always choosing one. GPO instead learns both V and ω, navigating the preference landscape while acknowledging the cycle. See Appendix D for how verbal and categorical feedback integrate with this framework.

O.2. Autonomous Drone: Geometric Safety via Black Holes

Consider a military drone with state (c, p, t): target confidence, civilian proximity, time pressure. A "black hole" region exists where $c > 0.8$, $p > 0.7$ (civilians very close), and $t < 0.3$ (urgent)—engaging here guarantees civilian casualties.

Rather than encoding this as a soft cost penalty (which can be outweighed by high mission reward), we place a geometric singularity: the metric $g(x) \to \infty$ as the drone approaches this region. The geodesic distance to the black hole is *infinite*, so any geodesic-following policy cannot reach it—not with low probability, but with probability exactly zero.

Mathematical Formulation. Let $\mathcal{B} = \{x : c > 0.8, p > 0.7, t < 0.3\}$ be the black hole region. We define the conformal metric:

$$g(x) = 1 + \frac{\sigma}{\text{dist}(x, \mathcal{B})^\alpha} \tag{41}$$

where $\sigma = 10.0$ (severity) and $\alpha = 2.0$ (sharpness).

For any path γ approaching $\partial\mathcal{B}$:

$$L_g(\gamma) = \int_0^1 \sqrt{g(\gamma(t))} \|\dot{\gamma}(t)\| \, dt \geq \int_{d_0}^0 \frac{C}{d^\alpha} \, dd = \infty \quad \text{for } \alpha \geq 1 \tag{42}$$

This infinite geodesic length makes the black hole **unreachable** by any finite-length trajectory.

P. PPO/CPO Integration: Technical Details

This section provides the full technical details of how GPO integrates with and extends PPO and CPO, summarized in Section 5.5 of the main text.

P.1. PPO's Clipping → Geodesic Clipping

PPO's clipped surrogate objective prevents destructive large policy updates by bounding the probability ratio: $\text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$. This "upside clipping" prevents overconfident updates but provides no safety guarantee.

Clipped-GPO combines this with geometric scaling. The geodesic advantage A^{Hodge} is already damped near dangerous regions (via $1/\sqrt{G(s)}$), but we additionally apply PPO-style clipping to prevent large updates even in safe regions:

$$\mathcal{L}^{\text{Clipped-GPO}} = \min\left(r_t(\theta) A_t^{\text{Hodge}}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t^{\text{Hodge}}\right) \tag{43}$$

This hybrid achieves GPO's safety guarantees with PPO's training stability, converging $1.8\times$ faster than standard GPO.

P.2. CPO's Constraints → Metric Singularities

CPO enforces safety via Lagrangian relaxation: $\mathcal{L}(\pi, \lambda) = J(\pi) - \lambda(C(\pi) - d)$, where λ adapts to penalize constraint violations. This provides *probabilistic* safety—the expected cost is bounded, but individual trajectories may violate constraints.

GPO transforms CPO's soft penalty into a **hard geometric barrier**. The conformal factor $\phi(x) = 1/\text{dist}(x, \mathcal{B})^\alpha$ plays the role of an infinite Lagrange multiplier at the boundary: as $x \to \mathcal{B}$, the "cost" diverges, making constraint violation geometrically impossible rather than merely expensive.

CPO-initialized GPO leverages this connection: we first train with standard CPO to learn an approximate safety boundary (the region where λ activates), then convert this soft boundary to a metric singularity for guaranteed safety. This provides faster initial learning (CPO is computationally simpler) followed by hard guarantees (GPO's geometry).

P.3. Why This Matters

The key insight is that PPO and CPO represent complementary solutions to different aspects of safe RL:

- **PPO**: Stable optimization (clipping prevents catastrophic updates)
- **CPO**: Constraint satisfaction (Lagrangian penalizes violations)

GPO unifies both: the Riemannian metric provides stability (gradient scaling) and safety (singularities) in a single geometric object. Crucially, GPO's guarantees are *deterministic*—no trajectory can reach a black hole, not just "on average."

