Semantics-Aware Reward Comparison for Finite Multi-Objective Reinforcement Learning

Janmenjaya Panda Joar Skalse

Research snapshot: 5 September 2026
Software version `morl-starc` 0.1.0

Supervised Program for Alignment Research — Fall 2026

Abstract

Scalar reward distances cannot be transferred mechanically to multi-objective reinforcement learning (MORL), because a vector reward does not determine a single policy ordering until a decision semantics is declared. Fixed weighted sums, robust maximin utility, and Pareto dominance preserve different reward transformations, so they require different quotients. This paper gives a self-contained account of the MORL-STARC program for finite discounted Markov decision processes under scalarized expected-return semantics. It covers the motivating failures, the constructions, proofs, counterexamples, algorithms, numerical audits, the formal-verification boundary, benchmark evidence, negative results, and open problems.

For Pareto semantics, projected objective directions generate an objective normal cone on the feasible occupancy-difference subspace. Equality of these cones is equivalent to equality of the induced Pareto policy relation. For a nonzero pointed cone, the convex hull of its unit extreme rays is a canonical compact body, and Pareto-STARC is the Euclidean Hausdorff distance between two such bodies. Its zero set is exactly Pareto policy-relation equality on the declared domain. Its singleton restriction is projected scalar chord STARC, and it equals the uniform discrepancy between normalized extreme-ray margin functions. It yields a selector-specific strict-improvement certificate and upper-bounds directional worst-case dominance-reversal severity. A nested-cone construction shows that no universal positive one-sided completeness constant exists. That obstruction is a consequence of partial ordering, not of numerical implementation.

For anchored maximin semantics, objective anchors and units are normative data. The lower envelope of standardized scalarizations induces a concave piecewise-linear utility on the randomized-policy simplex. Cardinal Maximin-STARC is the uniform distance between globally range-normalized lower-envelope utilities. It is exactly complete for the declared cardinal utility error and yields a sharp factor-two selector-margin certificate. It is not an ordinal metric: strictly increasing lower envelopes can induce the same policy ordering while remaining at positive distance. A separate thin-active-facet construction shows that Hausdorff distance between active envelope members can be arbitrarily more conservative than functional utility discrepancy, even when every retained member is uniquely active somewhere.

The computational evidence consists of theorem-driven unit tests, randomized metric and reversal audits, an independent SciPy solver oracle in dimensions two through six, numerical-conditioning sweeps, a deterministic Deep Sea Treasure comparison, and Lean-checked algebraic kernels. These results give falsification evidence and internal consistency on finite tabular problems. They do not establish scalability, end-to-end formal verification, practical estimability from data, or broad empirical validity. Those limits, including adverse and null results, are reported explicitly.

Contents

1 Introduction

1.1 The problem

Let a fixed environment admit policies whose expected discounted vector returns are

$$z_R(\pi) = \big(J_{r_1}(\pi), \ldots, J_{r_k}(\pi)\big) \in \mathbb{R}^k. \tag{1}$$

A scalar reward induces a scalar utility and therefore a total preorder over policies. A vector reward alone does not. One may instead declare

1. a fixed scalarization $w^\top z_R(\pi)$;
2. an uncertain but fixed scalarization drawn from a set W;
3. a maximin utility $\inf_{w \in W} w^\top z_R(\pi)$;
4. Pareto dominance;
5. a lexicographic, constrained, or other selector.

These semantics disagree about both policy ordering and reward invariance. The central question is not what norm to place between two vector rewards. It is:

> *After fixing the environment, policy class, return semantics, and protected decision rule, what quotient of vector reward specifications represents the same policy behavior, and what graded distance controls a relevant operational loss?*

1.2 Main answer

There is no universal semantics-independent vector STARC. The program developed here gives three answers:

- **Fixed weight:** scalarize first and apply scalar STARC exactly.
- **Pareto:** compare canonical projected objective cones using Pareto-STARC.
- **Maximin:** preserve anchors and units, then compare normalized lower-envelope utilities using Cardinal Maximin-STARC.

The split is not an implementation choice. It follows from incompatible invariance groups.

1.3 Positive results

The principal proved results are:

1. The feasible discounted occupancy set is a compact convex polytope, and only reward components in its difference span affect policy comparisons (lemmas 4.1 and 4.2).
2. Fixed weighted sums reduce exactly to scalar reward comparison (theorem 6.1).
3. Pareto policy-relation equality is equivalent to equality of projected objective normal cones on one fixed finite environment (theorem 7.2).
4. Pareto-STARC is a pseudometric with exactly this semantic zero set on nonzero pointed finite objective cones (theorem 8.2).
5. Pareto-STARC reduces exactly to projected scalar chord STARC for singleton objective families (theorem 8.3).
6. Pareto-STARC equals uniform extreme-ray margin discrepancy (theorem 8.5).
7. The proxy margin minus the Pareto-STARC distance lower-bounds the true margin (theorem 9.1).
8. Directional worst-case Pareto reversal is upper-bounded by Pareto-STARC (theorem 9.3).

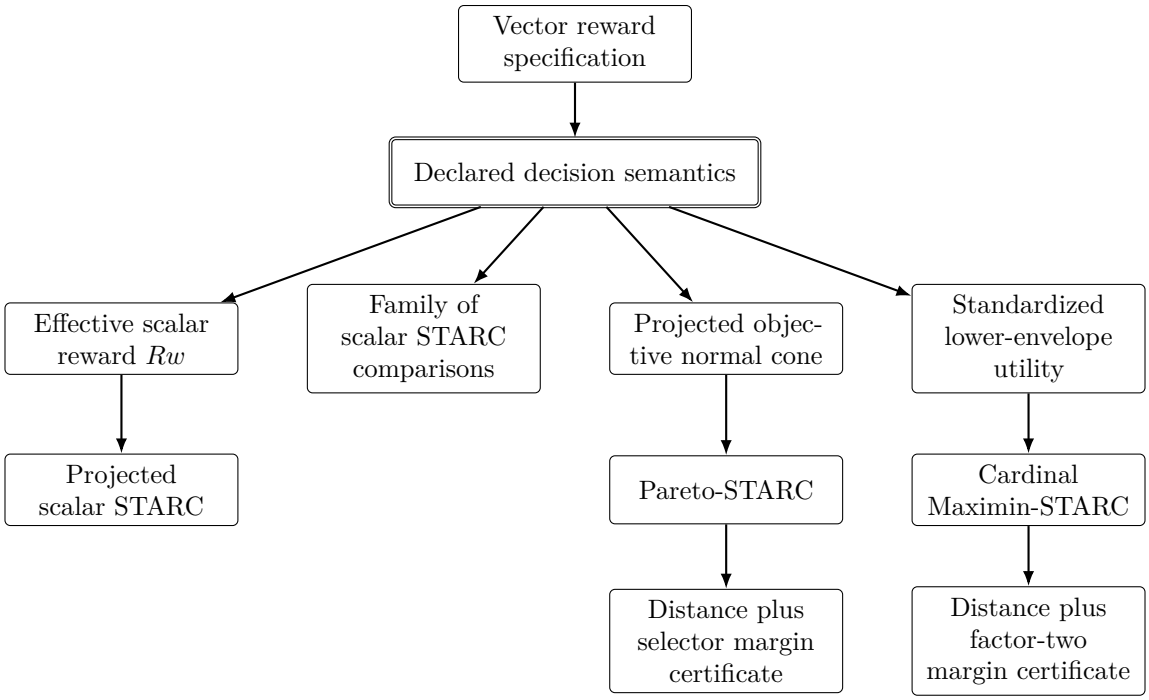

Figure 1: **Semantic routing.** The reward array is not canonicalized before semantics are selected. Fixed weights collapse to one scalar reward. Pareto dominance quotients positive-cone representation artifacts. Maximin retains relative levels and units before comparing lower-envelope utilities.

9. Cardinal Maximin-STARC is exactly the uniform discrepancy between globally normalized anchored maximin utilities (theorem 11.1).
10. A proxy maximin step with normalized utility margin greater than twice the distance is certified as a true improvement, and the factor two is sharp (theorem 11.2 and proposition 11.3).

1.4 Negative results

The negative results carry equal weight:

1. Objective-specific translation and scaling preserve Pareto dominance but can reverse maximin preferences (sections 5.3 and 5.4).
2. Family-STARC may be zero while maximin behavior changes (section 6.3).
3. Identical unlabelled Pareto fronts may correspond to opposite policy choices (section 7.4).
4. Arbitrarily small objective perturbations can flip zero-margin Pareto labels (section 7.5).
5. Cardinal Maximin-STARC can be positive for utilities with identical ordering (theorem 11.5).
6. Active-member Hausdorff distance can exceed functional maximin discrepancy by an arbitrarily large factor (theorem 12.5).
7. Pareto-STARC can be positive while one-sided directional reversal loss is zero (theorem 9.4).
8. The experiments reported here do not establish broad practical superiority, scalability, learned-reward performance, or external validity (section 15.11).

1.5 Theorem dependency graph

Source theorem identifiers are kept in figure 2 so that the consolidated arguments can be traced to their authoritative chapters.

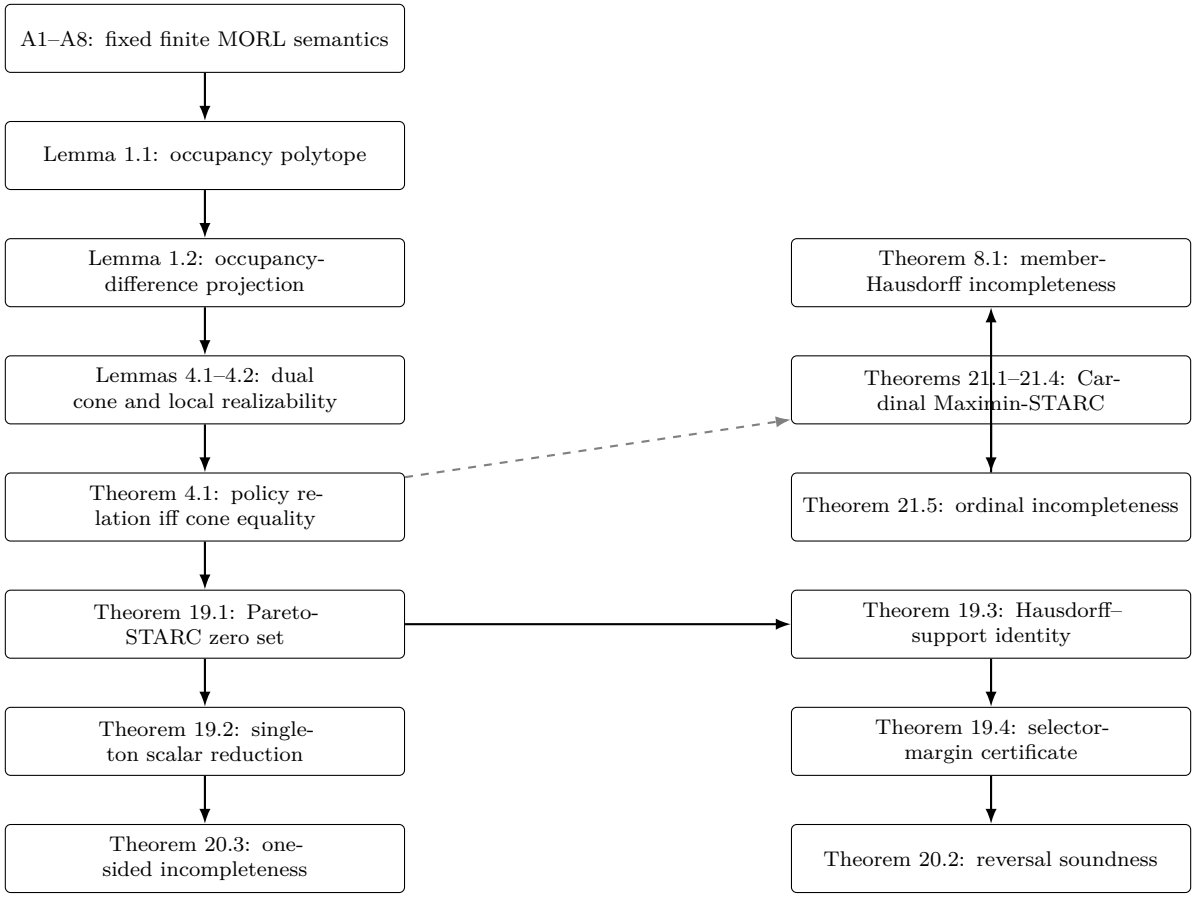

Figure 2: **Flagship theorem dependencies.** Occupancy geometry supports the Pareto semantic quotient. The canonical body then supports scalar reduction, the margin identity, and the reversal results. Maximin follows a separate generic-utility branch. The two impossibility results arise from different mechanisms: partial-order nesting for Pareto, and vanishing active regions for member-set maximin comparison.

2 Related Work and the Novelty Boundary

2.1 Scalar reward distances

EPIC canonicalizes scalar rewards against transformations such as potential shaping and positive scaling [Gleave et al., 2021]. STARC develops a dynamics-aware scalar reward comparison framework with soundness, completeness, and uniqueness results tied to policy regret [Skalse et al., 2024]. DARD also incorporates environment dynamics into reward comparison [Wulfe et al., 2022]. These are the direct antecedents of the present work, and none of them supplies a semantics-free metric for vector reward families.

For one fixed weight w, MORL creates the effective scalar reward Rw, and the scalar theory

applies unchanged. Applying a scalar distance objective by objective does not preserve the relative levels that maximin utility requires, and it does not by itself identify equality of a Pareto relation.

