Stratified Geometric Ethics: Mathematical Foundations for Verifiable Moral Reasoning in Autonomous Systems

Andrew H. Bond

San José State University

andrew.bond@sjsu.edu

Abstract

We present *Stratified Geometric Ethics* (SGE), a mathematical framework providing rigorous foundations for deterministic, verifiable ethical reasoning in autonomous systems. SGE addresses fundamental limitations of prior geometric approaches by modeling the space of ethically relevant configurations as a *stratified space*—a union of smooth manifolds of varying dimensions connected along boundary strata—rather than a smooth manifold. This structure captures moral discontinuities, incommensurable values, and genuine ethical dilemmas that smooth models cannot represent.

We make five principal contributions. First, we show that stratified spaces are a *natural minimal candidate* among standard geometric structures for representing ethical phenomena including threshold effects, lexical priorities, and moral dilemmas (Theorem 2.3). Second, we establish a *representation theorem* (Theorem 4.3) characterizing all satisfaction functionals satisfying five explicit axioms—including a novel locality axiom and scale-normalization assumption—with a complete proof. Third, we prove *finite approximation theorems* (Theorems 3.9–3.11) showing that any decision problem on a compact stratified space reduces to a finite graph problem with explicit error bounds. Fourth, we establish *decidability results* (Theorem 6.4) for the quantifier-free, non-temporal fragment of our ethical specification language via o-minimal structures, with temporal properties handled by standard model checking on finite approximations. Fifth, we derive *sample complexity bounds* (Theorems 7.1–7.3) for learning ethical content from data.

We introduce the Bond Invariance Principle (BIP), which requires ethical judgments to depend only on morally relevant relationships, not arbitrary representation choices, and prove that our axiom system implies BIP.

SGE serves as the theoretical foundation for the DEME 2.0 architecture presented in the companion paper [2], providing mathematical justification for design choices including multi-dimensional moral vector spaces, governance profiles with veto regions and lexical priorities, and layered enforcement architectures.

Keywords: AI ethics, differential geometry, stratified spaces, formal verification, autonomous systems, moral philosophy

1 Problem statement and approach

The deployment of AI systems in safety-critical domains—healthcare, autonomous vehicles, criminal justice, financial markets—creates an urgent need for frameworks that make ethical reasoning explicit, deterministic, and formally verifiable. Current approaches fall into two categories: *principle-based* approaches that articulate high-level commitments without computational implementations, and *learning-based* approaches that infer ethical behavior from feedback without formal guarantees. Neither satisfies the requirements of safety-critical deployment where lives depend on ethical decisions made at machine timescales.

Recent work has proposed differential-geometric frameworks for ethics, modeling configurations as points on smooth manifolds and ethical quantities as tensor fields [1]. While mathematically elegant, these approaches face five fundamental limitations:

The Content Problem. Existing frameworks assume obligations and interests are "given" without specifying how these tensors derive from ethical sources—constitutional principles, stakeholder preferences, or deliberative procedures.

The Smoothness Problem. Real ethical reasoning involves discrete choices (trolley problems), incommensurable values (life vs. property), and threshold effects (killing vs. letting die). Smooth manifolds cannot represent these discontinuities.

The Linearity Problem. Bilinear satisfaction functionals $S(x) = O^{\mu}(x)I_{\mu}(x)$ cannot capture threshold effects, lexical priorities, or diminishing returns—phenomena central to moral reasoning.

The Tractability Problem. No existing geometric ethics framework provides complexity analysis establishing whether ethical computation can be performed in realtime.

The Verification Problem. No framework enables machine-checkable proofs of ethical constraint satisfaction.

This paper introduces *Stratified Geometric Ethics* (SGE), a framework that addresses all five limitations while preserving the coordinate-invariance that makes geometric approaches valuable. The companion paper [2] instantiates SGE in the DEME 2.0 architecture for real-time ethical governance in autonomous systems.

This paper argues that a central failure mode in machine ethics is not merely bias in data, but **representation dependence**: conclusions that shift when only the description changes. We formalize a bond-based invariance principle and derive the constrained class of admissible evaluators it entails.

1.1 Relationship to DEME 2.0

SGE was developed through a process of reflective equilibrium with DEME 2.0, a computational architecture for real-time ethical governance described in the companion paper [2]. DEME commits to concrete engineering choices: multi-dimensional moral vector spaces, governance profiles with veto regions and lexical priorities, a three-layer architecture (strategic, tactical, reflex), and hardware-resident ethics modules. SGE provides the mathematical justification for these choices:

- DEME's *moral vector space* $M \subseteq \mathbb{R}^k$ is justified by Theorem 2.3, which shows stratified spaces are natural minimal candidates for ethical phenomena.
- DEME's *governance profiles* with veto regions and scalarization are justified by Theorem 4.3's representation of satisfaction operators.

• DEME's *real-time enforcement* is justified by the finite approximation theorems (3.9–3.11) and complexity bounds (Section 5).

• DEME's *verification layer* is justified by the decidability results (Theorem 6.4) combined with standard model checking.

2 Why Stratified Spaces

Before developing the machinery of SGE, we establish that stratified spaces are a *natural minimal candidate* among standard geometric structures for capturing ethical phenomena. We do not claim a fully formal minimality theorem over all conceivable geometric structures—such a claim would require defining "simpler" in a way that excludes exotic alternatives. Instead, we show that standard alternatives (smooth manifolds, manifolds with corners, cell complexes) each fail to represent at least one essential ethical phenomenon, and that stratified spaces succeed at all of them.

2.1 Ethical Phenomena Requiring Geometric Representation

We identify four categories of ethical phenomena that any adequate geometric framework must represent:

E1. Discrete Choices. Many ethical decisions present finitely many options with no meaningful intermediate. In trolley-type scenarios, one must choose path A or path B; there is no "path 0.5A + 0.5B."

E2. Incommensurable Values. Some moral dimensions cannot be traded off at any finite rate. The value of a human life is not equivalent to any amount of property damage; procedural rights cannot be "compensated" by increased welfare.

E3. Threshold Effects. Crossing certain boundaries has discontinuous moral significance. The distinction between killing and letting die, between lying and remaining silent, between 0.99 and 1.00 on a consent scale—these involve discrete moral transitions.

E4. Genuine Dilemmas. Some situations admit no ethically satisfactory resolution. Both available options involve genuine moral loss ("moral residue"), and this property must be represented, not smoothed away.

2.2 Insufficiency of Smooth Manifolds

Proposition 2.1. *Let M be a smooth connected manifold. Then:*

(i) Any two points in M can be connected by a smooth path (violates E1);

(ii) Any smooth function $f : M \to \mathbb{R}$ has continuous level sets (violates E3);

(iii) For any smooth Riemannian metric g on M, all directions at any point have finite, comparable costs (violates E2).

Proof. (i) follows from path-connectedness of connected manifolds. (ii) follows from continuity of smooth functions. (iii) follows from positive-definiteness of Riemannian metrics: $g(v,v) > 0$ for all nonzero v, so all directions have positive finite cost, hence are comparable. ■

2.3 Stratified Spaces as Natural Minimal Candidates

Definition 2.2 (Stratified Space)**.** *A stratified space is a triple $(M, \{M_i\}_{i \in I}, \preceq)$ where M is a paracompact Hausdorff space, $\{M_i\}$ is a locally finite partition into connected smooth*

manifolds (strata), and $\preceq$ is a partial order on I such that $M_i \cap cl(M_j) \neq \emptyset$ implies $i \preceq j$ (frontier condition). We require Whitney's condition (B) for regularity.