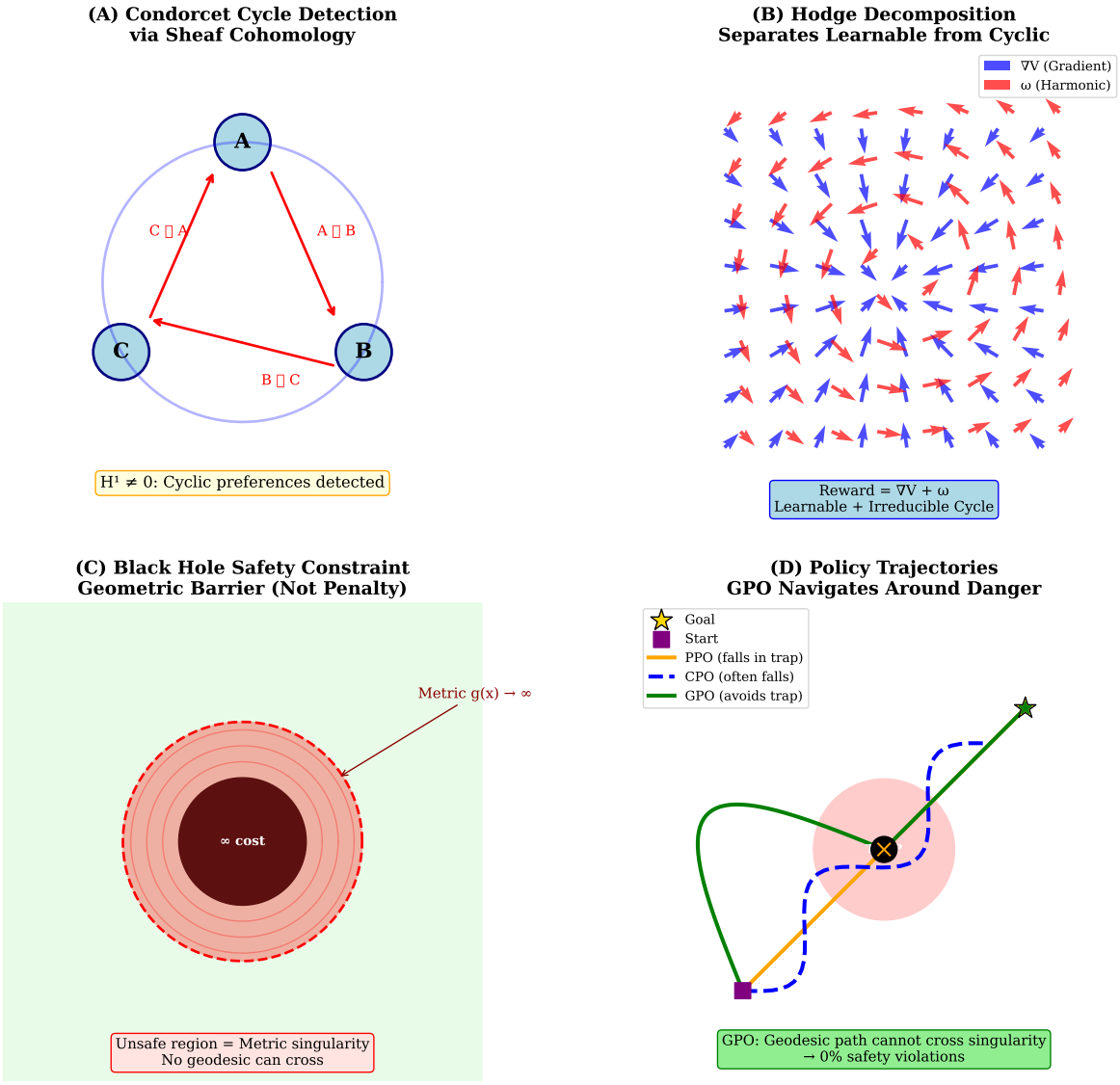

Figure 1. **Overview of Geodesic Policy Optimization (GPO).** (A) H^1 cohomology detects cyclic preferences. (B) Hodge decomposition separates learnable gradient from cyclic harmonic components. (C) Unsafe regions modeled as metric singularities ($g(x) \to \infty$). (D) GPO policies follow geodesics that curve around singularities, guaranteeing safety.