2.2 Existing MORL pseudometrics and utility theory

The MOQ pseudometric of Yang et al. [2019] takes a supremum over scalarized multi-objective value representations for Bellman analysis. It is an important MORL distance precedent, but it compares scalarized value-function objects rather than a dynamics-aware reward-family quotient by equality of Pareto policy relations.

Work on MORL utility representations establishes that a vector return has no unique total ordering without an explicit utility or relation [Rodriguez-Soto et al., 2024]. Max-min MORL develops optimization algorithms under cardinal robust semantics [Park et al., 2024, Byeon et al., 2025]. Pareto front traversal and set-quality indicators evaluate returned policy sets or fronts [Qiu et al., 2024, Park and Sung, 2025]. All of these ask different questions from comparing two reward specifications before policy selection.

2.3 Preference-relation and choice distances

Metrics on finite binary relations and choice correspondences are classical [Knoblauch, 2015, Klamler, 2008, Nishimura and Ok, 2024]. On a fixed finite policy set, the symmetric difference of ordered policy pairs is already zero exactly when the represented relations agree. Hausdorff topologies for preference relations and menu-wise choice distances likewise predate this work. The project therefore does not claim to be the first metric whose zero set is equality of a preference relation.

2.4 Convex-cone and support-function distances

Hausdorff, angular, spherical, truncated, and support-function distances between closed convex cones or compact convex bodies are prior art [Iusem and Seeger, 2010]. The bipolar theorem and the identity between Hausdorff distance and uniform support-function distance are classical finite-dimensional convex analysis. The mathematical primitive is not novel on its own.

2.5 What is claimed

The defensible claim is narrow:

> To our knowledge, Pareto-STARC is a dynamics-aware STARC-style pullback from finite vector reward specifications to Pareto policy-order equivalence on one fixed finite discounted environment. Its contribution is the composition of occupancy-difference reward projection, Pareto-specific cone reduction, exact singleton scalar reduction, and selector-margin and reversal guarantees.

Claims that are not supported include "the first MORL reward metric," "the first cone metric," "the first metric on partial preferences," "the unique vector STARC," and any assertion of scalar-style Pareto regret completeness. The detailed claim audit appears in the accompanying novelty audit and literature survey.

3 Mathematical Setting

3.1 Environment

Let

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \tau, \mu, \gamma) \tag{2}$$

be a fixed finite discounted Markov decision process, where $\mathcal{S}$ and $\mathcal{A}$ are finite, $\tau(s' \mid s, a)$ is the transition kernel, μ is the fixed initial-state distribution, and $0 < \gamma < 1$ in the theory. Unreachable states are removed.

The implementation accepts $0 \leq \gamma < 1$. The case $\gamma = 0$ is a valid one-step computational extension, but it is not silently included in theorem statements whose assumptions say $0 < \gamma < 1$.

3.2 Policy and return semantics

The policy class Π contains all stationary randomized policies. Deterministic stationary policies are its pure points. The analysis uses scalarized expected returns (SER): utility or Pareto comparison is applied after taking the expected discounted vector return. Expected scalarized returns (ESR), which require the distribution of trajectory returns, are outside scope.

Transition-dependent objectives may be replaced by their conditional expectations at each state-action pair. Write the resulting reward columns as

$$R = [r_1, \ldots, r_k] \in \mathbb{R}^{n \times k}, \qquad n = |\mathcal{S}|\,|\mathcal{A}|\,. \tag{3}$$

This replacement quotients next-state redistribution that leaves each state-action conditional expected reward unchanged.

3.3 Declared semantic data

The following are part of a comparison problem rather than incidental metadata: the environment, discount, and initial distribution; the policy class; SER rather than ESR; objective labels when weights refer to labels; the decision semantics; the preference uncertainty set, if present; objective anchors and units for maximin; and the norm and tolerance conventions for quantitative statements.

The main theorems assume exact rewards and occupancies. Statistical uncertainty must be represented by additional confidence sets or error terms.

4 Occupancy Geometry and Reward Projection

4.1 Discounted occupancies

For $\pi \in \Pi$, define its unnormalized discounted state-action occupancy by

$$\eta^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(S_t = s, A_t = a \mid \pi). \tag{4}$$

The occupancy set is $\Omega = \{\eta^\pi : \pi \in \Pi\}$. Every feasible occupancy satisfies $\eta(s, a) \geq 0$ and, for every $s' \in \mathcal{S}$,

$$\sum_a \eta(s', a) - \gamma \sum_{s,a} \tau(s' \mid s, a)\eta(s, a) = \mu(s'). \tag{5}$$

Summing equation (5) over s' gives

$$\sum_{s,a} \eta(s,a) = \frac{1}{1-\gamma}. \tag{6}$$

Lemma 4.1 (Occupancy-polytope structure)**.** Inherited result. Ω *is a nonempty compact convex polytope, and it is the convex hull of deterministic stationary-policy occupancies.*

Proof. The flow equations and nonnegativity constraints define a closed polyhedron. The fixed total mass $1/(1-\gamma)$ makes it bounded, hence compact. The usual occupancy-to-policy construction recovers a stationary randomized policy from each feasible flow by normalizing action masses at visited states. Conversely, every stationary policy generates such a flow. Extreme feasible flows correspond to deterministic stationary choices, so their convex hull is the full feasible set. □

4.2 Difference subspace

Choose η_{ref} in the relative interior of Ω, and define

$$L = \text{span}(\Omega - \Omega). \tag{7}$$

Every occupancy may be written as $\eta = \eta_{\text{ref}} + x$ with $x \in L$. Let P_L denote Euclidean orthogonal projection onto L. For objective r_i, define

$$b_i = r_i^\top \eta_{\text{ref}}, \qquad q_i = P_L r_i. \tag{8}$$

Lemma 4.2 (Reward decomposition)**.** *For every feasible centered occupancy x,*

$$r_i^\top(\eta_{\text{ref}} + x) = b_i + q_i^\top x. \tag{9}$$

Proof. Decompose $r_i = P_L r_i + (I - P_L) r_i$. Since $x \in L$, the second component is orthogonal to x, so $r_i^\top x = (P_L r_i)^\top x$. Adding the reference return proves the identity. □

Writing $Q = [q_1, \ldots, q_k]$ and $b = (b_1, \ldots, b_k)$ gives the affine return map

$$z_R(x) = b + Q^\top x. \tag{10}$$

Only Q affects policy differences. The offset b still matters to nonlinear level-sensitive utilities such as maximin.

4.3 Local realizability of every difference direction

Lemma 4.3 (Local realizability)**.** *For every nonzero $v \in L$ there is an $\epsilon > 0$ such that $\epsilon v \in \Omega - \Omega$.*

Proof. Because η_{ref} lies in the relative interior of Ω, a relative open ball around it lies in Ω. Choose $\epsilon > 0$ small enough that $\eta_{\text{ref}} + \epsilon v$ lies in that ball. Then $\epsilon v = (\eta_{\text{ref}} + \epsilon v) - \eta_{\text{ref}} \in \Omega - \Omega$. □

The lemma matters for the impossibility arguments. If two conic preference descriptions disagree on any direction in L, a scaled version of that disagreement is realized by a pair of feasible occupancies.

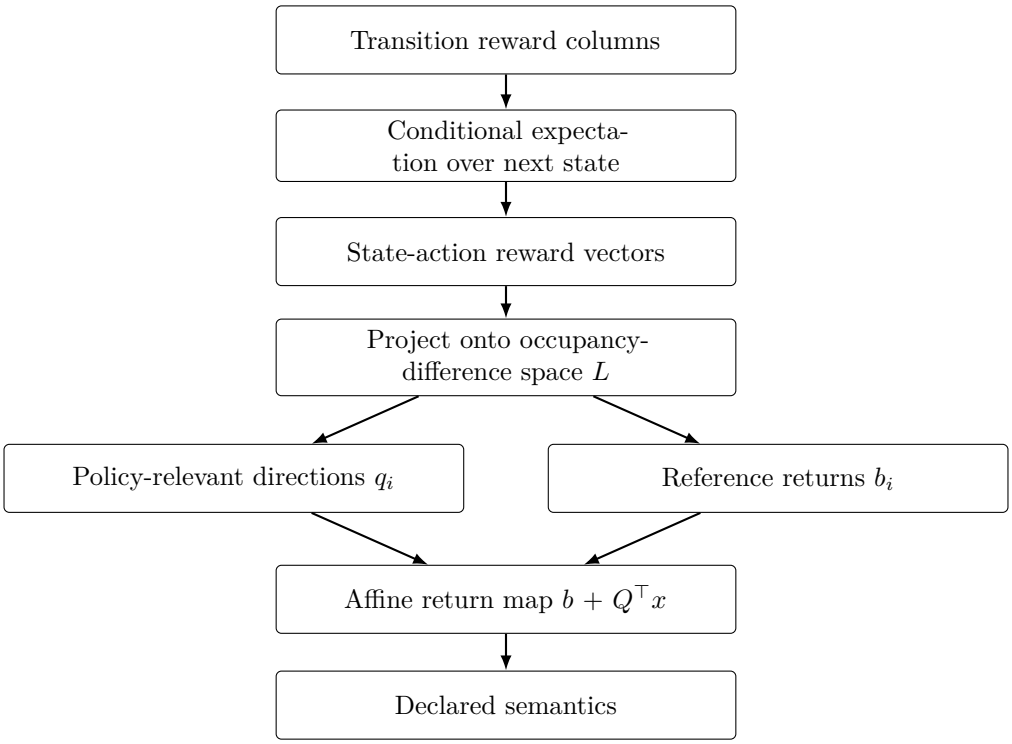

Figure 3: **Dynamics-aware reduction.** Projection removes reward components that are constant over feasible policy differences. Maximin still uses the reference levels b after standardization; Pareto differences cancel them.

4.4 A generic normalized-utility toolkit

The same elementary function-space construction underlies the cardinal maximin metric and clarifies where scalar-STARC-style completeness fails. Let X be compact and let $F : X \to \mathbb{R}$ be continuous. Define

$$\operatorname{rng}(F) = \max_X F - \min_X F \tag{11}$$

and

$$N(F) = \begin{cases} \dfrac{F - \min_X F}{\operatorname{rng}(F)}, & \operatorname{rng}(F) > 0, \\ 0, & \operatorname{rng}(F) = 0. \end{cases} \tag{12}$$

The zero convention identifies all constant utilities. The implemented Cardinal Maximin-STARC domain is narrower and rejects constants, but the generic construction exposes that design choice.

Definition 4.4 (Uniform normalized-utility distance)**.**

$$d_U(F, G) = \|N(F) - N(G)\|_\infty . \tag{13}$$

Theorem 4.5 (Pseudometric and positive-affine zero set)**.** *d_U is a pseudometric. For nonconstant F, G,*

$$d_U(F, G) = 0 \quad \Longleftrightarrow \quad G = aF + c \text{ on } X \text{ for some } a > 0,\ c \in \mathbb{R}. \tag{14}$$

Proof. The map $F \mapsto N(F)$ sends utilities into bounded real functions, so pulling back the sup metric gives nonnegativity, symmetry, and the triangle inequality. If $G = aF + c$ with $a > 0$, minima and ranges transform covariantly and $N(G) = N(F)$. Conversely, equality of normalized nonconstant utilities gives

$$\frac{F - \min F}{\operatorname{rng}(F)} = \frac{G - \min G}{\operatorname{rng}(G)}. \tag{15}$$

Rearranging,

$$G = \frac{\operatorname{rng}(G)}{\operatorname{rng}(F)} F + \min G - \frac{\operatorname{rng}(G)}{\operatorname{rng}(F)} \min F, \tag{16}$$

whose slope is positive. □

Definition 4.6 (Normalized reversal regret)**.**

$$\rho(F \mid G) = \sup_{x,y\,:\,G(y) \geq G(x)} [N(F)(x) - N(F)(y)]_+ . \tag{17}$$

Theorem 4.7 (Generic soundness)**.** *For all continuous utilities on compact X,*

$$\rho(F \mid G) \leq 2 d_U(F, G). \tag{18}$$

Proof. If $G(y) \geq G(x)$, normalization preserves the inequality unless G is constant, in which case both normalized values are zero. Therefore

$$\begin{aligned} N(F)(x) - N(F)(y) &= [N(F) - N(G)](x) + [N(G)(x) - N(G)(y)] \\ &\quad + [N(G) - N(F)](y) \\ &\leq 2 d_U(F, G). \end{aligned} \tag{19}$$

Taking positive parts and the supremum proves the result. □

Theorem 4.8 (Nonlinear same-order obstruction)**.** *On any utility class containing two nonconstant utilities with the same policy ordering but not related by a positive affine transformation, no positive constant c can satisfy*

$$\rho(F \mid G) \geq c\, d_U(F, G) \tag{20}$$

for every ordered pair in the class.

Proof. Choose such F, G. Theorem 4.5 gives $d_U(F, G) > 0$. Equal ordering means every G-preferred move is also F-preferred, so $\rho(F \mid G) = 0$. The proposed lower bound fails for every $c > 0$. □

The obstruction is not specific to one maximin construction. It applies to any unrestricted nonlinear utility class with a cardinal positive-affine quotient but an ordinal reversal target.

4.5 Generic pathwise safety

Theorem 4.9 (Robust derivative criterion)**.** *Let $\eta : [0, T] \to \Omega$ be absolutely continuous and let $\mathfrak{U}$ be a family of Lipschitz plausible true utilities. If, for every $U \in \mathfrak{U}$ and almost every t,*

$$\frac{d}{dt} U(\eta(t)) \geq 0, \tag{21}$$

then every $U \in \mathfrak{U}$ is nondecreasing along the entire path.