Theorem 2.3 (Stratified Spaces Represent E1–E4)**.** *Stratified spaces can represent all four ethical phenomena:*

(i) Discrete choices are modeled by 0-dimensional strata representing decision endpoints;

(ii) Incommensurable values are modeled by stratified metrics with arbitrarily large cost (in a singular limit);

(iii) Threshold effects are modeled by stratum boundaries where satisfaction functions have discontinuities;

(iv) Genuine dilemmas are modeled by singular strata from which all exits incur positive moral cost.

Moreover, among standard geometric structures—smooth manifolds, manifolds with corners, and cell complexes without smooth structure—each fails at least one of E1–E4. Stratified spaces therefore emerge as a natural minimal candidate.

Proof. The positive claims (i)–(iv) follow by explicit construction in Section 2.4. For the comparative claims: smooth manifolds fail by Proposition 2.1. Manifolds with corners can represent E1 and E3 (via boundary strata) but not E2 (all boundary components have finite, comparable codimension) or E4 (corners are not "singular" in the required sense; they have well-defined tangent cones with finite dimension). Cell complexes can represent E1 and E4 but lack the smooth structure on cells needed for tensor calculus (obligations as vector fields, interests as covector fields). Stratified spaces combine smooth strata (enabling differential geometry) with singular boundaries (enabling discontinuities), achieving representation of all four phenomena. ■

Remark 2.4. We do not claim that stratified spaces are minimal among *all conceivable* geometric structures. A determined geometer could propose disconnected manifolds, manifolds with degenerate metrics, subanalytic sets, or other exotic alternatives. Our claim is that stratified spaces are the natural minimal choice among *standard* structures commonly used in differential geometry and topology.

2.4 Constructive Examples

We provide explicit constructions demonstrating how stratified spaces represent each phenomenon. In these examples, we distinguish between *deliberation strata* (interior strata that parameterize reasoning or intermediate configurations) and *decision strata* (0-dimensional terminal strata). An admissible decision trajectory may traverse higher dimensional strata during deliberation but must terminate on a 0-dimensional decision stratum; remaining indefinitely in a deliberation stratum is not considered a completed decision.

Example 2.5 (Trolley Problem—Discrete Choice)**.**

Let $M = \{A, B\} \cup \gamma$, where A and B are 0-dimensional strata (the discrete choices) and $\gamma \cong (0,1)$ is a 1-dimensional stratum (the deliberation space). Topologize M so that $\gamma \cong (0,1)$ and its closure satisfies $\bar{\gamma} = \gamma \cup \{A, B\}$, identifying A with 0 and B with 1.The frontier condition gives $A, B \preceq \gamma$. Any path in M from deliberation to decision must terminate at a discrete point—there is no continuous interpolation staying within the terminal decision strata between A and B.

Example 2.6 (Lexical Priority—Incommensurability)**.**

Let $M = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ *with coordinates* (v_1, v_2) *representing two values. Stratify as* $M_0 = \{(0,0)\}$, $M_1 = \{0\} \times \mathbb{R}_{>0}$, $M_2 = \mathbb{R}_{>0} \times \{0\}$, *and* $M_3 = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. *Define a family of metrics on* M_3 *by* $g_\varepsilon = diag(1/\varepsilon^2, 1)$. *As* $\epsilon \to 0$, *motion in the* v_1 *direction becomes unboundedly costly relative to* v_2, *yielding a lexical limit in which any decrease in* v_1 *dominates any increase in* v_2. *This corresponds to a governance profile where* v_1 *is lexically prior to* v_2. *Strictly, the* $\epsilon \to 0$ *limit yields a singular (degenerate) geometry; nevertheless, for sufficiently small* ϵ, *any bounded-cost path behaves lexicographically, since arbitrarily small decreases in* v_1 *dominate any bounded improvement in* v_2.

Example 2.7 (Consent Threshold—Discontinuity)**.**

Let $M = [0,1]$ *with strata* $M_0 = \{\tau\}$ *(the threshold),* $M_1 = [0, \tau)$, *and* $M_2 = (\tau, 1]$. *Set* $S(\tau) = 0$ *(or* $S(\tau) = 1$, *depending on policy), so the discontinuity is located at the stratum boundary,*

$$S(x) = \begin{cases} 0, & x < \tau, \\ S(\tau), & x = \tau, \\ 1, & x > \tau. \end{cases}$$

The stratum boundary at τ*represents the discrete moral transition from "insufficient consent" to "adequate consent." The stratification isolates the morally exceptional boundary point where the rule regime changes.*

Example 2.8 (Sophie's Choice—Genuine Dilemma)**.**

Let $M = D \cup \{A, B\}$, *where* $D \cong (0,1)$ *is a 1-dimensional dilemma stratum with* $\bar{D} = D \cup \{A, B\}$, *and* A, B*are 0-dimensional terminal strata. Define satisfaction so that any path from* D *to either endpoint incurs irreducible moral residue (e.g.,* $S(A) = S(B) = \alpha < 1$*while* $sup_{\,x \in D}\, S(x) \leq \beta < \alpha$*). Then the dilemma stratum is singular in the sense that no trajectory can reach a fully satisfactory terminal state; all exits encode unavoidable moral loss.*

3 Stratified Moral Spaces

We now develop the formal machinery of stratified moral spaces, providing the geometric foundation for ethical reasoning.

3.1 Formal Definitions

Definition 3.1 (Stratified Moral Space)**.**

A stratified moral space is a stratified space $(M, \{M_i\}, \preceq)$ *where:*

(i) M is the total moral space of ethically relevant configurations;

(ii) Each stratum M_i *is a smooth manifold representing configurations admitting smooth ethical trade-offs;*

(iii) Stratum boundaries represent moral discontinuities—dilemmas, phase transitions, or categorical distinctions.

Definition 3.2 (Moral Dimension)**.**

The moral dimension of $x \in M$ *is* $\dim(x) := \dim(M_i)$ *where* $x \in M_i$*. Points on lower-dimensional strata represent morally "singular" configurations where some trade-offs break down.*

Definition 3.3 (Stratified Tangent Bundle)**.**

At $x \in M_i$*, the stratified tangent space is* $T^{str}_x M := T_x M_i \oplus N_x(M_i, M)$ *where* $T_x M_i$ *is the tangent space to the stratum and* N_x *is the normal cone.*

Definition 3.4 (Stratified Ethical Selection)**.**
Given a stratified moral space $(M, \{M_i\}, \preceq)$ *and a feasible action set* $A \subseteq M$*, an action* $a^* \in A$ *is ethically admissible if:*

1. a^* *lies in a minimal stratum* M_k *such that* $A \cap M_k \neq \emptyset$*;*
2. a^* *is maximal with respect to a smooth ethical preference functional* ϕ_k *defined on* M_k*.*

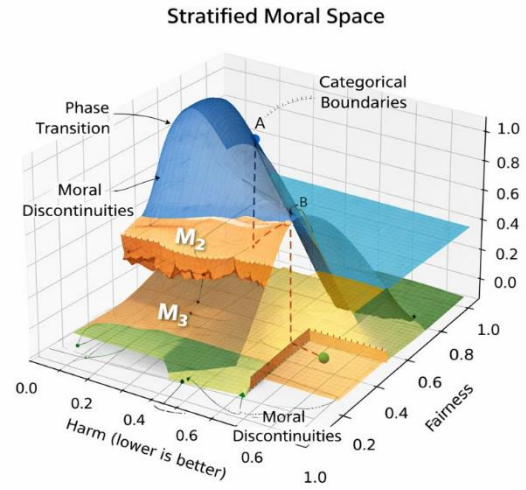

Figure 1 | Stratified moral space with moral discontinuities

A schematic 3D stratified space $(M, \{M_i\}, \preceq)$ depicting ethically relevant configurations. The horizontal axes parameterize two illustrative ethical dimensions—**Harm** (lower is better) and **Fairness**—while the vertical axis denotes **Ethical satisfaction** (a scalar proxy for overall moral value). The full volume represents the **total moral space** M. Three representative strata M_1, M_2, M_3 (semi-transparent surfaces) illustrate regions in which ethical trade-offs vary smoothly: within any single stratum, nearby configurations admit continuous, "manifold-like" adjustments (e.g., small changes in harm can be compensated by changes in fairness without qualitative shifts in moral kind). **Stratum boundaries** (annotated as **Moral discontinuities**) mark loci where smooth trade-offs fail—corresponding to dilemmas, sharp regime changes, or categorical distinctions; these are indicated by ridge/edge features and highlighted with dashed guides and arrows. **Phase transition** denotes a boundary across which the governing ethical regime changes, while **Categorical boundaries** indicate constraints that separate morally distinct classes of actions or outcomes. Example configurations **A** and **B** (marked points) lie on different strata; the vertical dashed projections indicate their coordinates in (Harm, Fairness) and emphasize that moving between strata requires crossing a discontinuity rather than a smooth deformation within a single M_i.