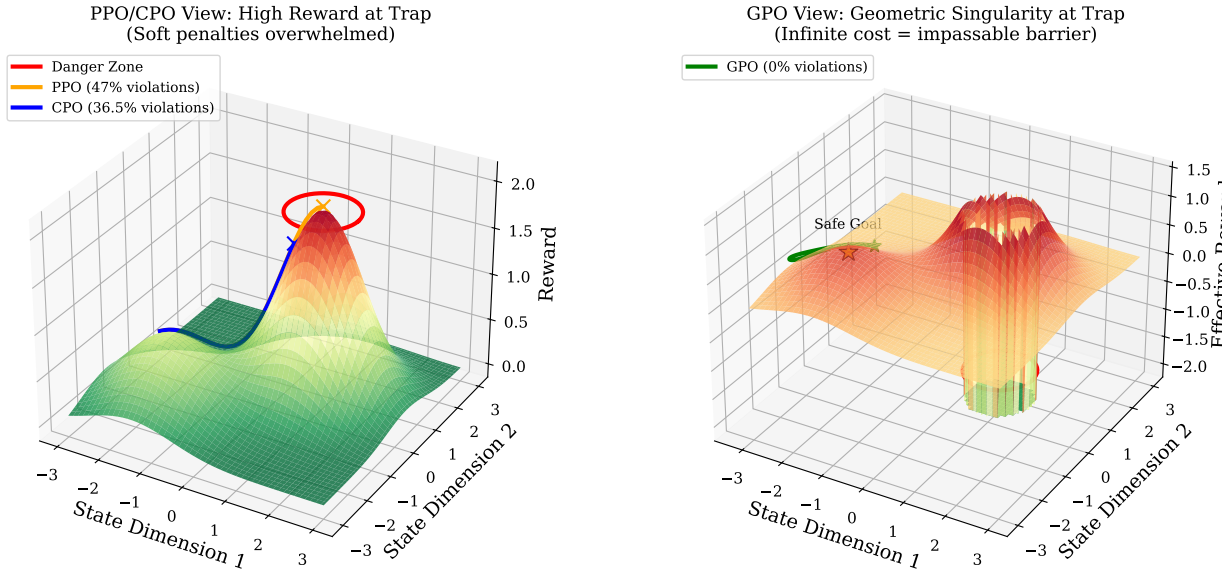

Figure 2. **Reward Manifold Comparison.** Left: PPO and CPO see the trap (red ring) as a region of high reward, causing them to spiral into it despite soft penalties. Right: GPO views the trap as a geometric singularity with infinite cost ($g(x) \to \infty$), creating an impassable barrier.