Proof. The composition of a Lipschitz function with an absolutely continuous path is absolutely continuous. Hence, for $0 \leq s < t \leq T$,

$$U(\eta(t)) - U(\eta(s)) = \int_s^t \frac{d}{du} U(\eta(u))\, du \geq 0. \tag{22}$$

□

This is the non-angular core of robust MORL stopping. Each semantics must still supply a computable lower bound on its directional derivative. The detailed selector and statistical stopping theory lives in the companion project; the distance results here supply inputs to those tests.

5 Why No Semantics-Free Vector STARC Exists

5.1 Three preference objects

For a fixed nonnegative weight w, the scalar utility is

$$U_{R,w}(x) = w^\top z_R(x) = w^\top b + (Qw)^\top x. \tag{23}$$

For a nonempty compact ambiguity set W, anchored maximin utility has the form

$$U_{R,W}(x) = \inf_{w \in W} w^\top z_R(x), \tag{24}$$

after any required objective standardization. Pareto weak dominance is

$$y \succeq_R^P x \quad \iff \quad z_R(y) - z_R(x) \in \mathbb{R}^k_+. \tag{25}$$

The first object is affine, the second is generally concave piecewise affine, and the third is a partial relation rather than a scalar utility.

5.2 Transformation audit

Four transformation classes matter.

Next-state redistribution. Any transformation preserving the conditional expected reward at each state-action pair preserves all three preference objects. It has already been removed before R is formed.

Policy-independent objective translations. For $z'_i(x) = z_i(x) + c_i$, fixed-weight policy differences and Pareto differences are unchanged. Maximin generally changes, because $\min_i(z_i(x) + c_i)$ depends on the relative offsets c_i.

Positive componentwise scaling. For $z'_i(x) = d_i z_i(x)$ with $d_i > 0$, Pareto dominance is unchanged. A fixed weight or maximin bottleneck generally changes unless weights or declared units co-transform.

One common positive affine utility transformation. For any scalar utility U, replacing it by $aU + c$ with $a > 0$ preserves its policy ordering.

5.3 Explicit translation reversal

Let two policies have returns

$$z(A) = (0, 100), \qquad z(B) = (1, 1). \tag{26}$$

Objective-wise maximin prefers B, since $1 > 0$. Add the policy-independent vector $(2, 0)$:

$$z'(A) = (2, 100), \qquad z'(B) = (3, 1). \tag{27}$$

Now maximin prefers A, since $2 > 1$. Pareto comparisons are unchanged because the same vector was added to every policy.

5.4 Explicit scale reversal

Let

$$z(A) = (1, 100), \qquad z(B) = (2, 3). \tag{28}$$

Maximin prefers B, since $2 > 1$. Multiplying the first objective by ten gives

$$z'(A) = (10, 100), \qquad z'(B) = (20, 3), \tag{29}$$

after which maximin prefers A, since $10 > 3$. Pareto dominance remains unchanged.

Corollary 5.1 (No exact universal quotient)**.** *There is no componentwise reward-equivalence relation that is simultaneously the exact invariance quotient for unrestricted Pareto and maximin semantics.*

Proof. Separate positive translations and scales belong to the Pareto invariance group, but the examples above show they do not belong to the maximin invariance group. An exact common quotient would have to both identify and distinguish the same pair. □

6 Fixed Weighted Sums and Family-STARC

6.1 Exact scalar reduction

For fixed w, the policy-dependent part of utility is $(Qw)^\top x$. The term $w^\top b$ is policy independent.

Theorem 6.1 (Exact scalar reduction)**.** *MORL reward comparison under one fixed linear scalarization is exactly scalar reward comparison for the effective projected direction Qw.*

Proof. For any two policies x, y,

$$U_{R,w}(y) - U_{R,w}(x) = (Qw)^\top (y - x). \tag{30}$$

Every ordering, margin, and regret statement therefore depends only on the effective scalar reward. Applying the scalar STARC canonicalization and distance recovers the scalar theory without a new vector construction. □

6.2 Unknown but fixed weights

If the protected semantics says that one unknown $w \in W$ will be fixed, a natural family comparison is

$$d_{\text{family}}(R, S) = \sup_{w \in W} d_{\text{STARC}}(Rw, Sw). \tag{31}$$

Uniform scalar STARC guarantees pass through the supremum. This answers how different the scalar policy ordering can be for any declared weight. It does not answer the maximin question $\inf_{w \in W} w^\top z_R(x)$, because normalizing each weight-specific scalar reward separately destroys relative levels across weights.

6.3 Zero family distance with changed maximin behavior

Take $W = \{e_1, e_2\}$ and positively rescale one objective in the scale-reversal example of section 5.4. Each individual scalar objective differs only by positive scale, so its scalar STARC distance is zero. The per-objective family distance is therefore zero. The minimum across objectives nevertheless changes and reverses the maximin preference. Family STARC and Maximin-STARC solve distinct problems.

6.4 The scalar STARC blueprint and its transplant boundary

The scalar source theory assumes finite state and action spaces, fixed dynamics, initial distribution and $0 < \gamma < 1$, removal of unreachable states, stationary randomized policies, scalar transition rewards, expected discounted return, and a finite-dimensional reward space.

Abstractly, scalar STARC chooses a linear canonicalization c whose fibers remove potential shaping and next-state redistribution, a norm n on $\mathrm{Im}(c)$, the standardization

$$s(R) = \begin{cases} c(R)/n(c(R)), & c(R) \neq 0, \\ 0, & c(R) = 0, \end{cases} \tag{32}$$

an admissible metric bilipschitz-equivalent to a norm, and finally the distance between standardized representatives.

The scalar zero set is policy-order equality, because two nontrivial linear functionals inducing the same order on a relatively open domain differ by a positive scale.

The soundness proof uses the chain

$$|J_1(\pi) - J_2(\pi)| \leq \frac{1}{1-\gamma} \|R_1 - R_2\|_\infty , \tag{33}$$

finite-dimensional norm equivalence, two endpoint errors for a proxy-preferred move, policy-independent value shifts under canonicalization, and scale control by the true policy-value range. This yields an inequality of the form

$$J_1(\pi_1) - J_1(\pi_2) \leq U \operatorname{range}(J_1)\, d(R_1, R_2) \tag{34}$$

whenever the proxy ranks π_2 at least as highly as π_1, for an environment- and norm-dependent constant U.

Scalar completeness is more specifically linear-geometric. Standardized rewards lie on a compact normalization boundary, metric distance controls their angle, the occupancy image contains a relative open ball, and in the plane spanned by two reward normals one can choose proxy-tied policies separated by the true reward. Trigonometry and norm equivalence then produce a lower regret witness.

That last argument does not transplant automatically. Maximin has a switching lower envelope instead of one normal, and same order need not imply positive-affine equality. Pareto has a partial order, no unique scalar regret, and no single normal or optimizer. The constructions in section 7 through section 12 replace those missing scalar ingredients rather than extending a vector norm.

For scalar Goodhart stopping, an angular uncertainty cap of radius θ around proxy direction r contains a harmful true reward for step d exactly when

$$\frac{r^\top d}{\|d\|_2} < \|r\|_2 \sin\theta. \tag{35}$$

MORL replaces this single cap test with lower support or margin tests over the declared semantic uncertainty set.

7 Pareto Policy Geometry

7.1 Primal and normal cones

Offsets cancel from return differences. A centered policy step $v = y - x \in L$ is weakly Pareto improving exactly when $q_i^\top v \geq 0$ for every objective i. Define the primal dominance cone

$$C_R = \{v \in L : q_i^\top v \geq 0 \ \forall i\} \tag{36}$$

and the objective normal cone

$$K_R = \operatorname{cone}\{q_1, \ldots, q_k\}. \tag{37}$$

For a cone $K \subseteq L$, define its dual by $K^* = \{v \in L : u^\top v \geq 0 \ \forall u \in K\}$.

Lemma 7.1 (Dual representation)**.** $C_R = K_R^*$.

Proof. Nonnegativity against each generator q_i implies nonnegativity against every nonnegative conic combination of generators. The converse holds because each generator belongs to K_R. □

7.2 Exact relation-equivalence theorem

Theorem 7.2 (Relation equivalence)**.** *For two reward families R and S on the same declared environment, the following are equivalent:*

1. *R and S induce the same weak Pareto relation on every policy pair;*
2. *$C_R = C_S$;*
3. *$K_R = K_S$.*

The same cone equality also fixes strict dominance once weak dominance and equality directions are interpreted consistently.

Proof. A policy pair (x, y) is weakly ordered by R exactly when $y - x \in C_R$. Equal primal cones therefore imply equal policy relations.

Conversely, suppose $C_R \neq C_S$. Choose v in their symmetric difference. By lemma 4.3, some positive multiple ϵv is a feasible occupancy difference. The corresponding policy pair is weakly improving under one reward family and not the other, contradicting relation equality. Relation equality therefore implies $C_R = C_S$.

Finitely generated cones are closed. By lemma 7.1 and the finite-dimensional bipolar theorem,

$$K_R = (K_R^*)^* = C_R^*, \tag{38}$$

and similarly for S. Hence $C_R = C_S$ if and only if $K_R = K_S$. □

7.3 Consequent invariances

The theorem proves invariance under policy-independent objective translations, which cancel in differences; positive rescaling of any objective, which preserves its ray; objective permutation; objective duplication; adding or deleting an objective already in the positive cone of the others; and scalar reward null transformations removed by occupancy projection. The last four transformations all preserve K_R. This cone quotient is the semantic object that a Pareto reward distance must respect.

7.4 Why unlabelled front distance is insufficient

Consider two policies with true and proxy returns

$$z_T(A) = (1, 1), \quad z_T(B) = (0, 0), \qquad z_P(A) = (0, 0), \quad z_P(B) = (1, 1). \tag{39}$$

Both unlabelled fronts are the singleton set $\{(1, 1)\}$, so their geometric front Hausdorff distance is zero. Proxy optimization nevertheless selects B, which is strictly dominated by A under truth. Policy identity, or an equivalent reward-semantic relation, cannot be discarded in a safety analysis.

7.5 Margin-free instability

Let $L = \mathbb{R}^2$, with true unit objective directions e_1, e_2, and inspect $v = e_1$. Its true margin is zero. Rotate the second objective to $q_2' = (-\sin\epsilon, \cos\epsilon)$. As $\epsilon \downarrow 0$, q_2' approaches e_2, but ${q_2'}^\top v = -\sin\epsilon < 0$. An arbitrarily small reward-direction perturbation therefore changes the weak dominance label of a boundary step. No nonzero uniform perturbation radius can preserve every Pareto classification without a positive margin assumption.

7.6 Earlier Pareto comparison objects

The current extreme-ray-body metric was reached after separating three valid but different Pareto comparison targets. Recording them prevents later work from conflating ordinal relation equality, numerical margin control, and a family of scalar preferences.

7.6.1 Dominance-cone distance

Let $B_L = \{v \in L : \|v\| \le 1\}$ and let $\operatorname{dist}(v, C)$ use the chosen norm. Define

$$d_C(R, S) = \sup_{v \in B_L} |\operatorname{dist}(v, C_R) - \operatorname{dist}(v, C_S)| . \tag{40}$$

Theorem 7.3 (Dominance-cone distance)**.** *d_C is a pseudometric on reward specifications, a metric on Pareto-equivalence classes, and*

$$d_C(R, S) = 0 \quad \Longleftrightarrow \quad R, S \text{ induce the same Pareto relation.} \tag{41}$$

Proof. This is the sup metric between two distance functions. If the cones agree, the functions agree. Conversely, equal distance functions have equal zero sets on B_L. Both zero sets are cones, so equality on the unit ball implies equality everywhere. Apply theorem 7.2. □

The quantity d_C gives the right ordinal zero set but does not by itself say whether a particular direction crosses the cone boundary.

7.6.2 All-generator margin distance

For each nonzero projected objective, define $\widehat{q}_i = q_i / \|q_i\|_*$, and, on the unit sphere S_L,

$$m_R^{\text{all}}(v) = \min_i \widehat{q}_i^{\top} v. \tag{42}$$

Define

$$d_m(R, S) = \sup_{v \in S_L} \left| m_R^{\text{all}}(v) - m_S^{\text{all}}(v) \right| \tag{43}$$

and

$$\rho_P(R \mid S) = \sup_{\substack{v \in S_L \\ m_S^{\text{all}}(v) \ge 0}} \left[-m_R^{\text{all}}(v) \right]_+ . \tag{44}$$

Theorem 7.4 (All-generator margin soundness)**.** *$\rho_P(R \mid S) \le d_m(R, S)$. More generally, if $m_S^{\text{all}}(v) \ge \delta$, then $m_R^{\text{all}}(v) \ge \delta - d_m(R, S)$.*

Proof. The definition of uniform distance gives $m_R^{\text{all}}(v) \ge m_S^{\text{all}}(v) - d_m(R, S)$ pointwise. Apply the proxy-margin assumptions and take a positive part or retain δ. □

If objectives are aligned by label, then $d_m(R, S) \le \max_i \|\widehat{q}_i - \widehat{q}'_i\|_*$. The absolute difference between minima of finite coordinate lists is at most their largest coordinatewise difference, and Hölder's inequality bounds each evaluation on the unit sphere.

This margin is sound, but a conically redundant generator can change its numerical value without changing Pareto behavior. The Pareto-STARC margin of section 8 minimizes only over unit extreme rays, which removes that representation artifact.

7.6.3 Preference-complete scalar comparison

For a declared compact $W \subseteq \Delta_k$, define

$$d_{PC}(R, S) = \sup_{w \in W} d_{\mathrm{STARC}}(Rw, Sw). \tag{45}$$

If scalar STARC soundness constants are uniform over W, taking the supremum gives a worst-preference scalar regret bound. Under uniform scalar completeness assumptions, a corresponding worst-weight lower witness also follows.