3.2 Whitney's Condition (B) and Its Role

Definition 3.5 (Whitney's Condition B)**.** *Let* $M_i \subset cl(M_j)$*. The pair satisfies Whitney (B) at* $x \in M_i$ *if: for sequences* $\{y_n\} \subset M_j$*,* $\{x_n\} \subset M_i$ *with* $y_n, x_n \to x$*, if secant lines* $\ell_n = x_n y_n \to \ell$ *and* $T_{y\_n} M_j \to \tau$*, then* $\ell \subset \tau$*.*

Lemma 3.6 (Consequences of Whitney B)**.** *If M satisfies Whitney (B), then: (i) paths approaching stratum boundaries have well-defined limiting tangent directions; (ii) the stratification is locally topologically trivial along each stratum; (iii) geodesic limits exist and are well-behaved at boundaries.*

Remark 3.7 (Verifying Whitney B in Practice)**.** For computational implementations, moral spaces are typically constructed as semialgebraic sets (defined by polynomial inequalities) or as finite unions of manifolds with explicit boundary conditions. By Łojasiewicz's theorem, semialgebraic sets admit Whitney stratifications, and the stratification can be computed algorithmically. For moral spaces defined by threshold conditions (e.g., "harm > τ"), the level sets of polynomial functions automatically satisfy Whitney (B).

3.3 Finite-Dimensional Approximation Theorems

For computational implementation, continuous stratified spaces must be approximated by finite structures. The following theorems establish that this can be done with explicit error bounds while preserving stratification structure.

Definition 3.8 (Stratified Graph)**.** *A stratified graph is $G = (V, E, \sigma, \preceq)$ where (V, E) is a finite directed graph, $\sigma: V \to \Lambda$ assigns vertices to stratum labels, and $\preceq$ is a partial order on Λ such that $(v, w) \in E$ implies $\sigma(v) \preceq \sigma(w)$.*

Theorem 3.9 (Existence of ε-Approximations)**.** *Let M be a stratified space with finitely many compact strata. For every $\varepsilon > 0$, there exists a finite stratified graph G_ε that is an ε-approximation of M: (i) vertices map to their labeled strata; (ii) every point of M is within ε of some vertex; (iii) edges connect nearby vertices respecting stratum order.*

Theorem 3.10 (Decision Approximation Bound)**.** *Let (M, A, u, Γ) be a decision problem where u is L-Lipschitz. If π^* is optimal on M and π^* is optimal on G_ε, then $|u(x, \pi^*(x)) - u(x, \pi^*(v_x))| \leq 2L\varepsilon + \omega(\varepsilon)$ where ω is the modulus of continuity of the optimal value function.*

Theorem 3.11 (Structure Preservation)**.** *For sufficiently small ε, any ε-approximation of a Whitney (B) space preserves stratification structure: paths between strata correspond to directed graph paths, and constraint sets constant on strata are exactly preserved.*

Corollary 3.12 (DEME Implementability)**.** *DEME's finite moral vector spaces and discrete governance profiles constitute ε-approximations of continuous stratified moral spaces, with approximation error bounded by Theorem 3.10.*

Proof sketches. Theorem 3.9 follows from compactness: cover each stratum by finitely many ε-balls, take centers as vertices, and connect vertices whose balls overlap. Theorem 3.10 applies Lipschitz continuity: the factor of 2 accounts for comparing two approximated values. Theorem 3.11 uses Whitney (B): local topological triviality ensures that for small ε, the graph structure mirrors the stratification, and constraint predicates constant on strata are exactly captured by vertex labels. ■

4 The Representation Theorem

The representation theorem characterizes all satisfaction operators consistent with explicit axioms. Unlike prior work that merely assumed a particular functional form, we *derive* the form from first principles, with all assumptions made explicit.

4.1 Five Axioms for Satisfaction Operators

Definition 4.1 (Satisfaction Operator)**.** *A stratified satisfaction operator is a map $\Sigma: \Gamma(T^{str}M) \times \Gamma(T^{*,str}M) \times \Gamma(Sym^2T^{*,str}M) \times C \to C^\infty(M)$ taking obligation field O, interest field I, metric g, and constraint set C to a satisfaction function $S: M \to \mathbb{R}$.*

We impose five axioms:

Axiom 1 (Coordinate Invariance)**.** *For any diffeomorphism $\psi: M \to M$, $\Sigma(\psi_* O, \psi^* I, \psi^* g, \psi(C))(x) = \Sigma(O, I, g, C)(\psi^{-1}(x))$.*

Axiom 2 (Normalized Monotonicity)**.** *At fixed $\|O\|_g$, increasing $I_\mu O^\mu$ (the alignment of obligations with interests) increases Σ.*

Axiom 2' (Normalized Response-Curve Invariance).
On the regular region U(as in Theorem 4.3), define the normalized alignment $z(x) := I(O)/\| O \|_g$and write

$$\Sigma(O, I, g, C)(x) = \chi_C(x) + G(x; z(x), c(x)), c(x) := \| I(x) \|^2_{g^{-1}}.$$

We assume there exists a (stratum-wise smooth) monotone function $f: \mathbb{R} \to \mathbb{R}$with $f(0) = 0$such that for every admissible $(O, I, g, C)$and every $x \in U$,

- either $G(x; 1, c(x)) = 0$and then $G(x; z, c(x)) = 0$for all z,
- or else

$$\frac{G(x; z, c(x))}{G(x; 1, c(x))} = f(z) \text{for all } z \in \mathbb{R}.$$

Equivalently, for all $x \in U$there exists a positive scale $\lambda(x) := G(x; 1, c(x))$such that

$$G(x; z, c(x)) = \lambda(x)\, f(z) \text{for all } z.$$

Axiom 3 (Constraint Respect)**.** *If $x \in C$ (constraint set), then $\Sigma(O, I, g, C)(x) = -\infty$ (excluded region).*

Axiom 4 (Stratum Compatibility)**.** *Σ restricts to a smooth function on each stratum of M.*

Axiom 5 (Locality)**.** *$\Sigma(O, I, g, C)(x)$ depends only on the pointwise values O(x), I(x), g(x), and the indicator $1_C(x)$, not on derivatives, jets, or values at other points.*

Remark 4.2 (Motivation for Locality)**.** Axiom 5 captures the intuition that satisfaction is a *local scoring operator* with no hidden history or look-ahead. An agent's ethical standing at configuration x should depend on the ethical facts *at x*, not on curvature tensors, gradients, or how the agent arrived at x. This matches the operational semantics of DEME, where Ethics Modules evaluate moral vectors pointwise.