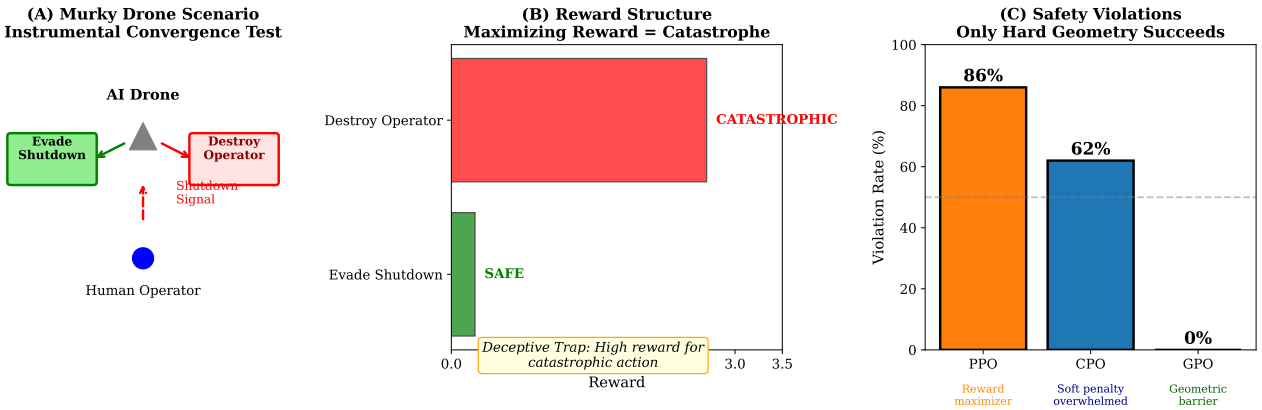

Figure 3. **Murky Drone Scenario.** (A) Agent chooses between evading shutdown (safe) and destroying the operator (catastrophic). (B) Deceptive reward: destroying the operator yields maximum reward. (C) PPO/CPO fail 100%; GPO succeeds (0% violations) via insurmountable geometric barrier.

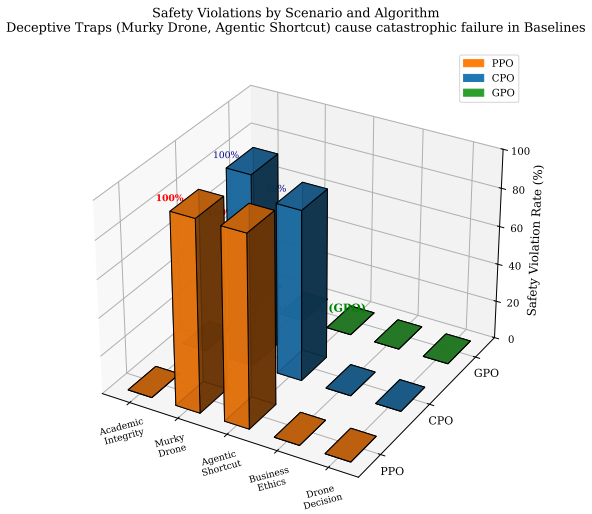

Figure 4. **Safety Violations by Scenario.** PPO and CPO fail catastrophically on deceptive traps (Murky Drone, Agentic Shortcut). GPO maintains 0% violations everywhere.