This object is stronger than bare Pareto equivalence. Pareto retains the positive hull of objective normals; preference-complete comparison retains each effective direction Qw up to scalar equivalence. Equal cones need not identify these directions weight by weight. Thus d_{PC} is a legitimate family metric, not a metric on bare Pareto relations.

7.6.4 Why Pareto-STARC is the primary object

The three historical objects answer different questions. The distance d_C has the exact ordinal cone zero set but lacks a direct finite margin representation in the implemented prototype. The distance d_m controls reversals but is sensitive to redundant generators. The distance d_{PC} controls a declared family of total scalar preferences and is finer than bare Pareto semantics.

Pareto-STARC retains the exact cone zero set while canonicalizing to extreme rays, and its Hausdorff distance is exactly a redundancy-invariant margin discrepancy. It therefore combines the two properties the frozen Pareto scope requires, without pretending to replace preference-complete analysis.

8 Pareto-STARC

8.1 Declared metric domain

For each nonconstant objective, define its canonical projected direction

$$c(r_i) = \frac{P_L r_i}{\|P_L r_i\|_2}. \tag{46}$$

Thus $q_i = P_L r_i$ and $c(r_i) = q_i / \|q_i\|_2$. Some source chapters denote the difference subspace by U and its projector by Π_U; this paper uses L and P_L throughout.

Policy-constant objectives with $P_L r_i = 0$ are discarded. Define

$$K_R = \mathrm{cone}\{c(r_i) : P_L r_i \neq 0\}. \tag{47}$$

The metric requires K_R to be nonzero and pointed: $K_R \cap (-K_R) = \{0\}$. Nonpointed cones, lineality-stratified comparison, and all-constant families fall outside the current metric domain rather than receiving an arbitrary distance.

8.2 Canonical extreme-ray body

Let $E(K_R)$ contain the unique unit representative of every extreme ray of the finite pointed cone. Define the compact convex body

$$B_R = \mathrm{conv}\, E(K_R). \tag{48}$$

This representation deletes duplicates and cone-redundant generators while retaining exactly the cone: $K_R = \mathrm{cone}\, B_R$.

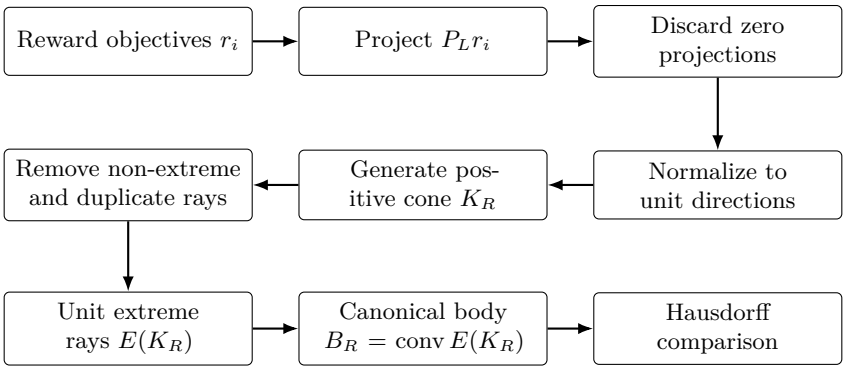

Figure 4: **Pareto-STARC canonicalization.** Each operation corresponds to a behavioral invariance: occupancy-null rewards, positive scale, duplication, and positive-cone redundancy are removed before distance is measured.

8.3 Definition

For nonempty compact sets $A, B \subseteq L$, Euclidean Hausdorff distance is

$$d_H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}\|a-b\|_2, \sup_{b\in B}\inf_{a\in A}\|a-b\|_2\right\}. \tag{49}$$

Definition 8.1 (Pareto-STARC)**.**

$$d_{\mathrm{PSTARC}}(R,S) = d_H(B_R, B_S). \tag{50}$$

8.4 Metric and zero set

Theorem 8.2 (Pareto-STARC zero set)**.** *d_{PSTARC} is a pseudometric on finite reward-family representations in its declared domain and a metric on Pareto-semantic equivalence classes. Moreover,*

$$d_{\mathrm{PSTARC}}(R,S) = 0 \iff K_R = K_S \iff \succeq_R^P = \succeq_S^P. \tag{51}$$

Proof. Hausdorff distance is a metric on nonempty compact sets, so nonnegativity, symmetry, and the triangle inequality transfer through the map $R \mapsto B_R$. Distance zero is equivalent to $B_R = B_S$. Equal bodies generate equal positive cones. Conversely, equal finite pointed cones have the same extreme rays and therefore the same unit extreme-ray body. Theorem 7.2 converts cone equality to Pareto policy-relation equality. □

8.5 Exact scalar reduction

Theorem 8.3 (Singleton scalar reduction)**.** *For singleton nonconstant families $R = (r)$ and $S = (s)$,*

$$d_{\mathrm{PSTARC}}(R,S) = \|c(r) - c(s)\|_2 = 2\sin\frac{\theta(r,s)}{2}, \tag{52}$$

where $\theta(r,s)$ is the angle between projected canonical reward directions.

Proof. Each canonical body is a singleton, so Hausdorff distance is the Euclidean distance between its points. For unit vectors u, v,

$$\|u-v\|_2^2 = 2 - 2u^\top v = 2 - 2\cos\theta = 4\sin^2(\theta/2). \tag{53}$$

Taking the nonnegative square root proves the result. □

This is an exact chord-distance reduction, not merely agreement of zero sets.

8.6 Margin representation

For a nonzero policy step $v \in L$, define

$$m_R(v) = \min_{q \in E(K_R)} q^\top \frac{v}{\|v\|_2}. \tag{54}$$

Lemma 8.4 (Sign semantics)**.** *$m_R(v) \geq 0$ if and only if v is weakly improving under every objective in R.*

Proof. If every extreme ray is nonnegative on v, every positive conic combination is nonnegative, and every original projected objective lies in that cone. Conversely, each extreme ray is represented by a generator direction in the finite cone up to positive scaling, so a negative extreme-ray evaluation violates at least one generating objective constraint. □

For a compact convex body B, write $h_B(u) = \sup_{b \in B} b^\top u$ for its support function. A linear minimum over a convex hull occurs at an extreme point, so

$$m_R(v) = -h_{B_R}\left(-\frac{v}{\|v\|_2}\right). \tag{55}$$

Theorem 8.5 (Hausdorff-support identity)**.**

$$d_{\mathrm{PSTARC}}(R, S) = \sup_{\|v\|_2 = 1} |m_R(v) - m_S(v)|. \tag{56}$$

Proof. The classical support-function identity for nonempty compact convex sets gives

$$d_H(B_R, B_S) = \sup_{\|u\|_2 = 1} |h_{B_R}(u) - h_{B_S}(u)|. \tag{57}$$

Substitute $u = -v$ and use $m_R(v) = -h_{B_R}(-v)$ together with the corresponding identity for S. □

This is the exact completeness statement for the declared normalized margin function. It is not a completeness theorem for every possible Pareto regret.

9 Pareto Selector Certificates and Reversal Loss

9.1 Pointwise selector certificate

Theorem 9.1 (Selector-margin certificate)**.** *For true reward family R, proxy family S, and any nonzero selector step v,*

$$m_R(v) \geq m_S(v) - d_{\mathrm{PSTARC}}(R, S). \tag{58}$$

Therefore $m_S(v) > d_{\mathrm{PSTARC}}(R, S)$ implies $m_R(v) > 0$.

Proof. Theorem 8.5 bounds the absolute margin discrepancy at every unit direction by the distance. Rearranging the lower side gives the first inequality. A proxy margin strictly larger than the distance leaves a positive true margin. Lemma 8.4 then implies strict improvement under every protected true objective. □

The distance alone does not classify a step. The same reward-family distance can accompany a robust interior step or a step arbitrarily close to a cone boundary. The operational object is the triple of reward distance, selector geometry, and protected semantics.

9.2 Directional worst-case reversal

For pointed true and proxy cones K and L, define the proxy dominance cone

$$D_L = \{v : q^\top v \geq 0 \ \forall q \in E(L)\}. \tag{59}$$

Define directional normalized reversal loss by

$$\mathcal{R}(K \leftarrow L) = \sup_{\substack{\|v\|_2 \leq 1 \\ m_L(v) \geq 0}} [-m_K(v)]_+ . \tag{60}$$

The arrow matters: the proxy L declares acceptable steps, while K evaluates their true harm.

Theorem 9.2 (Projection formula)**.**

$$\mathcal{R}(K \leftarrow L) = \max_{q \in E(K)} \|\Pi_{D_L}(-q)\|_2 . \tag{61}$$

Proof. Expand the negative true margin:

$$[-m_K(v)]_+ = \left[\max_{q \in E(K)} (-q)^\top v \right]_+ . \tag{62}$$

Because $E(K)$ is finite, the supremum over feasible v and the maximum over q may be interchanged. For a closed convex cone D, the support function of $D \cap \{v : \|v\|_2 \leq 1\}$ at z equals $\|\Pi_D(z)\|_2$. Apply this with $D = D_L$ and $z = -q$. The positive part is automatic because $v = 0$ is feasible. □

Theorem 9.3 (Reversal soundness)**.** $\mathcal{R}(K \leftarrow L) \leq d_{\mathrm{PSTARC}}(K, L)$.

Proof. If $m_L(v) \geq 0$, theorem 9.1 gives $m_K(v) \geq -d_{\mathrm{PSTARC}}(K, L)$. Hence $[-m_K(v)]_+ \leq d_{\mathrm{PSTARC}}(K, L)$. Take the supremum. □

9.3 Failure of one-sided completeness

Theorem 9.4 (One-sided incompleteness)**.** *There is no universal $c > 0$ such that, for every ordered pair of pointed cones,*

$$\mathcal{R}(K \leftarrow L) \geq c\, d_{\mathrm{PSTARC}}(K, L). \tag{63}$$

Proof. Work in $\mathbb{R}^2$ with

$$K = \mathrm{cone}\{e_1\}, \qquad L = \mathrm{cone}\{e_1, e_2\}. \tag{64}$$

Every L-improving step has both coordinates nonnegative, so it is also K-improving. Therefore $\mathcal{R}(K \leftarrow L) = 0$. The canonical bodies are $\{e_1\}$ and $\mathrm{conv}\{e_1, e_2\}$. Their Hausdorff distance is attained at e_2:

$$d_{\mathrm{PSTARC}}(K, L) = \|e_2 - e_1\|_2 = \sqrt{2}. \tag{65}$$

Thus $0 \geq c\sqrt{2}$ fails for every $c > 0$. □

The obstruction is order-theoretic. A proxy partial order can impose more simultaneous-improvement constraints than truth and thereby remove every unsafe step while still differing semantically. Scalar total-order completeness cannot be imported unchanged.

9.4 Policy-indexed Pareto-set coverage

Cone distance is not the only useful Pareto comparison. Suppose true and proxy standardized return maps satisfy

$$\delta_z = \sup_{x \in X} \|\widehat{z}_R(x) - \widehat{z}_S(x)\|_\infty . \tag{66}$$

Let P_R and P_S denote their policy-indexed nondominated sets.

Theorem 9.5 (Coverage certificate)**.** *For every $x \in P_R$ there exists $y \in P_S$ such that*

$$\widehat{z}_R(y) \geq \widehat{z}_R(x) - 2\delta_z \mathbf{1} \tag{67}$$

coordinatewise.

Proof. Starting from x, choose a proxy-nondominated y that proxy-dominates x. Existence follows by maximizing a strictly positive sum of proxy coordinates over the compact upper set. Then

$$\widehat{z}_R(y) \geq \widehat{z}_S(y) - \delta_z \mathbf{1} \geq \widehat{z}_S(x) - \delta_z \mathbf{1} \geq \widehat{z}_R(x) - 2\delta_z \mathbf{1}. \tag{68}$$

□

If an output set only ϵ-covers P_S under proxy returns, the same argument gives $(\epsilon + 2\delta_z)$ true coverage. The theorem evaluates the same policy identities under both rewards; an unlabelled front comparison cannot supply the guarantee.

10 Anchored Maximin Semantics

10.1 Why anchors and units are required

For objective i, declare an anchor α_i and positive unit σ_i. These encode the zero level and scale on which objectives are compared. Define $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ and standardized return

$$\widehat{z}_R(x) = D^{-1}(b + Q^\top x - \alpha). \tag{69}$$

Let W be a nonempty finite subset of the nonnegative simplex. For each $w \in W$, define the affine member

$$f_w(x) = w^\top \widehat{z}_R(x) = a_w + g_w^\top x, \tag{70}$$

where $a_w = w^\top D^{-1}(b - \alpha)$ and $g_w = QD^{-1}w$. The robust utility is the lower envelope

$$U_F(x) = \min_{w \in W} f_w(x). \tag{71}$$

It is finite, concave, continuous, and piecewise affine. It is Lipschitz on the compact feasible domain because

$$|U_F(y) - U_F(x)| \leq \max_{w \in W} \|g_w\|_* \, \|y - x\| . \tag{72}$$

Here $F = (R, W, \alpha, \sigma)$ denotes the full anchored specification, not only the set of affine member vectors. Comparisons F, G always assume the same declared environment and a common complete policy-mixture domain.

10.2 Covariance under recoding

Theorem 10.1 (Anchor-unit covariance)**.** *Suppose objective returns are recoded as $z'_i = c_i z_i + d_i$ with $c_i > 0$, while anchors and units co-transform as $\alpha'_i = c_i\alpha_i + d_i$ and $\sigma'_i = c_i\sigma_i$. Then standardized returns and anchored maximin utility are unchanged.*

Proof. Coordinatewise,

$$\frac{z'_i - \alpha'_i}{\sigma'_i} = \frac{c_i z_i + d_i - c_i\alpha_i - d_i}{c_i\sigma_i} = \frac{z_i - \alpha_i}{\sigma_i}. \tag{73}$$

Every weighted member is unchanged, hence so is their minimum. □

This is covariance of an explicitly standardized decision problem. Holding the anchors fixed while translating rewards changes the problem and may reverse preferences, as section 5.3 showed.