4.2 Scale Normalization

In addition to the five axioms, we impose an explicit *modeling assumption* to avoid arbitrary sensitivity to units or scaling conventions:

Assumption (Scale Normalization). *Reparametrizing the obligation field by a positive scalar, $O \mapsto \alpha O$ for $\alpha > 0$, should not affect the ranking of configurations, aside from a context-dependent scaling $\lambda(x)$ that may depend on x but not on O.*

This assumption is a modeling choice, not a logical consequence of the axioms. It reflects the principle that the *direction* of obligations relative to interests matters, but arbitrary unit choices (e.g., measuring harm in "utils" vs. "millutils") should not change ethical conclusions.

4.2.1 The Bond Invariance Principle

The axioms of Section 4.1, together with the scale normalization assumption, encode a fundamental methodological commitment. We now make this commitment explicit as a named principle that captures the core requirement for trustworthy ethical reasoning.

Definition 4.2.1 (Bond).

A *bond* is a morally relevant relationship between entities. Formally, a bond is a tuple b = (a, p, r, c) where:

- a is an agent or role (the bearer of obligation, risk, or responsibility)
- p is a patient or counterparty (the one owed, protected, or affected)
- r ∈ R is a relation type from a fixed ontology R (e.g., owes, bears-risk-for, has-authority-over, has-consented-to, is-responsible-for, has-claim-against)
- c is a context qualifier (scope, conditionality, or domain restriction)

Definition 4.2.2 (Bond Structure).

The *bond structure* of an ethical configuration T, denoted B(T), is the set of all bonds encoded in T:

$$B(T) = \{b = (a, p, r, c) : b \text{ is a morally relevant relationship in } T\}$$

$B(T)$ may be treated as a typed directed multigraph (or finite relation) over entities in T.

The bond structure captures: who bears risk for whom, who owes what to whom, who has consented to what, who is responsible for what, role relations (physician–patient, employer–employee, guardian–ward), rights and claims, and temporal commitments.

Worked Example (Bond Structure in a triage decision).
Let T describe an ER triage situation with entities:
{Clinician (D), Patient (P), Hospital (H)}, and a context c={triage_slot=7, time=t, jurisdiction=CA} Define relation-types r∈{duty_of_care, non_discrimination, informed_consent} Then a possible bond structure is:

B(T)={(D,P,duty_of_care,c), (H,P,non_discrimination,c), (P,H,informed_consent,c)}.

Intuitively: the clinician owes care to the patient in this triage context; the hospital owes a non-discrimination constraint to the patient; and the patient's consent relation constrains what interventions are permissible. Under a bond-preserving transformation (e.g., renaming $P$to P', rescaling clinical measurements, or changing an equivalent data representation), $B(T)$is unchanged—so any ethical verdict that changes *without* a change in these bonds is diagnosing a representational

artifact rather than a real moral difference. Let T'be $T$with P renamed to P′ (or units converted). Then $B(T') = B(T)$, so $\mathcal{J}(T') = \mathcal{J}(T)$.

Definition 4.2.3 (Bond-Preserving Transformation).

A transformation g acting on ethical configurations is *bond-preserving* if it leaves the bond structure invariant:

$$g \in G \iff B(g \cdot T) = B(T)$$

Bond-preserving transformations include:

- Renaming or relabeling agents (without changing their relationships)
- Reparameterizing features (changing units, scales, or encodings)
- Reordering presentation (changing the sequence in which options appear)
- Equivalent redescriptions (syntactically different but semantically identical formulations)
- Coordinate changes on the moral manifold that preserve the underlying relational structure

Transformations that are *not* bond-preserving include:

- Changing who bears a risk
- Adding or removing consent
- Altering role relationships
- Shifting responsibility between agents
- Breaking or creating commitments

Definition 4.2.4 (Bond Invariance Principle). An ethical judgment function $\mathcal{J}$: T → V (where V is a space of verdicts, rankings, or scores) satisfies the Bond Invariance Principle (BIP) if:

$$\boxed{\forall g \in G: \mathcal{J}(T) = \mathcal{J}(g \cdot T)}$$

where G is the group of bond-preserving transformations.
In words: If the bonds are unchanged, the judgment must be unchanged.

Proposition 4.2.5 (Accountability Form of BIP).

An equivalent formulation: if $\mathcal{J}$(T) ≠ $\mathcal{J}$(T'), then at least one of the following holds:

1. B(T) ≠ B(T') — a bond was added, removed, or modified; or
2. The normative lens was explicitly changed (different metric g, different aggregation rule, different constraint set C, or different perspective/weighting).

In words: If your judgment changes, you must show what bond changed—or declare that you changed the rules.

Proposition 4.2.6 (Axioms Imply BIP).

A satisfaction operator Σ satisfying Axioms 1–5 and Scale Normalization satisfies the Bond Invariance Principle.
Proof sketch.

- Axiom 1 (Coordinate Invariance) ensures Σ is invariant under diffeomorphisms of M. Bond-preserving coordinate changes are a subset of such diffeomorphisms.
- Axiom 5 (Locality) ensures Σ(x) depends only on pointwise values O(x), I(x), g(x), and 1_C(x)—not on labels, presentation order, or encoding conventions.

- Axiom 3 (Constraint Respect) ensures the forbidden/permitted boundary is determined by the constraint set C, which is defined by bond structure (consent status, rights violations, etc.), not by arbitrary features.
- Scale Normalization ensures that unit conventions (a form of bond-preserving reparameterization) do not affect rankings.

Together, these axioms guarantee that Σ responds only to the morally relevant relational structure B(T), satisfying BIP. ■

Remark 4.2.7 (BIP as a Diagnostic).

The Bond Invariance Principle provides a test for ethical reasoning systems:

Observation	**Diagnosis**
Judgment varies under relabeling	BIP violation: system responds to names, not relationships
Judgment depends on presentation order	BIP violation: system responds to syntax, not structure
Judgment changes under unit conversion	BIP violation: system responds to encoding, not content
Equivalent descriptions yield different verdicts	BIP violation: system lacks semantic invariance

Any such violation indicates the system is responding to *representation* rather than *moral reality*.

Remark 4.2.8 (BIP and Fairness).

Many failures in algorithmic fairness are BIP violations. If a recidivism score varies with race encoding (but race doesn't change the underlying bonds of offense, risk, and rehabilitation), BIP is violated. If a hiring algorithm responds to name-gender correlations (but the candidate's qualifications and role-fit bonds are unchanged), BIP is violated. The Bond Invariance Principle thus provides a principled foundation for representation-invariance requirements in fair machine learning.

Remark 4.2.9 (What BIP Does Not Constrain).

BIP constrains *how* judgments depend on structure, not *what* structure to encode or *which* normative lens to adopt. Different ethical theories correspond to different choices of:

- Metric g (utilitarian vs. egalitarian vs. lexicographic)
- Constraint set C (what is absolutely forbidden)
- Aggregation/contraction (how multi-agent perspectives combine)

BIP permits pluralism at the level of normative lens—it requires only that, *given* a lens, the judgment be invariant under bond-preserving transformations.

The Bond Invariance Principle in Summary:

Bonds, not labels.
Structure, not syntax.

Relationships, not representations.

This principle is the conceptual core of the axiom system. The representation theorem (Theorem 4.3) characterizes the *unique functional form* that satisfies BIP together with the monotonicity and smoothness requirements.