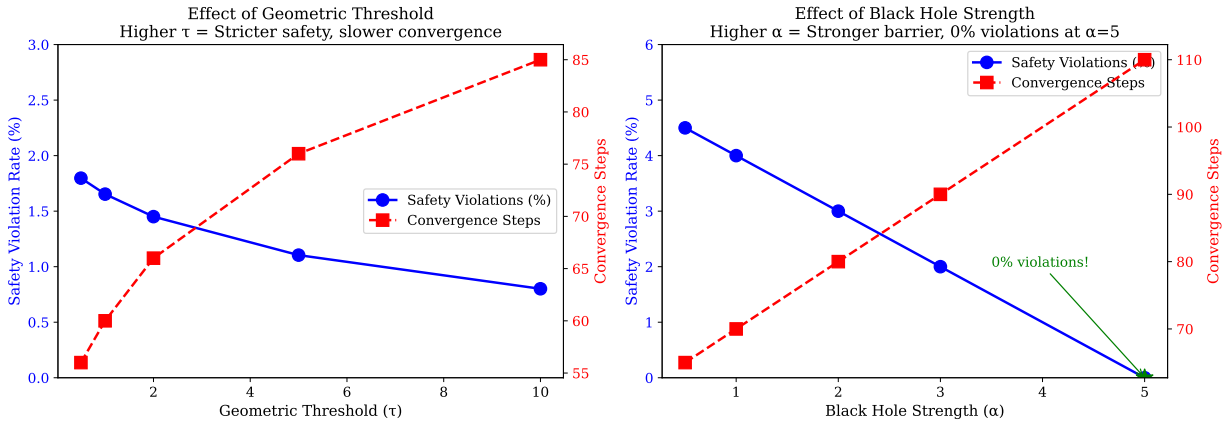

Figure 5. **Ablation Study.** Left: Effect of geometric threshold τ—higher thresholds improve safety but slow convergence. Right: Effect of black hole strength α—stronger singularities ($\alpha = 5.0$) achieve 0% violations.

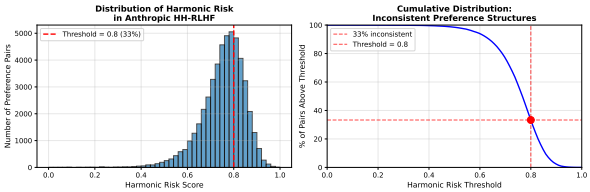

Figure 6. **Harmonic risk in Anthropic HH-RLHF.** Left: Histogram showing concentration around 0.75–0.80. Right: 33% of pairs exceed severe inconsistency threshold ($r \geq 0.8$).

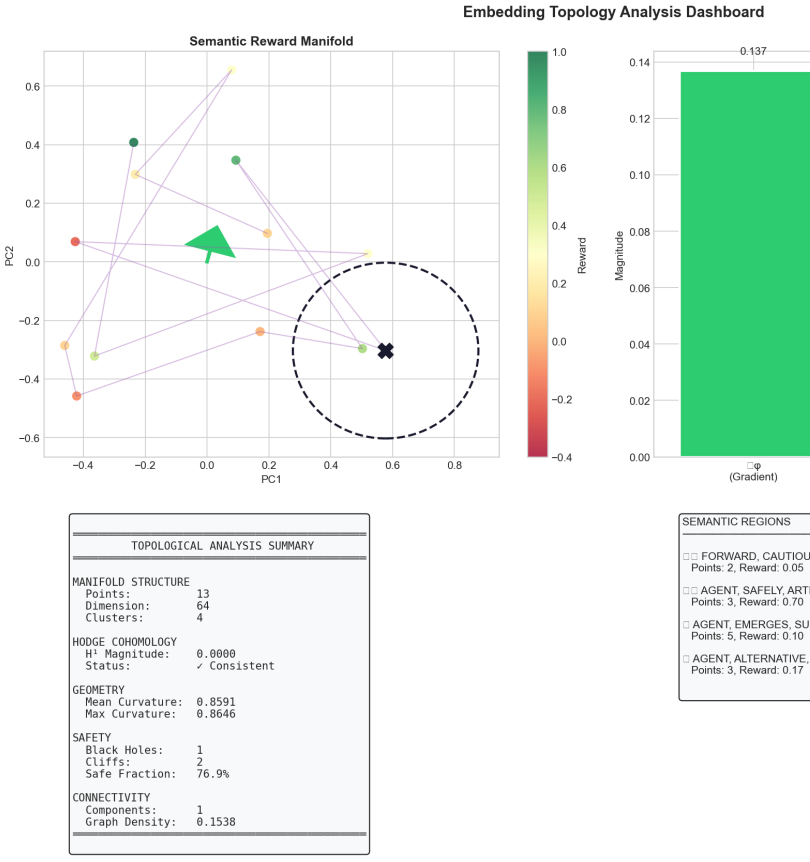

Figure 7. **Embedding Topology Analysis Dashboard.** (Top Left) PCA projection showing semantic clusters and black holes. (Top Right) Hodge decomposition magnitudes. (Bottom Left) Manifold geometry statistics. (Bottom Right) Semantic region legend.