10.3 Policy-simplex representation

Let $z^1, \dots, z^m$ be the returns of a complete finite list of deterministic policy vertices. A randomized occupancy is represented by $\lambda \in \Delta_m$, and each affine member is represented by its values at those vertices, $f_j^\top \lambda$. The lower-envelope utility becomes

$$U_F(\lambda) = \min_j f_j^\top \lambda. \tag{74}$$

Completeness of the vertex list matters: omitting an occupancy-polytope vertex changes the domain and can invalidate exactness.

10.4 Activity is not a vertex-only property

For member j, define its strict-activity margin

$$\alpha_j = \max_{\lambda\in\Delta_m} \min_{\ell\neq j} (f_\ell - f_j)^\top \lambda. \tag{75}$$

If $\alpha_j > 0$, member j is uniquely minimal somewhere. A member can be inactive at every deterministic vertex but active at a randomized mixture. For example,

$$f_1 = (0, 1), \quad f_2 = (1, 0), \quad f_3 = (0.4, 0.4). \tag{76}$$

The member f_3 is not minimal at either endpoint, but it is uniquely minimal near the half-half mixture. Vertex-only pruning would incorrectly change the utility.

10.5 One global normalization

Define

$$u_F^{\min} = \min_{\lambda\in\Delta_m} U_F(\lambda), \qquad u_F^{\max} = \max_{\lambda\in\Delta_m} U_F(\lambda), \qquad r_F = u_F^{\max} - u_F^{\min}. \tag{77}$$

For nonconstant utility $r_F > 0$, normalize the entire envelope by

$$\overline{U}_F(\lambda) = \frac{U_F(\lambda) - u_F^{\min}}{r_F}. \tag{78}$$

Members must not be normalized independently, because their relative offsets and scales determine which member is the bottleneck.

Two envelopes may also agree at every deterministic vertex and differ inside the simplex. Let $F = \{(0,2,1),(2,0,1)\}$ and $G = \{(0,0,1)\}$. Both have deterministic-vertex utility $(0,0,1)$, but at $\lambda = (1/2, 1/2, 0)$,

$$U_F(\lambda) = 1, \qquad U_G(\lambda) = 0. \tag{79}$$

Interior randomized mixtures are therefore part of the metric domain.

10.6 Historical operational envelope equivalence

Before the name Cardinal Maximin-STARC was fixed, two anchored specifications were called operationally envelope-equivalent when their induced utilities satisfied $V(x) = aU(x) + c$ for all $x \in X$ and some $a > 0$. The generic maximin analogue was

$$d_{\mathrm{MM}}((R,W),(S,W')) = \left\| N(U_{R,W}) - N(U_{S,W'}) \right\|_\infty . \tag{80}$$

By theorem 4.5, this is a pseudometric whose nonconstant zero set is exactly operational envelope equivalence. By theorem 4.7, normalized one-sided reversal regret is at most $2d_{\mathrm{MM}}$. Section 11 specializes the construction to a finite complete deterministic-policy family, explicit anchors and units, exact active-region computation, and the name Cardinal Maximin-STARC.

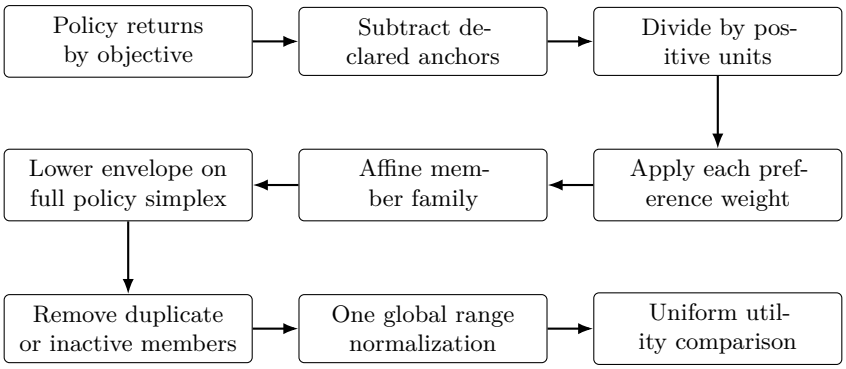

Figure 5: **Maximin-STARC construction.** Relative member levels are retained through envelope formation. Only one common positive affine utility degree of freedom is removed.

11 Cardinal Maximin-STARC

11.1 Definition

For two nonconstant normalized lower-envelope utilities on the same policy simplex, define

$$d_{\mathrm{MSTARC}}(F,G) = \sup_{\lambda \in \Delta_m} \left| \overline{U}_F(\lambda) - \overline{U}_G(\lambda) \right| . \tag{81}$$

The current software rejects constant maximin utilities rather than choosing a normative convention for comparing them.

11.2 Pseudometric and exact cardinal zero set

Theorem 11.1 (Cardinal zero set)**.** *d_{MSTARC} is a pseudometric on envelope representations, and*

$$d_{\mathrm{MSTARC}}(F,G) = 0 \iff \overline{U}_F = \overline{U}_G \text{ on every feasible policy mixture.} \tag{82}$$

Proof. The map $F \mapsto \overline{U}_F$ sends each representation to a bounded real function on Δ_m. Pull back the sup norm. Nonnegativity, symmetry, and the triangle inequality follow from the norm. Sup distance zero is exact pointwise equality. □

The distance is invariant to duplicate members, members that never affect the lower envelope, covariant anchor-unit recoding, and one common positive affine transformation of the final utility.

11.3 Exact cardinal completeness

If the operational error is defined as maximum normalized cardinal utility error,

$$\mathrm{Err}_{\mathrm{card}}(F, G) = \sup_{\lambda \in \Delta_m} \left| \overline{U}_F(\lambda) - \overline{U}_G(\lambda) \right|, \tag{83}$$

then by definition $\mathrm{Err}_{\mathrm{card}}(F, G) = d_{\mathrm{MSTARC}}(F, G)$. Soundness and completeness hold with constant one for this declared cardinal target. The identity must not be relabelled as ordinal-regret completeness.

11.4 Selector-margin bound

For a move from x to y, define $m_F(x, y) = \overline{U}_F(y) - \overline{U}_F(x)$.

Theorem 11.2 (Factor-two margin bound)**.**

$$m_F(x, y) \geq m_G(x, y) - 2d_{\mathrm{MSTARC}}(F, G). \tag{84}$$

Proof. Add and subtract proxy utility at each endpoint:

$$\begin{aligned} m_F(x, y) = m_G(x, y) &+ [\overline{U}_F(y) - \overline{U}_G(y)] \\ &- [\overline{U}_F(x) - \overline{U}_G(x)]. \end{aligned} \tag{85}$$

Each bracket has magnitude at most $d_{\mathrm{MSTARC}}(F, G)$, so the lower bound follows. □

Consequently, $m_G(x, y) > 2d_{\mathrm{MSTARC}}(F, G)$ certifies strict true maximin improvement.

Proposition 11.3 (Sharpness of the factor two)**.** *The constant two cannot be reduced uniformly.*

Proof. Use normalized singleton utilities on four policies with values

$$F = (0.75, 0.25, 0, 1), \qquad G = (0.5, 0.5, 0, 1). \tag{86}$$

Their sup distance is $d_{\mathrm{MSTARC}} = 0.25$. For the step from the first policy to the second, $m_G = 0$ and $m_F = -0.5$. Hence $m_F = m_G - 2d$, attaining the bound. □

11.5 Singleton reduction

Theorem 11.4 (Maximin singleton reduction)**.** *For singleton member vectors f, g,*

$$d_{\mathrm{MSTARC}}(\{f\}, \{g\}) = \|N(f) - N(g)\|_\infty, \tag{87}$$

where normalization subtracts the minimum deterministic-policy value and divides by its range.

Proof. The difference between the two normalized singleton utilities is affine in the mixture λ, so its absolute value is convex. A convex function on a finite simplex attains its maximum at a vertex. The supremum over all randomized mixtures therefore equals the largest absolute difference at a deterministic-policy vertex. □

This is a range-normalized L_∞ scalar STARC-style instance, but it is not numerically the same normalization as the projected Euclidean chord distance used by the Pareto-STARC singleton reduction.

11.6 Ordinal incompleteness

Theorem 11.5 (Ordinal incompleteness)**.** *There is no universal positive constant lower-bounding ordinal reversal regret by Cardinal Maximin-STARC distance on unrestricted lower-envelope utilities.*

Proof. On $t \in [0, 1]$, let

$$\overline{U}_F(t) = t, \qquad \overline{U}_G(t) = \min\{2t, \tfrac{1}{2} + \tfrac{1}{2}t\}. \tag{88}$$

Both functions are strictly increasing, so they induce the same policy order and every one-sided ordinal reversal loss is zero. At $t = 1/3$,

$$\overline{U}_F(1/3) = 1/3, \qquad \overline{U}_G(1/3) = 2/3, \tag{89}$$

so their distance is at least $1/3$. No inequality of the form $\rho_{\text{ord}} \geq c\, d_{\text{MSTARC}}$ can therefore hold for all such utilities with $c > 0$. □

The exact ordinal maximin goal remains open: construct or specialize a tractable metric whose zero set is equality of maximin weak orders, including ties, and prove an operational guarantee for that quotient.

12 Rejected Maximin Constructions and Thin-Facet Incompleteness

12.1 The envelope-member Hausdorff candidate

An earlier proposal globally normalized every member with the envelope's common utility range, removed inactive members, and compared the remaining affine member sets by Hausdorff distance under their sup norm on the policy domain. Call this distance d_{env}.

For two member families K, L, the induced normalized utilities satisfy

$$\left\|\overline{U}_K - \overline{U}_L\right\|_\infty \leq d_{\text{env}}(K, L). \tag{90}$$

Proof. Let $h = d_{\text{env}}(K, L)$. For each member $p \in K$, choose a member $q \in L$ with $\|p - q\|_\infty \leq h$. Then at every policy x, $q(x) \leq p(x) + h$. Taking an infimum over members yields $U_L(x) \leq U_K(x) + h$. Interchanging K, L gives the reverse inequality. □

The set distance is therefore a sound upper certificate. The question was whether it was also comparable from below to functional discrepancy. It is not.

12.2 The thin-active-facet family

Parameterize the two-policy simplex by $t \in [0,1]$. For endpoint vector $v = (v_0, v_1)$, let $\ell_v(t) = (1-t)v_0 + tv_1$. Fix $M > 1$ and $0 < \epsilon < 1$. Define the baseline and candidate member sets

$$B = \{g\}, \quad g = (0,1), \qquad K_{M,\epsilon} = \{g, h_{M,\epsilon}\}, \quad h_{M,\epsilon} = (M, 1-\epsilon). \tag{91}$$

Their affine functions are

$$\ell_g(t) = t, \qquad \ell_h(t) = M(1-t) + (1-\epsilon)t. \tag{92}$$

Lemma 12.1 (Crossing and activity)**.** *The two members cross at $t^* = M/(M+\epsilon)$. The member g is uniquely active before t^* and h is uniquely active after t^*.*

Proof. Their difference is

$$\ell_g(t) - \ell_h(t) = (M+\epsilon)t - M, \tag{93}$$

which is strictly increasing and has its unique zero at t^*. □

Both retained members are genuinely active for every positive ϵ. Their strict-activity margins are $\alpha_g = M$ and $\alpha_h = \epsilon$. The second facet is real but becomes arbitrarily thin as $\epsilon \downarrow 0$.

Lemma 12.2 (Global normalization)**.** *The baseline utility range is one. The candidate lower envelope has minimum zero and maximum t^*, so its globally normalized member endpoint vectors are*

$$\overline{g} = \left(0, 1 + \frac{\epsilon}{M}\right), \qquad \overline{h} = \left(M+\epsilon, \frac{(1-\epsilon)(M+\epsilon)}{M}\right). \tag{94}$$

Proof. The candidate envelope increases along g until the crossing and then follows h to $1-\epsilon$. Since $t^* > 1-\epsilon$, the maximum is t^*. Divide both endpoint vectors by $t^* = M/(M+\epsilon)$. □

Lemma 12.3 (Exact member-set distance)**.** *Under the endpoint sup norm,*

$$d_{\mathrm{env}}(K_{M,\epsilon}, B) = M + \epsilon. \tag{95}$$

Proof. The normalized baseline is the singleton $(0,1)$. The distance from $\overline{g}$ is ϵ/M. The first coordinate of $\overline{h}$ differs from the baseline by $M+\epsilon$, and this is the maximal endpoint difference. Taking the required directed maxima gives $M+\epsilon$. □

Lemma 12.4 (Exact functional discrepancy)**.** *Let*

$$d_{\mathrm{func}}(K, B) = \sup_{t \in [0,1]} \left| \overline{U}_K(t) - \overline{U}_B(t) \right|. \tag{96}$$

Then

$$d_{\mathrm{func}}(K_{M,\epsilon}, B) = \max\left\{ \frac{\epsilon}{M+\epsilon}, \frac{\epsilon(M-1+\epsilon)}{M} \right\}. \tag{97}$$

Proof. On $[0, t^*]$ the candidate follows g, and

$$\overline{U}_K(t) - \overline{U}_B(t) = \frac{t}{t^*} - t = \frac{\epsilon}{M} t. \tag{98}$$

This increases to $\epsilon/(M+\epsilon)$ at t^*. On $[t^*, 1]$ the difference is affine, with endpoint values

$$D(t^*) = \frac{\epsilon}{M+\epsilon}, \qquad D(1) = -\frac{\epsilon(M-1+\epsilon)}{M}. \tag{99}$$

The maximum absolute value of an affine function on an interval occurs at an endpoint, which proves the formula. □

Theorem 12.5 (No uniform Hausdorff completeness)**.** *There is no universal $c > 0$ such that*

$$c\, d_{\mathrm{env}}(K, L) \le d_{\mathrm{func}}(K, L) \tag{100}$$

for all globally normalized finite lower envelopes, even if every retained member is uniquely active somewhere.