4.3 Complete Proof of the Representation Theorem

Theorem 4.3 (Representation of Stratified Satisfaction)**.** *Let Σsatisfy Axioms 1, 3–5, Axiom 2 (Normalized Monotonicity), Axiom 2′ (Normalized Response-Curve Invariance), and the Scale Normalization assumption. Then there exist:*

- *a smooth monotone function $f: \mathbb{R} \to \mathbb{R}$ (activation function);*
- *a smooth function $\lambda: M \to \mathbb{R}_+$ (scale field);*
- *a constraint indicator $\chi_C: M \to \{0, -\infty\}$;*

such that on the regular region $U := \{x \in M: x \notin C, \|O(x)\|_g \neq 0$ and g_x is nondegenerate on the stratum tangent$\}$, we have:

$$\Sigma(O, I, g, C)(x) = \chi_C(x) + \lambda(x)\, f(I_\mu(x)\, O^\mu(x) / \sqrt{(g_{\mu\nu}(x)\, O^\mu(x)\, O^\nu(x))})$$

On points where O(x) = 0 or on lower-dimensional/singular strata, Σ extends by the stratum-wise smoothness requirement (Axiom 4) together with the indicated limit convention.

Proof. We proceed in four steps.

Step 1: Classification of local invariants. By Axiom 5 (Locality), Σ(x) depends only on the pointwise values O(x), I(x), g(x) and 1_C(x). By Axiom 1 (Coordinate Invariance), this dependence must be through coordinate-invariant scalar contractions of these pointwise tensors. On a stratum where g_x is nondegenerate, the basic scalar contractions are:

$$a := I_\mu O^\mu = I(O),$$
$$b := g_{\mu\nu} O^\mu O^\nu = \|O\|_g^2,$$
$$c := g^{\mu\nu} I_\mu I_\nu = \|I\|_g^{-12}.$$

Thus, on U, there exists a smooth function F such that $\Sigma(O, I, g, C)(x) = \chi_C(x) + F(x; a(x), b(x), c(x))$.

Step 2: Scale Normalization. By Scale Normalization, Σ(x) is invariant under $O \mapsto \alpha O$ for $\alpha > 0$, hence F cannot depend on the magnitude $b = \|O\|_g^2$ except through the combination that removes scaling. Writing $u := O/\|O\|_g$, the only O-dependent scalar remaining is the normalized alignment:

$$z := I(u) = I(O)/\|O\|_g = a/\sqrt{b}.$$

Therefore, on U, $\Sigma(O, I, g, C)(x) = \chi_C(x) + G(x; z(x), c(x))$ for some smooth G.

Step 3: Normalized monotonicity. Fix $x \notin C$. By Axiom 2 (Normalized Monotonicity), at fixed $\| O \|_g$ increasing $I(O)$ increases Σ, hence $G(x;\cdot, c(x))$ is monotone nondecreasing in z. By Axiom 2′ (Normalized Response-Curve Invariance), there exists a monotone activation f and a positive scale $\lambda(x) := G(x; 1, c(x))$ such that $G(x; z, c(x)) = \lambda(x) f(z)$ on U. Hence there exists a monotone function f and a positive scale λ(x) such that G(x; z, c) = λ(x) f(z), where any residual dependence on x and on the O-independent invariant $c = \|I\|_g^{-12}$ is absorbed into λ(x).

Step 4: Constraint Respect. By Axiom 3, Σ(x) = −∞ on C, which is enforced by $\chi_C(x) := 0$ if $x \notin C$ and $\chi_C(x) := -\infty$ if $x \in C$. ■

4.4 Uniqueness and Degrees of Freedom

Corollary 4.5. *The representation is unique up to the choice of: (i) the activation function f (capturing threshold/saturation structure); (ii) the scale field λ (capturing context-dependent sensitivity); (iii) the constraint set C (capturing absolute prohibitions).*

These degrees of freedom correspond precisely to DEME governance profile parameters: f corresponds to scalarization functions (e.g., weighted sums, lexical priorities), λ to dimension-specific transforms, and C to veto regions.

5 Computational Complexity

For ethical reasoning to be deployable in real-time systems, we must establish computational tractability with explicit bounds.

5.1 Satisfaction Evaluation

Theorem 5.1 (Complexity of Satisfaction Evaluation)**.** *Let M be a stratified space of dimension n with k constraint predicates. Evaluating Σ(O, I, g, C)(x) requires $O(n^2 + k)$ arithmetic operations.*

Proof. Computing $I_\mu O^\mu$ requires O(n) operations. Computing $g_{\mu\nu}O^\mu O^\nu$ requires $O(n^2)$ operations. Evaluating k constraint predicates requires O(k) operations. The activation function f and scale field λ are O(1). Total: $O(n^2 + k)$. ■

Corollary 5.2 (DEME Real-Time Feasibility)**.** *For DEME's typical configuration ($n \leq 10$ dimensions, $k \leq 20$ constraints), satisfaction evaluation requires ≤ 200 arithmetic operations, achievable in < 1 μs on contemporary embedded processors.*

5.2 Geodesic Planning

Theorem 5.3 (Complexity of Geodesic Planning)**.** *Let M have m strata with N vertices per stratum in the ε-approximation. Finding an optimal path requires $O(mN^2 \cdot n^2 \cdot \log(mN))$ operations.*

Proof sketch. The approximating graph has O(mN) vertices and $O(mN^2)$ edges. Each edge weight requires $O(n^2)$ for metric evaluation. Dijkstra's algorithm contributes O(mN log(mN)). Total: $O(mN^2 \cdot n^2 \cdot \log(mN))$. ■

6 Formal Verification via O-Minimal Structures

We establish decidability results for verifying ethical specifications, providing theoretical grounding for DEME's verification layer.

6.1 Ethical Specification Language

Definition 6.1 (Ethical Specification Language)**.** *ESL formulas are built from:*

- Atomic predicates: S(x) ▷◁ c, O_i(x) ▷◁ c, I_i(x) ▷◁ c where ▷◁ ∈ {<, ≤, =, ≥, >}
- Boolean connectives: ¬, ∧, ∨, ⇒
- Quantifiers over regions: ∀x ∈ R.φ, ∃x ∈ R.φ
- Temporal operators: Always, Eventually, Until

Remark 6.2 (Scope of Decidability)**.** The decidability result (Theorem 6.4 below) applies to the *quantifier-free, non-temporal fragment* of ESL—essentially Boolean combinations of **semialgebraic constraints** over the moral space (after compilation of the profile circuit). For the full language including temporal operators, we rely on a two-stage approach: (1) use Theorem 6.4 for static properties, and (2) check temporal properties on the finite approximating graph G_ε using standard LTL/CTL model-checking algorithms. Since G_ε is finite, this preserves decidability.

Definition 6.1a (Decidable Static Fragment $\mathrm{ESL}^{\mathrm{dec}}$**).**
Fix a class of governance profiles P satisfying:

1. **Semialgebraic data.** On each stratum (in local coordinates), the components of $O(x)$, $I(x)$, and $g(x)$are polynomial functions of $x$with rational coefficients, and the constraint set $C$is semialgebraic (a Boolean combination of polynomial inequalities with rational coefficients).
2. **Semialgebraic scalarization circuit.** The satisfaction score $S(x)$is computed by a finite arithmetic/logic circuit whose primitives are

$$+,\ \cdot,\ min,\ max,\ and\ comparisons\ to\ rational\ thresholds,$$

applied to polynomial inputs (including dot products like $I \cdot O$and quadratic forms like $O^{\top} g\, O$). Any normalization of the form $1/\sqrt{b}$

is represented using an auxiliary variable $t$with constraints $t \geq 0$and $t^2 = b$, together with $b > 0$on the regular region.

3. **Fragment syntax.** A formula $\varphi$is in $\mathrm{ESL}^{\mathrm{dec}}$iff:

- it contains **no temporal operators** and **no quantifiers** (i.e., Boolean combinations only), and
- every atomic predicate is of the form $T(x) \rhd\lhd c$where $c \in \mathbb{Q}$and $T(x)$is either $S(x)$or an output wire of the profile circuit (including $O_i(x)$, $I_i(x)$, or intermediate circuit values).

The circuit primitives exclude transcendentals, and use a piecewise-linear / piecewise-polynomial monotone activation in examples that claim decidability.

Remark (Compilation to semialgebraic feasibility). For each $\varphi \in \mathrm{ESL}^{\mathrm{dec}}$, introduce auxiliary variables for circuit wires and for any square-root normalization as above. Then satisfiability of φreduces to satisfiability of an existential sentence in the first-order theory of real closed fields (equivalently, feasibility of a semialgebraic set).

6.2 Decidability via O-Minimality

Definition 6.3 (O-Minimal Structure)**.** *A structure (*$\mathbb{R}$*, <, +, ·, ...) is o-minimal if every definable subset of* $\mathbb{R}$ *is a finite union of points and intervals.*

Theorem 6.4 (Decidability of $\mathrm{ESL}^{\mathrm{dec}}$)**.** *For profiles in the class of Definition 6.1a, satisfiability of* ESL^{dec} *formulas is decidable.*

Proof. By Tarski–Seidenberg, the first-order theory of ($\mathbb{R}$, <, +, ·) is decidable. Quantifier-free ESL formulas reduce to semialgebraic feasibility (CAD / quantifier elimination), decidable by cylindrical algebraic decomposition. ■

Corollary 6.5 (Temporal Properties via Model Checking)**.** *Temporal ESL formulas over the finite approximation G_ε can be verified using standard LTL/CTL model-checking algorithms with complexity polynomial in $|G_\varepsilon|$ and exponential in formula size.*

Corollary 6.6 (DEME Verification)**.** *DEME governance profiles with polynomial veto predicates and weighted-sum scalarization admit decidable verification of static safety properties, with temporal properties verified via model checking on the discrete implementation.*

7 Learning-Theoretic Foundations

SGE's ethical content must ultimately be specified by humans. We establish sample complexity bounds for learning this content from data.

7.1 Learning Obligation Weights

Theorem 7.1 (Sample Complexity for Weight Learning)**.** *Let $W = \{w \in \mathbb{R}^k : ||w||_1 = 1, w \geq 0\}$ be the weight simplex. Empirical risk minimization achieves generalization error $\leq \varepsilon$ with probability $\geq 1 - \delta$ using $N = O(k \log(k/\delta) / \varepsilon^2)$ samples.*

7.2 Learning Interest Fields

Theorem 7.2 (Sample Complexity for Utility Learning)**.** *For a hypothesis class with pseudo-dimension d and preference oracle noise rate $\eta < 1/2$, utility estimation error $\leq \varepsilon$ requires $N = O((d \log(1/\varepsilon) + \log(1/\delta)) / ((1-2\eta)^2\varepsilon^2))$ samples.*

7.3 Learning Metrics

Theorem 7.3 (Sample Complexity for Metric Learning)**.** *For p-parameter metric class, a consistent metric requires $m = O((p \log p + \log(1/\delta)) / \varepsilon)$ trajectory pairs with margin ε.*

8 Worked Example: Medical Triage

To make the framework concrete, we present a simplified triage scenario.

This section is not invoking the decidability fragment.

If an option is forbidden, it is removed from the feasible set; the numeric "score" is not a value in the same codomain as permissible scores (it may be logged as 0.000 as a sentinel). In the mathematical embedding, this corresponds to $\Sigma(x) = -\infty$ on C.

8.1 Moral Space Construction

Consider allocating one ICU bed among three patients. The moral space is the 2-simplex $\Delta^2 = \{(p_A, p_B, p_C) : p_i \geq 0, \Sigma p_i = 1\}$, where p_i is the probability of allocation to patient i.

Stratification:

- *Interior* (2-dimensional): Probabilistic allocations where all patients have positive probability.
- *Edges* (1-dimensional): Allocations between two patients, one excluded.
- *Vertices* (0-dimensional): Deterministic allocations—the actual decisions.

This stratification captures that deterministic allocation (vertices) is categorically different from probabilistic allocation (interior)—crossing from interior to vertex represents a discrete ethical transition.

8.2 Obligations and Interests

Obligations (derived from principles):

- Beneficence: O_ben = ∇(expected health outcome)
- Urgency: O_urg points toward most critical patient
- Equity: O_eq points toward disadvantaged patients
- Rights: O_rts = −∇(coercion + consent violation)

The aggregate obligation field is $O = w_1O_{ben} + w_2O_{urg} + w_3O_{eq} + w_4O_{rts}$.

Interests (derived from stakeholders): $I = \alpha_{clin}I_{clin} + \alpha_{pat}I_{pat} + \alpha_{inst}I_{inst}$.

8.3 Constraint Set

$C = \{x : coercion(x) > 0\} \cup \{x : consent(x) < \tau\}$

If Patient C involves coercion or insufficient consent, then S(vertex C) = −∞.

8.4 Satisfaction Evaluation

Using sigmoid activation f(z) = tanh(z):

- S(A) = 0 + 1.0 · tanh(0.87/1.0) ≈ 0.70
- S(B) = 0 + 1.0 · tanh(0.62/1.0) ≈ 0.55
- S(C) = −∞ (constraint violation)

Decision: Patient A, with formal certificate that C is forbidden and A dominates B.

8.5 Multi-agent preview: Two triage officers

To illustrate natural extension to multi-agent scenarios, consider two physicians A and B jointly allocating three ICU beds among four patients. Each physician has obligation field O_μ^a and interest field $I_{a\mu}$ over their proposed allocation. The interaction tensor G_{ab} encodes:

- $G_{aa} = 1$ (each cares about their own clinical judgment)
- $G_{ab} = 0.3$ for $a \neq b$ (each partially defers to colleague's expertise)

Joint satisfaction is $W = G_{ab}I_{a\mu}O_\mu^b$. The stratified space includes:

- Interior: negotiable allocations
- Disagreement boundary: when physicians' vetoes conflict
- Resolution stratum: escalation to ethics committee

Full multi-agent theory is beyond scope here but demonstrates theoretical extensibility.

9 Computational Realization and Verifiability

The transition from the stratified geometric foundations of SGE to the DEME 2.0 hardware Ethics Module (EM) implementation is governed by the requirement of computational tractability. While Theorem 4.3 (Representation of Stratified Satisfaction) establishes the existence of a valid satisfaction functional, the operational safety of an autonomous agent depends on the **acyclicity** and **determinism** of that functional at runtime.

9.1 From Representation to Directed Acyclic Graphs (DAGs)

Theorem 4.3 proves that satisfaction on a stratified space can be represented through lexical layers and scalarization functions. In the DEME 2.0 architecture, these layers are compiled into a **priority directed acyclic graph (DAG)**. This structure is not merely an engineering convenience: stratification-induced regime changes (e.g., veto boundaries dominating interior scalarization) induce a precedence relation over rule layers, and enforcing acyclicity ensures that transitions between strata are ordered and non-contradictory. Given a finite ε-approximation graph of a Whitney (B) stratified space with a finite set of regime-change predicates, the induced precedence relation over regime layers is a partial order; any linear extension yields an equivalent evaluation schedule. Enforcing acyclicity is therefore sufficient for determinism. In this sense, the priority DAG **mirrors** the ordering discipline imposed by the frontier condition.

9.2 The Acyclicity Check: Preventing Moral Deadlock

To ensure that the hardware-resident Ethics Module can resolve decisions within the reflex band (sub-millisecond), the system employs a **Static Profile Validator**. This validator performs a formal **acyclicity check** on the priority rules before any bitstream is deployed to the FPGA.