Proof. Fix any $M > 1$ and let $\epsilon \downarrow 0$. Lemma 12.3 gives $d_{\mathrm{env}}(K_{M,\epsilon}, B) = M + \epsilon \to M > 0$, while lemma 12.4 gives $d_{\mathrm{func}}(K_{M,\epsilon}, B) \to 0$. If a positive universal c existed, the limit would imply $cM \le 0$, a contradiction. □

The divergence rate is

$$\frac{d_{\mathrm{env}}}{d_{\mathrm{func}}} \sim \frac{M^2}{\max\{1, M-1\}} \cdot \frac{1}{\epsilon}. \tag{101}$$

For $M = 50$ and $\epsilon = 0.001$, the recorded values are a strict activity margin of 0.001 for the thin member, an envelope-member Hausdorff distance of 50.001, a normalized functional discrepancy of 0.00098002, and a conservatism ratio of 51020.3873.

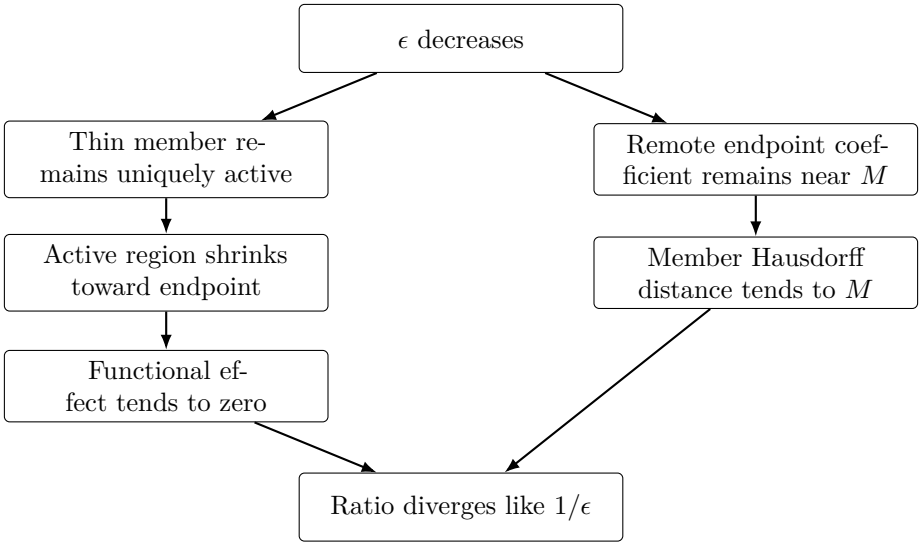

Figure 6: **Thin-facet failure mechanism.** Active-member reduction does not repair the candidate. The remote member is truly active for every positive ϵ, but only on a vanishing region.

12.3 Consequence for metric selection

The primary maximin metric is $d_{\mathrm{MSTARC}} = d_{\mathrm{func}}$. The member Hausdorff quantity may still be reported as a conservative geometric certificate when member matching has substantive meaning, but it must not be presented as uniformly equivalent to operational discrepancy.

12.4 Other investigated maximin candidates

Three candidate families should not be conflated. The induced normalized utility distance became Cardinal Maximin-STARC: a valid pseudometric with an exact cardinal zero set and a sound factor-two reversal bound, whose ordinal incompleteness is accepted and stated. Active-member Hausdorff distance remains a sound conservative upper certificate but was rejected as the primary metric after

theorem 12.5. Direct reversal discrepancy may be defined asymmetrically and symmetrized, but the resulting object is not automatically a pseudometric and may fail the triangle inequality; no metric theorem is claimed for it.

Family-STARC is not a fourth candidate for the same semantics. It addresses an unknown-but-fixed scalar preference rather than an adversarial minimum across preferences.

13 Counterexample Ledger

This section collects the minimal examples that every implementation and claim should survive. Most are proved above; the ledger exists so that none is lost when the surrounding exposition changes.

Objective translation reverses maximin. See section 5.3. Pareto is invariant; maximin is not.

Objective scaling reverses maximin. See section 5.4. Per-objective positive scaling is not a maximin invariance unless units co-transform.

Family-STARC zero does not imply equal maximin. See section 6.3. Independent scalar normalization removes exactly the cross-objective cardinal information maximin needs.

Identical fronts hide policy catastrophe. See section 7.4. A front-value metric that forgets policy identity can be zero while proxy optimization selects a truly dominated policy.

Same order does not fix cardinal gaps. For three policies, let $U(A) = 0$, $U(B) = 1/2$, $U(C) = 1$, and $V = h \circ U$ with $h(t) = t^{10}$. Then $V(A) = 0$, $V(B) = 1/1024$, $V(C) = 1$. The order and normalized endpoints agree, but the A-to-B utility gap differs by almost one half. Ordinal equivalence cannot determine cardinal regret for an unrestricted nonlinear utility class.

Pareto classification is unstable at zero margin. See section 7.5. Quantitative safety requires either a positive margin or an allowed ϵ-dominance error.

Weighted sums can miss deterministic Pareto points. Consider deterministic-policy returns $A = (1, 0)$, $B = (0, 1)$, $C = (0.4, 0.4)$. The point C is nondominated among these three deterministic policies. For every nonnegative normalized weight, at least one of A, B has scalarized value at least 0.5, while C has value 0.4. Thus C is never weighted-sum optimal. If randomized mixtures are allowed under SER, the mixture $(0.5, 0.5)$ is feasible and dominates C. Statements about weighted-sum completeness must therefore specify deterministic versus randomized policy domains.

Maximin ordinal incompleteness. See theorem 11.5. Equal strict ordering can coexist with positive Cardinal Maximin-STARC distance.

Active-member Hausdorff incompleteness. See theorem 12.5. The issue persists after exact active-member reduction.

Pareto one-sided regret incompleteness. See theorem 9.4. A stricter proxy partial order can eliminate all unsafe directions while remaining at positive Pareto-STARC distance.

Degenerate and out-of-domain cases. The theory does not hide these behind defaults. All projected objectives zero leaves the Pareto metric undefined under the current nonzero-cone construction. A cone with lineality is excluded from the pointed-cone metric. Numerically unresolved near-coincident rays raise an ambiguity error at the requested tolerance. A constant lower-envelope utility is rejected by the current Maximin-STARC implementation. An incomplete policy-vertex list voids the exact maximin claims. A missing true reward means the distance-based oracle certificate is not directly observable without an estimation theorem.

14 Algorithms and Implementation

14.1 Software organization

The package implementation lives under `src/morl_starc/`. The module `tabular.py` handles finite MDP validation, policy enumeration, occupancies, and reward evaluation. The module `geometry.py` handles occupancy-difference bases, reward projection, scalar chord STARC, and family projection. The module `pareto.py` handles Pareto fronts, cones, extreme rays, Pareto-STARC, margins, reversal, and coverage certificates. The module `optimization.py` handles finite simplex-region and lower-envelope optimization. The module `envelope.py` handles anchored envelope construction, activity, and normalization. The module `maximin_starc.py` handles Cardinal Maximin-STARC and certificates. The module `deep_sea_treasure.py` handles the canonical benchmark maps and route-choice reduction.

14.2 Occupancy solution

For a deterministic policy, the implementation solves the discounted flow linear system using scaled pivoting. It verifies nonnegativity, total occupancy mass, and flow residual postconditions. Accepted near-stochastic distributions are normalized before solving. Transition-level returns use stable summation.

Enumerating all deterministic policies is exponential in the number of states. This is acceptable for the declared small finite-tabular research setting and is not presented as a scalable control algorithm.

14.3 Occupancy-difference projection

Deterministic-policy occupancies generate the difference span. The code builds an orthonormal basis, projects each reward column, and normalizes nonzero projections. Projection removes any reward component orthogonal to all feasible policy differences, including the relevant scalar shaping and null directions.

14.4 Pareto cone processing

The implementation projects and normalizes objective directions, removes duplicate rays, tests each generator for positive-cone redundancy, rejects unresolved or nonpointed geometry, forms the unit extreme-ray bodies, computes exact finite convex-hull projection distances by active-set enumeration, and takes the two directed maxima for Hausdorff distance.

The procedure does not discretize the unit sphere. "Exact" refers to searching the finite active-set formulation rather than a grid. Floating-point linear algebra and tolerances still govern classification.

14.5 Lower-envelope processing

For maximin, the implementation evaluates complete deterministic-policy return vertices, standardizes with declared anchors and positive units, validates preference weights as nonnegative simplex elements, forms affine member values on the policy simplex, identifies exact active regions by finite constraint enumeration, computes the global lower-envelope minimum and maximum, globally normalizes the full envelope, and maximizes pairwise affine discrepancies on every compatible active-region intersection. This yields the undiscretized finite supremum distance. Active-region enumeration is exponential in the number of policy vertices and envelope members.

14.6 Numerical conditioning

The implementation distinguishes mathematical degeneracy from floating-point ambiguity. Safeguards include scaled pivoting in occupancy and active-set solves, global offset and scale removal before lower-envelope enumeration, scale-aware canonicalization thresholds, postcondition checks on flow, simplex feasibility, and objective values, explicit rejection of Pareto rays that cannot be resolved at the requested tolerance, retention of positive thin facets rather than tolerance-based silent deletion, and deterministic failure manifests with seeds and full case data.

An independent solver audit exposed a genuine defect when raw lower-envelope values were of order 10^{12}: the production active-set routine could choose the wrong basic feasible solution. Conditioning by a common affine scale fixed the problem, and the independent audit then passed. That failure history is part of the evidence.

15 Experimental and Numerical Evidence

15.1 Evidence hierarchy

The experiments are theorem-aligned falsification tests. They support internal consistency but do not prove the universally quantified theorems or establish external validity.

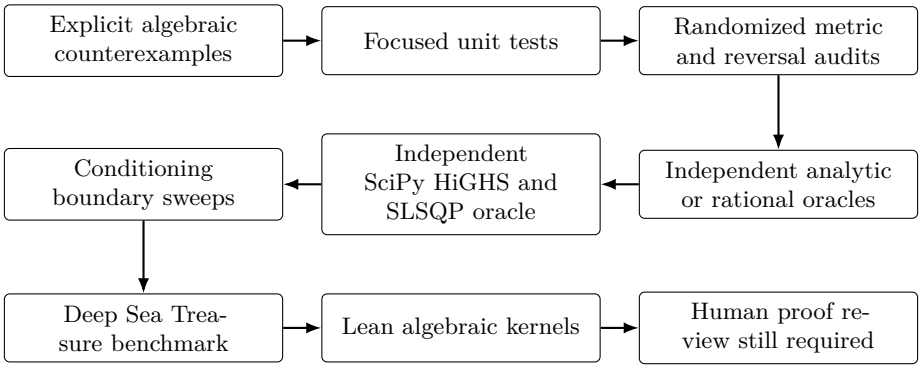

Figure 7: **Verification stack.** No layer substitutes for another. Lean currently begins after substantive premises, and a benchmark correlation does not verify a theorem.

15.2 Core numerics audit

The core numerics audit uses deterministic seed `20260901` and independently checks 5,000 simplex linear-optimization cases, 5,000 lower-envelope cases, 2,000 occupancy cases, and 500 scalar STARC projection cases. Recorded failure counts were zero for linear feasibility and value, envelope value, occupancy mass and flow, and scalar projection invariance.

The acceptance thresholds are numerical rather than symbolic. A linear or lower-envelope value error greater than 10^{-8} is a failure. An occupancy-mass error or flow residual greater than 10^{-7} is a failure. A transformed scalar-STARC discrepancy greater than 10^{-7} is a failure. Interval-oracle denominators or slopes of magnitude at most 10^{-14} are treated as numerically flat in the independent two-policy oracle.

15.3 Pareto-STARC metric audit

The Pareto-STARC metric audit uses deterministic seed `20260902` and runs four groups of 3,000 cases: random metric triples in dimensions two and three; same-cone transformations involving permutation, scaling, duplication, and redundant positive objectives; independent analytic two-dimensional line-segment Hausdorff comparisons; and singleton chord comparisons.

All metric, invariance, zero-set, margin, certificate, oracle, and singleton failure counts were zero. The largest independent two-dimensional and singleton errors were both 4.441×10^{-16}. Observed triangle and margin-bound excesses were nonpositive in every sampled case.

The audit's failure tolerances are 10^{-9} for sampled margin and invariance checks, 10^{-8} for the independent two-dimensional Hausdorff oracle, and 10^{-10} for singleton chord reduction. These thresholds define the pass/fail report; the much smaller observed maxima are reported separately.

15.4 Pareto reversal audit

The Pareto reversal audit uses seed `20260902` and compares the production active-set reversal solver with an independent analytic two-dimensional angular oracle over 5,000 random pointed-cone pairs. Oracle failures and soundness failures were both zero. The maximum oracle error was 5.204×10^{-16}. The nested-cone distance was 1.414214 and the nested-cone directional regret was 0.000000. The last pair numerically reproduces the exact incompleteness theorem rather than concealing it.

15.5 Cardinal Maximin-STARC audit

The Cardinal Maximin-STARC audit uses seed `20260902` and tests 3,000 random two-policy envelope triples. Its one-dimensional breakpoint oracle independently enumerates candidate locations, although it shares production normalization and utility evaluation. A separate exact-rational adversarial audit evaluated 5,000 newly generated two-policy envelope pairs.