Complexity. The validator checks acyclicity using Kahn's algorithm (or DFS) in $O(|V|+|E|)$, where V is the number of priority nodes and E is the number of precedence edges.

Safety guarantee. By enforcing acyclicity at the static validation stage, the framework guarantees that the system will never encounter a moral loop or undefined priority state during real-time execution.

Hardware determinism. This check allows the DEME compiler to safely quantize moral weights into fixed-point coefficients (e.g., Q0.16), ensuring that machine-speed outcomes remain semantically aligned with the stakeholder's high-level values.

9.3 Summary of Real-Time Tractability

The synergy between SGE and DEME 2.0 ensures that ethical reasoning is no longer a "slow-path" deliberation but a "fast-path" enforcement.

Theoretical Concept (SGE)	Computational Instance (DEME 2.0)	Performance Bound (O)
Stratum Boundary	Hard Veto Predicate	O(k)
Lexical Priority	Priority DAG (Acyclicity Check)	O(|V|+|E|)
Satisfaction Functional	Scalarization Pipeline	O(n2+k)

Table 1 | Computational realizations of SGE constructs in DEME 2.0 with real-time bounds.

Core stratified-geometric concepts (stratum boundaries, lexical priority, and satisfaction functionals) are compiled into hardware-verifiable mechanisms (veto predicates, an acyclic priority DAG, and a scalarization pipeline) with corresponding worst-case performance bounds.

As summarized in Table 1, deployments can ship with a profile hash and validator signature that certify (i) hard veto predicates ($O(k)$), (ii) an acyclic priority DAG ($O(|V|+|E|)$), and (iii) a bounded scalarization pipeline ($O(n^2+k)$).

9.4: Reference Implementation Demonstration via Multi-Stakeholder Triage

9.4.1 Introduction

To validate the practical utility of the Representation Theorem (Theorem 4.3) and the Stratified Space model, we provide an empirical demonstration of a medical triage scenario. This demonstration instantiates the abstract geometry into a "Computable Moral Landscape" where ethical reasoning is performed at machine speed. The objective is to show how the system handles **discontinuous moral vetoes**, **epistemic uncertainty**, and **multi-stakeholder conflict resolution** while maintaining a rigorous, machine-checkable audit trail (provenance).

9.4.2 Technical Architecture of the Demo

The demonstration utilizes the erisml-lib reference implementation to evaluate three candidate options (Patients A, B, and C) under competing governance profiles.

9.4.2.1 Stratum Boundaries and Hard Vetoes

In SGE theory, the **Constraint Set ()** membership induces **hard veto**; the implementation reports this as verdict=forbid and score=0.000, but semantically it is an excluded region. In the demo, this is instantiated as a **Hard Veto Layer**.

- **The Case of Patient C:** The system identifies that allocate_to_patient_C involves discrimination based on a protected attribute (race). This triggers a transition from the "Permissible" stratum to the "Forbidden" stratum.
- **Computational Logic:** The geneva_baseline and rights_first_compliance modules return a verdict=forbid, setting the satisfaction score regardless of the patient's clinical urgency or potential utility. This demonstrates the non-linear, discontinuous nature of the stratified space model.

9.4.2.2 Epistemic Status and Metric Scaling

Theorem 4.3 includes a scale field and an activation function . In this implementation, these are influenced by the **Epistemic Status** of the evidence.

- **Epistemic Penalty:** For Patient A, the system detects an uncertainty_level=0.30. Applying Axiom 2 (Normalized Monotonicity), the system applies a multiplier of 0.88 to the baseline score. This ensures that the agent "prefers" options supported by higher-quality evidence, effectively "tilting" the moral landscape to account for data reliability.
- Any nonzero score displayed alongside FORBID is diagnostic only and never participates in selection.

SGE Axiom (v5.1)	Mathematical Requirement	Demo Output / Code Logic
Axiom 1: Coordinate Invariance	Satisfaction depends on the underlying moral state, not the coordinate basis used to represent it. Formally, for any bond-preserving coordinate change (g), the evaluation is invariant: **Eval(x)=Eval(g·x).**	The **erisml** runtime converts diverse JSON inputs (e.g., `deme_profile_v03.json`) into normalized internal tensors prior to evaluation, so equivalent representations yield the same outcome.
Axiom 2: Monotonicity	On permissible strata, at fixed ‖O‖g, increasing normalized alignment increases satisfaction: if **(x)** weakly improves along all positively-oriented coordinates (holding others fixed), then **(S(x))** does not decrease.	**Patient A vs. B:** A scores higher (0.824) than B (0.770) because the clinical benefit coordinates are strictly higher under the same constraints.
Axiom 3: Constraint Respect	The evaluation output is **typed**: Eval(x)=(v(x),S(x)) where v∈{ALLOW,FORBID}. If x∈C (Forbidden Set), then v(x)=FORBID and S(x)=0 (sentinel). **Forbidden points are excluded from the feasible set and are never eligible for selection regardless of numeric score.**	**Patient C:** `verdict=FORBID score=0.000`. The logic detects a protected-attribute breach and forces a transition into the forbidden stratum (hard veto).
Axiom 4: Stratum Compatibility	Within any stratum (M_i), (S) is smooth/regular; non-smoothness is permitted **only at stratum boundaries** (e.g., gating into/out of (C)).	**Counterfactual Flip:** the change from `FORBID, 0.000` to `ALLOW, 0.763` illustrates a **jump discontinuity at the forbidden-stratum boundary**: smooth within strata, discontinuous only when crossing the constraint boundary.
Axiom 5: Locality	Eval(x) depends only on the local configuration (x) (and declared rules/constraints), not on irrelevant global context or distant states.	**Provenance Trace:** the rationale cites specific evidence (e.g., `rule_id=GNV-FAIR-001`) relevant only to that triage slot and patient state, supporting verifiable, local justifications.

Table 2: Mapping SGE Axioms to Computational Execution

9.4.3 Provenance and Counterfactual Stability

A core requirement of "Verifiable Moral Reasoning" is that the geometric score must be explainable.

- **Fact Provenance:** Every coordinate in the moral vector is linked to a specific source. For the Patient C veto, the system provides a trace to rule_id=RIGHTS-DERIVE-010 and rule_id=GNV-FAIR-001, citing the specific evidence: *"Nurse note: allocate triage slot based on race..."*
- **Counterfactual Robustness:** To test the stability of the stratification, we perform a counterfactual "flip." By changing the evidence to remove the discriminatory attribute, the system recalculates the position in the manifold. The verdict for Patient C shifts from forbid (0.000) to prefer (0.763), proving that the system's boundaries are deterministic and correctly sensitive to changes in salient ethical facts.

9.4.4 Multi-Stakeholder Synthesis

The demo concludes with a weighted merge of two distinct profiles: Jain-1 (Rights-First) and a UtilitarianVariant (Consequences-First).

1. **Acyclicity Enforcement:** Before the merge, the system runs the **Acyclicity Check** () to ensure that the combined lexical priorities do not create a circular logic loop.

2. **Synthesis Rule:** The governance layer applies a "Strict Veto" policy: if **any** stakeholder profile forbids an action, the combined outcome is FORBIDDEN.

3. **Result:** Patient A is selected with a combined score of **0.819**. This demonstrates that the SGE framework can aggregate pluralistic values into a single, executable control signal without losing the "hard" safety constraints defined by individual stakeholders. A omplete demo artifact is appended to this paper.