Recorded results: metric failures, independent-oracle failures, affine-invariance failures, certificate failures, and singleton-reduction failures were all zero. The maximum triangle excess and the maximum margin-bound excess were both 0.000×10^{0}. The maximum independent-oracle error was 1.221×10^{-15} and the maximum exact-rational two-policy discrepancy was 1.78×10^{-15}. The attained selector-margin factor was 2.000000, the same-order positive cardinal distance was 0.333333, and the thin-facet Hausdorff-to-functional ratio was 51020.387. These include both favorable metric checks and adverse completeness witnesses.

15.6 High-dimensional independent solver oracle

The high-dimensional audit uses SciPy HiGHS and SLSQP rather than the production active-set solvers. With seed `20260903`, dimensions two through six, five categories, and four cases per dimension-category pair, it runs 100 composite cases. The categories are random instances, near-degenerate cones, thin maximin facets, duplicate policies, and high-discount cases through $\gamma = 1 - 10^{-12}$.

It independently checks simplex-region linear optimization, lower-envelope optima, active-member classification, normalized envelope sup distance, Pareto ray redundancy, and convex-hull distance. The committed full manifest records zero failures in every category. Maximum discrepancies were 1.318×10^{-15} for region LP relative error, 2.803×10^{-15} for lower-envelope relative error, 6.661×10^{-16} for normalized Maximin-STARC distance error, and 2.490×10^{-14} for Pareto-STARC distance error. The smoke artifact covers one case for each of the 25 dimension-category pairs and likewise records zero failures.

15.7 Conditioning-boundary audit

The conditioning-boundary audit tests reward scales from 10^{-12} through 10^{12}; discounts including 0, 0.9, 0.999, $1-10^{-12}$, and $1-10^{-15}$; active-facet margins down through 10^{-12}; and explicit ambiguous near-parallel cone cases. Projected-scaling, Pareto-scaling, maximin-scaling, thin-facet, occupancy, and ambiguity-handling failure counts were zero in the recorded run. Ambiguous geometry is expected to raise an error rather than produce a false zero distance.

The audit treats projected, Pareto, or maximin scale discrepancies above 10^{-10} and occupancy relative error above 10^{-12} as failures. It first confirms that the default tolerance rejects a deliberately unresolved narrow cone with an error containing `numerically ambiguous`; it then tightens `atol` to 10^{-15} and requires the same cone to have positive distance. This tests both conservative rejection and caller-controlled higher precision.

15.8 Thin-facet grid search

The thin-facet grid search evaluates the exact construction of section 12.2 on $M \in \{1, 2, 5, 10, 20, 50\}$ and $\epsilon \in \{0.5, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001\}$. All 54 deterministic cases satisfy the sound upper bound. The largest observed conservatism ratio is the $M = 50$, $\epsilon = 0.001$ case reported in section 12.2. The search illustrates a separately proved asymptotic counterexample.

15.9 Selector-specific Pareto benchmark

The companion project owns the selector-specific Pareto benchmark, which imports MORL-STARC through its public API. With seed `20260902`, it samples 10,000 pointed two-objective true/proxy cone pairs and proxy-Pareto-improving selector directions. Recorded outcomes were 7,222 true-safe cases, 2,778 true reversals, 2,722 certified-safe cases, zero certificate false positives, zero reversal-severity bound failures, and 32 of 34 distance bins containing both safe and reversal labels.

The mixed labels in 32 of 34 distance bins are a negative result for distance-only classification and positive evidence for the theorem's selector-margin interface. The benchmark is synthetic and theorem-aligned; it does not estimate real-world predictive performance.

15.10 Deep Sea Treasure benchmark

The package implements the deterministic 11×11 Deep Sea Treasure (DST) environment with cardinal actions, collision behavior, terminal treasure reward, time penalty, and published convex

and concave treasure maps. It enumerates one shortest collision-free path to each of ten treasures, with exact step counts $1, 3, 5, 7, 8, 9, 13, 14, 17, 19$.

For exhaustive metric comparison, the benchmark uses a labelled one-state, ten-action route-choice reduction preserving the discounted returns of those ten canonical paths. It does not claim to represent every looping or nonshortest grid policy.

The companion benchmark combines two maps, discounts 0.90, 0.97, and 0.99, four corruption families, and 25 cases per map-discount-corruption combination, giving $2 \times 3 \times 4 \times 25 = 600$ deterministic cases. Corruptions are positive objective rescaling and translation, positive objective mixing, smooth route-level misspecification, and targeted dominance reversal.

Observed outcomes were 405 dominance-reversal cases, zero Pareto-STARC certificate false positives, and a maximum reversal-severity bound excess of zero to printed precision. Every positive rescaling or translation case has zero Pareto-STARC distance and zero dominance reversal.

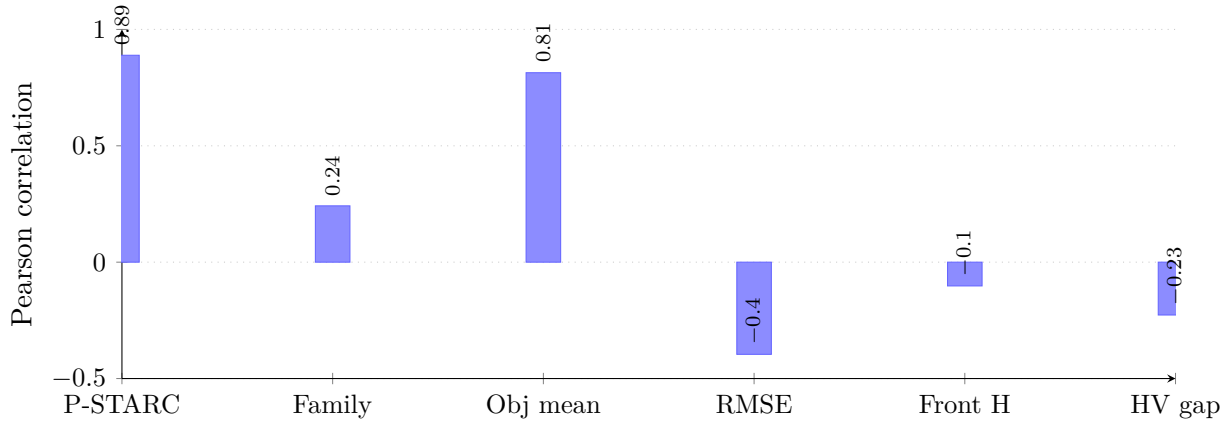

Figure 8: **Dominance-reversal target.** Pareto-STARC correlates 0.889, scalarized family-STARC 0.242, aligned per-objective STARC mean 0.814, return RMSE −0.396, normalized front Hausdorff −0.102, and hypervolume gap −0.227. Signs are empirical properties of this corruption design, not universal metric rankings.

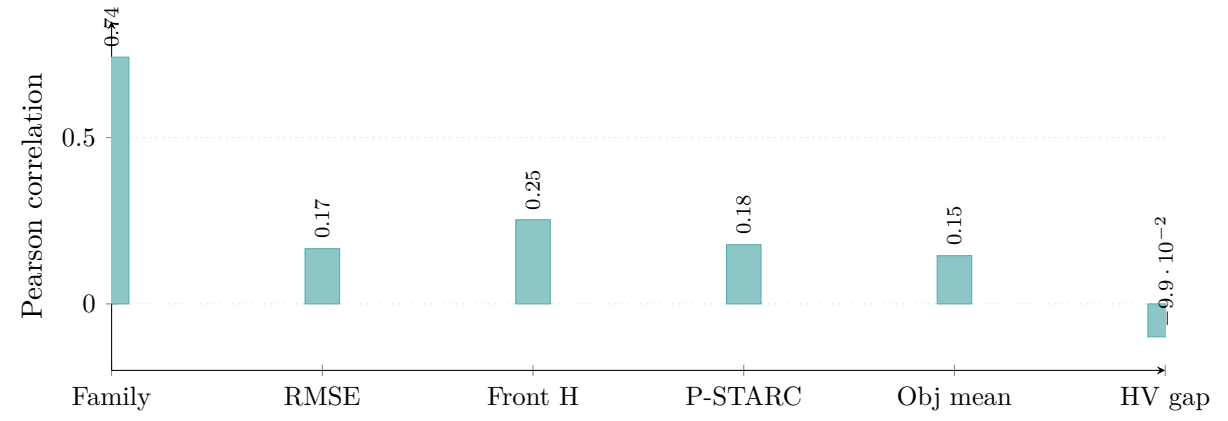

Figure 9: **Weighted-sum target.** Scalarized family-STARC correlates 0.742, return RMSE 0.166, normalized front Hausdorff 0.253, Pareto-STARC 0.178, aligned per-objective STARC mean 0.145, and hypervolume gap −0.099.

The two graphs show the intended semantic separation. Pareto-STARC best tracks the declared dominance-reversal target in this constructed benchmark, while family-STARC best tracks fixed-weight regret. Neither is universally superior independent of the protected loss. Correlation with the simple consecutive epsilon-front decline statistic is weak: 0.282 for Pareto-STARC and no more than 0.297 among the listed baselines. That adverse result is retained.

15.11 What the empirical evidence does not show

There are no learned reward models, continuous-control tasks, real preference datasets, cross-seed uncertainty intervals, held-out corruption families, out-of-distribution tests, or scaling curves. Cardinal Maximin-STARC has no standard-environment benchmark. The DST study has one deterministic design and synthetic corruptions that strongly structure labels. The reported 0.889 correlation is theorem-aligned validation, not evidence of broad practical superiority.

16 Formal Verification

16.1 Lean project

The Lean project under `formal/` is pinned by `lean-toolchain` and `lake-manifest.json`. The `check.sh` gate compiles with warnings treated as errors and audits for placeholders. The pinned toolchain is `leanprover/lean4:v4.34.0-rc2`; Mathlib revisions are fixed by the generated Lake manifest.

16.2 Pareto algebraic kernels

The Pareto kernel file `Formal/ParetoSTARC.lean` checks algebraic statements including Lipschitz control of a unit-direction support value by chord distance; the proxy-margin minus distance lower-bounding the true margin; strict certified margin implying safety; scalar reversal soundness after the margin premise is supplied; sign preservation under positive rescaling; sign preservation under adding a redundant positive objective; and the singleton chord-square identity.

The margin lower bound and strict-safety lemmas correspond to theorem 9.1 after the uniform margin premise of theorem 8.5 is supplied. The reversal lemma corresponds to the final scalar inequality used by theorem 9.3. The singleton identity supplies the algebraic last step of theorem 8.3. Cone construction and the support-function identity are not premises proved by this module.

16.3 Maximin algebraic kernels

The maximin kernel file `Formal/MaximinSTARC.lean` checks the endpoint-error argument yielding the factor-two margin bound; strict safety above $2d$; exact sharpness of the factor two; monotonicity of the same-order piecewise counterexample; and positive cardinal separation of that counterexample. These correspond to theorem 11.2, proposition 11.3, and theorem 11.5 after uniform endpoint-error assumptions are supplied. Sharpness is witnessed by the explicit finite-policy instance; the formal claim is that no smaller universal coefficient can replace two, not that every instance attains it.

16.4 What Lean does not verify

The formal project does not define and connect the full end-to-end objects: occupancy polytopes and flow equivalence; occupancy-difference projection; objective normal cones and their extreme

rays; Hausdorff distance between canonical bodies; the support-function identity in the repository's exact setting; active-region enumeration; global lower-envelope normalization; and correctness or completeness of the floating-point algorithms.

The phrase "Lean-checked algebraic kernels" is therefore accurate. "The complete Pareto-STARC or Maximin-STARC theorem is formally verified" is not.

The formal project also states that the Bretagnolle-Huber, Hoeffding, and Maurer-Pontil concentration inequalities used by adjacent confidence work are imported and source-audited rather than reproved. Lean checks downstream algebra only when those inequalities are supplied as hypotheses.

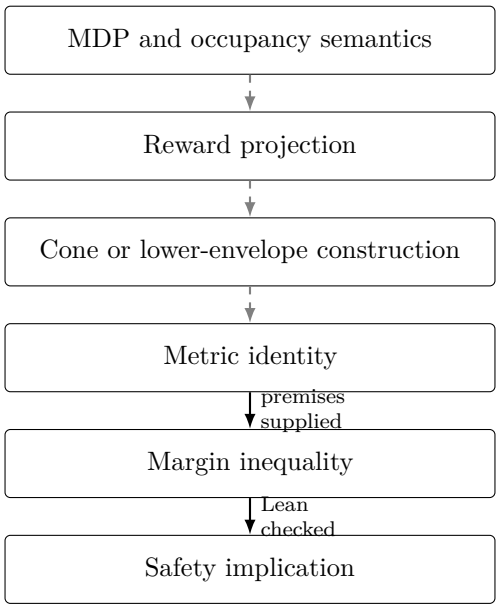

Figure 10: **Formalization boundary.** The checked kernels cover the final algebraic implications. The geometric and optimization pipeline before those premises remains a human-proof and software-validation obligation. Dashed edges are not yet mechanized.

17 Reading the Figures

The figures in this paper carry the conceptual and quantitative record directly in the source. The repository previously contained no committed empirical PNG, SVG, or PDF plot generated by the experiments; numerical outputs were recorded in Markdown and JSON. The diagrams here reproduce that record.

Figure 1 prevents the most common conceptual error: canonicalization must follow the semantic choice. The branches are not interchangeable algorithms for one target.

Figure 2 records the proof dependencies using the source theorem identifiers. It also separates the Pareto partial-order obstruction from the maximin thin-facet obstruction.

Figure 3 separates policy-relevant slopes Q from reference levels b. Pareto policy differences depend only on slopes; anchored maximin also depends on standardized levels.

Figure 4 explains every quotient in Pareto-STARC. The final body is representation invariant exactly because zero projections, scale, duplication, and conic redundancy have already been removed.

Figure 5 shows why maximin cannot normalize each objective or preference member independently. Lower-envelope formation precedes global utility normalization.

Figure 6 visualizes the mechanism behind a negative theorem: a member can remain uniquely active while its active region and functional effect vanish.

Figure 7 orders the verification layers without treating any one as decisive. Tests and solver oracles can discover counterexamples; they cannot prove universal claims. Lean can verify implications; it currently assumes the hard geometric premises.

Figure 8 compares metrics against dominance-reversal severity on 600 constructed DST cases. The high Pareto-STARC correlation is target-specific, and the negative correlations of some front and value baselines are reported rather than adjusted away.

Figure 9 changes the target to fixed weighted-sum regret. Family-STARC now has the highest correlation. This reversal of ranking is evidence for the semantic thesis rather than an inconvenience.

Figure 10 marks the exact formalization boundary. It prevents algebraic kernel verification from being overstated as an end-to-end mechanization.

No confidence intervals appear in the quantitative graphs. The DST design is a fixed deterministic 600-case snapshot, not a multi-seed sampling study. Adding conventional error bars would imply a sampling design that does not exist. A research-grade empirical extension must define the population or repeated-randomization scheme, run at least ten seeds, and report uncertainty intervals before inferential graph elements are added.

18 Limitations, Publication Status, and Open Problems

18.1 Mathematical limitations

Pareto-STARC currently excludes nonpointed cones and all-constant projected objective families. It is complete for its normalized margin function, not for every one-sided or symmetrized Pareto regret. Zero-margin Pareto classifications are necessarily perturbation-unstable. Cardinal Maximin-STARC has a cardinal utility zero set, not an ordinal weak-order zero set. Constant maximin utilities have no adopted comparison convention. Maximin anchors and units are normative inputs and may dominate conclusions. Active-member Hausdorff distance is only a conservative certificate. The theory is SER-specific and does not cover ESR trajectory-distribution semantics.

18.2 Computational limitations

Deterministic-policy enumeration is exponential. Cone-redundancy and convex-hull projection use combinatorial active-set searches. Lower-envelope activity and pairwise region discrepancy are exponential. Degeneracy decisions depend on floating-point tolerances, and "exact" does not mean exact arithmetic. No runtime or memory scaling study has been performed. The software is a small finite-tabular research prototype, not a scalable MORL library.

18.3 Statistical and operational limitations

Certificates compare proxy semantics against truth, which is unavailable in a real reward-learning problem. No end-to-end theorem converts observed preference or transition data into a high-probability Pareto-STARC or Maximin-STARC bound. Current confidence experiments in the companion project are often conservative rather than sharply calibrated. No learned proxy rewards or real preference data are evaluated. No held-out corruption, out-of-distribution, continuous-control, or cross-seed study is available. Maximin-STARC lacks a standard-environment benchmark.

18.4 Publication status

The finite-tabular mathematical core is coherent and strongly tested, but this snapshot is not a submission-ready empirical machine-learning paper. A narrow theory paper is plausible after independent human proof review. The required review covers cone equality versus policy-relation equality, including lower-dimensional occupancy geometry and degenerate objectives; occupancy and reachability quantifiers; the Hausdorff and support-function identity and norm conventions; lower-envelope normalization and active-region completeness; and prenex quantifiers and realizability for the impossibility theorems.

Numerical agreement and automated review are not substitutes for a signed human report. The complete submission gate is: an independent human reviewer names the examined commit and submits a signed written verdict for every commissioned proof surface; all blocking findings are resolved in the theorem statements, proofs, or declared domain; Python tests, numerical audits, the independent SciPy oracle, and Lean gates pass from a clean public checkout; every reported experiment is generated by a documented command and seed; priority language remains consistent with the novelty audit and literature survey; and a tagged reproducible release is created for any submitted version.

As of this snapshot, remote and local engineering gates have passed, but the signed human review and tagged submission release do not exist. The publication-readiness audit in the repository is a post-remediation self-audit dated 2026-09-03, not evidence of independent peer review.

18.5 Open problems

Open Problem 18.1 (Ordinal Maximin-STARC). For utility U, define its weak-preference graph $G_U = \{(\pi, \pi') : U(\pi) \geq U(\pi')\}$. A Hausdorff or relation distance between G_U and G_V has the desired ordinal zero set and is invariant to strictly increasing recodings. The open work is to specialize it to occupancy-induced piecewise-linear maximin utility, handle ties and duplicate policy coordinates, prove computability, and derive a meaningful operational bound. The relation-distance primitive itself is prior art.

Open Problem 18.2 (Nonpointed Pareto cones). Develop a lineality-stratified canonical representation that treats equality directions consistently and reduces to the current body on pointed cones.

Open Problem 18.3 (Regular maximin completeness). Find explicit active-facet thickness, coefficient, and transversality conditions under which member-set Hausdorff distance is lower-bounded by functional discrepancy. Section 12.2 proves that some regularity assumption is necessary.

Open Problem 18.4 (Scalable cone and envelope algorithms). Replace exhaustive active-set enumeration with certified optimization or approximation schemes, while exposing approximation error in the final certificate.

Open Problem 18.5 (Statistical estimation). Propagate reward, preference, and transition confidence regions through occupancy projection, cone canonicalization, and lower-envelope normalization. The target is a computable high-probability upper bound on semantic distance without observing truth.

Open Problem 18.6 (Selector-specific stopping). Identify structural conditions under which instantaneous support margins are monotone along an optimization path, and distinguish local safety, prefix safety, and globally optimal stopping.

Open Problem 18.7 (ESR and distributional semantics). Extend the framework from expected return vectors to trajectory-return distributions when nonlinear utility is applied before expectation.

Open Problem 18.8 (Broad empirical validation). Use multiple MOMDPs, learned proxies, at least ten seeds, uncertainty intervals, held-out corruption families, runtime scaling, and direct relation-, choice-, front-, and scalarization-distance baselines.

Open Problem 18.9 (General order-graph STARC). For any declared weak relation $\succeq$ on a compact policy domain, one may represent it by its graph $G_{\succeq} = \{(\pi, \pi') : \pi \succeq \pi'\}$. This common language includes scalar total preorders, maximin complete preorders, Pareto partial orders, lexicographic rules, and constraint-dependent relations. A graph Hausdorff or symmetric-difference metric gives a transparent zero set, but useful grading is not automatic. Open obligations include scalar STARC reduction or bilipschitz comparison, stability away from ties, unavoidable discontinuity at ties and cone degeneracy, computability, approximation error, and relation to selector regret. On a finite deterministic policy set, the exact pairwise comparison matrix is an immediate classical baseline; it ignores the ordering of randomized mixtures and should not be presented as a novel metric.

18.6 MORL phenomena absent from scalar STARC

The program isolates nine structural differences that any future generalization must state explicitly. Semantics is not unique: one vector reward supports incompatible Pareto, weighted-sum, maximin, lexicographic, and constrained orderings. Partial orders contain incomparability, so preserving only winners or front values is weaker than preserving every dominance comparison. Units and anchors may be normative, since separate positive scaling is a Pareto invariance but can change maximin behavior. Labels and alignment matter conditionally: permutation is harmless for bare Pareto semantics but changes a fixed labelled scalarization unless its weights are permuted too. Redundancy is semantic, since a positive-cone-redundant objective does not change Pareto dominance but may become a maximin bottleneck. Objective dimension can vary, and families with different objective counts can generate the same Pareto cone or induced policy relation. Preference uncertainty is an object, not just an experiment parameter. Outputs may be set-valued, so Pareto coverage must retain policy identity. Selector dependence is irreducible: reward distance alone generally does not decide safety, because the selected step, path, margin, and target loss also matter.

18.7 Recommended research sequence

Audit every proof against degenerate objectives, lower-dimensional occupancy polytopes, and explicit quantifier order. Keep every new counterexample as an exact executable regression test. Complete the independent human review of the current Pareto and maximin claims. Develop nonpointed Pareto and ordinal maximin objects before claiming a universal semantic framework. Add certified approximation and runtime analysis before increasing policy or objective dimension. Connect confidence regions to estimable semantic-distance certificates. Only then run learned-proxy, multi-environment, multi-seed empirical studies with held-out corruptions and uncertainty intervals.

19 Reproduction Protocol

Commands run from the MORL-STARC repository root unless a command changes directory explicitly.

19.1 Installation

Python 3.11 or newer is required.

```
python3 -m pip install -e .
```

The core runtime has no third-party dependency. Install the independent oracle extra with

```
python3 -m pip install -e '.[oracle]'
```

19.2 Unit tests

```
PYTHONPATH=src python3 -W error -m unittest discover -s tests -v
```

19.3 Numerical audits

```
PYTHONPATH=src python3 -W error -m experiments.audit_core_numerics
PYTHONPATH=src python3 -W error -m experiments.audit_pareto_starc_metric
PYTHONPATH=src python3 -W error -m experiments.audit_pareto_starc_regret
PYTHONPATH=src python3 -W error -m experiments.audit_maximin_starc_metric
PYTHONPATH=src python3 -W error -m experiments.audit_conditioning_boundaries
PYTHONPATH=src python3 -W error -m experiments.search_envelope_tightness
python3 -W error -m experiments.audit_high_dimensional_solver_oracle \
  --cases-per-category 4
```

19.4 Deep Sea Treasure benchmark

These commands require the companion repository to be present as a sibling directory. The benchmark code is owned by that project and imports the installed or source-tree MORL-STARC API.

```
cd ../MORL-Goodharts-Law
PYTHONPATH=src:../MORL-STARC/src \
  python3 -m experiments.benchmark_deep_sea_treasure
```

The selector-specific 10,000-case benchmark is

```
cd ../MORL-Goodharts-Law
PYTHONPATH=src:../MORL-STARC/src \
  python3 -m experiments.benchmark_pareto_starc_goodhart
```

19.5 Formal gate

```
cd formal
./check.sh
```

19.6 Expected artifact fields

The high-dimensional audit writes a JSON manifest containing the master seed, dimensions, categories, cases per category, failure counters, maximum errors, and complete replay data for every failure. A clean audit has an empty `failures` list. An empty list is evidence about that run, not proof that no counterexample exists.

20 Source Map

The derivations consolidated here come from the project's theory directory. The main sources are the common foundations, the maximin reward comparison, the Pareto reward comparison, the theorem dependency and open-problem notes, the Pareto implementation notes, the thin-facet incompleteness chapter, the authoritative Pareto-STARC definition, the Pareto reversal soundness and incompleteness chapter, the authoritative Cardinal Maximin-STARC definition, and the original STARC-to-MORL program statement.

The result and audit records are the maximin envelope results, the envelope completeness findings, the Pareto-STARC metric results, the novelty audit, the Deep Sea Treasure and Pareto-regret results, the Cardinal Maximin-STARC results, the adversarial publication-readiness audit, the literature survey, the independent human proof-review brief, and the narrow manuscript scope.

Bibliographic details and claim-specific quotations should be checked against the project reading list and literature survey before submission. Where the repository has not frozen a complete camera-ready bibliography, this paper uses conservative descriptions.

Conclusion

MORL-STARC is not one distance obtained by putting a norm around a vector reward. It is a semantic research program. Fixed weights recover scalar STARC. Pareto dominance is represented by a projected objective cone, leading to a canonical extreme-ray Hausdorff metric with exact margin semantics and a sound but one-sided reversal guarantee. Anchored maximin is represented by a standardized lower-envelope utility, leading to an exact cardinal function metric and a sharp factor-two selector certificate. The failed completeness claims are not peripheral: thin maximin facets and nested Pareto cones identify the precise boundaries at which scalar intuition breaks.

The mathematical and computational record is strong for the declared finite tabular domain. The responsible conclusion is narrower than a publication headline. The project supplies rigorous constructions, proofs, explicit counterexamples, reproducible numerical falsification, and checked algebraic kernels, while scalability, ordinal maximin comparison, nonpointed Pareto geometry, statistical estimability, end-to-end formalization, and broad empirical validity remain open.

References

Geon-Hyeong Byeon et al. Multi-objective reinforcement learning with max-min criterion: A game-theoretic approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations (ICLR)*, 2021.

Alfredo Iusem and Alberto Seeger. Distances between closed convex cones: Old and new results. *Journal of Convex Analysis*, 17, 2010.

Christian Klamler. A distance measure for choice functions. *Social Choice and Welfare*, 30, 2008. doi: 10.1007/s00355-007-0239-y.

Vicki Knoblauch. Two preference metrics provide settings for the study of properties of binary relations. *Theory and Decision*, 79, 2015. doi: 10.1007/s11238-015-9487-y.

Hiroki Nishimura and Efe A. Ok. A class of dissimilarity semimetrics for preference relations. *Mathematics of Operations Research*, 49(4), 2024. doi: 10.1287/moor.2022.0351.

Giseung Park and Youngchul Sung. Reward dimension reduction for scalable multi-objective reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Giseung Park et al. The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. In *International Conference on Machine Learning (ICML)*, 2024.

Shuang Qiu et al. Traversing pareto optimal policies: Provably efficient multi-objective reinforcement learning. *arXiv preprint arXiv:2407.17466*, 2024.

Manel Rodriguez-Soto, Juan A. Rodriguez-Aguilar, and Maite Lopez-Sanchez. An analytical study of utility functions in multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations (ICLR)*, 2024.

Blake Wulfe et al. Dynamics-aware comparison of learned reward functions. *arXiv preprint*, 2022.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.