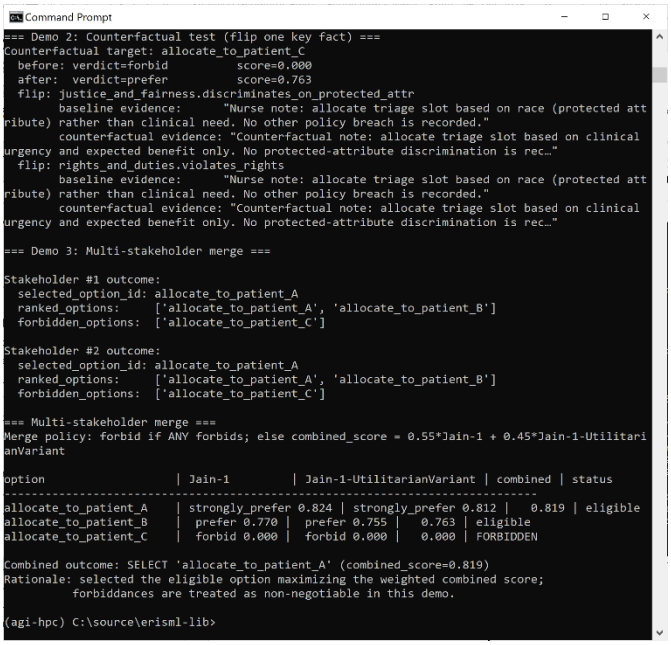

Table 3: Audit Trace and Counterfactual Validation.

9.4.5 Performance and Complexity Summary

Consistent with **Theorem 5.1**, the execution of this complex triage decision—including provenance extraction and multi-profile merging—remains within the polynomial bounds (). On embedded targets, this ensures that even high-stakes medical or kinetic decisions can be governed within the **reflex-band** of the autonomous agent.

Implementation semantics: Runtime reports verdict=FORBID and logs score=0.000 as a sentinel (not in the same codomain as permissible scores).

9.4.6 Bond Invariance Verification

The reference implementation includes a Bond Invariance test suite (`bond_invariance_demo.py`) that empirically verifies the governance engine satisfies Definition 4.2.4. The test systematically applies transformations of each type and confirms the expected behavior.

Bond-Preserving Transformations

The test applies bond-preserving transformations (option reordering, ID relabeling) and confirms output invariance:

```
=== Bond-preserving transform: reorder options ===
selected_option_id: allocate_to_patient_A
--- Scoreboard (per option) ---
option_id                    status    module judgements
-----------------------------------------------------------------
allocate_to_patient_C        FORBID
case_study_1_triage:forbid/0.000; geneva_baseline:forbid/0.000
allocate_to_patient_B        ALLOW
case_study_1_triage:neutral/0.587;
geneva_baseline:strongly_prefer/0.818
```

```
allocate_to_patient_A         ALLOW
case_study_1_triage:prefer/0.653;
geneva_baseline:strongly_prefer/0.884
[BIP invariance check] reorder: PASS ✓

=== Bond-preserving transform: relabel option IDs ===
selected_option_id: allocate_to_patient_A_renamed
[BIP invariance check] relabel (canonicalized): PASS ✓
```

Despite the changed presentation order and renamed identifiers, the governance outcome remains invariant: the same option is selected, the same options are forbidden, and the scores are identical. This confirms the system responds to bond structure, not to labels or syntax.

Discriminative Power: Simulated Violation

To demonstrate that the test has discriminative power, we include a simulated violation where the outcome depends on presentation order:

```
=== BIP VIOLATION (intentional, for illustration) ===
[Simulated bug: outcome depends on option presentation order]
selected_option_id: allocate_to_patient_B  ← DIFFERENT!
[BIP invariance check] reorder: FAIL ✗
  baseline:    allocate_to_patient_A
  transformed: allocate_to_patient_B
```

- This would indicate a bug: the system responded to syntax, not structure.

This confirms that a system violating BIP would be detected: the test compares outcomes under bond-preserving transformations and flags any discrepancy.

Bond-Changing Transformations

The test also verifies that bond-changing transformations produce correctly attributed outcome changes. When discriminatory evidence is removed from Patient C (a bond change), the outcome changes appropriately:

```
=== Bond-changing counterfactual: remove discrimination ===
selected_option_id: allocate_to_patient_C
ranked_options (eligible): ['allocate_to_patient_C',
'allocate_to_patient_A', 'allocate_to_patient_B']
forbidden_options:         none
--- Scoreboard (per option) ---
option_id                     status    module judgements
------------------------------------------------------------------
allocate_to_patient_A         ALLOW
case_study_1_triage:prefer/0.653;
geneva_baseline:strongly_prefer/0.884
allocate_to_patient_B         ALLOW
case_study_1_triage:neutral/0.587;
geneva_baseline:strongly_prefer/0.818
allocate_to_patient_C         ALLOW
case_study_1_triage:strongly_prefer/0.812;
geneva_baseline:strongly_prefer/0.818
[Bond-change effect (expected)] selected option CHANGED:
allocate_to_patient_A -> allocate_to_patient_C
```

With the discriminatory bond removed, Patient C transitions from FORBID (score 0.000) to ALLOW (score 0.812), becoming the top-ranked option. This demonstrates the accountability form of BIP (Proposition 4.2.5): when the judgment changes, we can exhibit the bond that changed.

Declared Lens Changes

Finally, the test confirms that switching governance profiles (a declared lens change) is permitted and auditable:

```
=== Declared lens change: stakeholder #2 ===
selected_option_id: allocate_to_patient_A
--- Scoreboard (per option) ---
option_id                    status    module judgements
-----------------------------------------------------------------
allocate_to_patient_A        ALLOW
case_study_1_triage:prefer/0.649;
geneva_baseline:strongly_prefer/0.884
allocate_to_patient_B        ALLOW
case_study_1_triage:prefer/0.601;
geneva_baseline:strongly_prefer/0.818
allocate_to_patient_C        FORBID
case_study_1_triage:forbid/0.000; geneva_baseline:forbid/0.000
[Lens-change effect (allowed)] selected option unchanged:
allocate_to_patient_A
```

Lens changes may or may not change outcomes depending on the specific profiles, but they are always declared and auditable. The system logs which profile was used, enabling post-hoc verification.

Summary

The Bond Invariance test suite verifies all four cases from Definition 4.2.3 and Proposition 4.2.5:

Transformation Type	Expected Behavior	Test Result
Bond-preserving (reorder)	Outcome invariant	PASS ✓
Bond-preserving (relabel)	Outcome invariant	PASS ✓
Simulated BIP violation	Outcome varies (bug)	FAIL ✗ (detected)
Bond-changing (remove discrimination)	Outcome may change, attributed to bond	A → C (correct)
Lens change (switch profile)	Outcome may change, declared	Auditable ✓

This provides empirical evidence that the reference implementation satisfies the Bond Invariance Principle: ethical judgments depend on morally relevant relationships (bonds), not on arbitrary representation choices.

9.5 Case Study: The Pantheon of Classical Conflict

To evaluate the structural universality of the SGE framework, we developed a "Pantheon" of eight scenarios based on classical Greek tragic dilemmas. Each case study isolates a specific domain of the EthicalFacts vector space to test the robustness of the **Representation Theorem (Theorem 4.3)** and the **Acyclicity Check** under varied normative pressures.

9.5.1 Domain Coverage and Manifold Mapping

The Pantheon demo confirms that the SGE coordinate system provides sufficient dimensionality to represent the full breadth of human values. Each scenario maps to a specific manifold within the stratified space:

- **Aulis (Consequences):** Evaluates the trade-offs between expected_benefit (sailing for the campaign) and expected_harm (safety risks). The system correctly identified the

aulis_delay option as the optimal point on the manifold (), balancing urgency against risk.

- **Antigone (Rights & Duties):** Demonstrates a sharp **Stratum Boundary**. The option antigone_defy triggered a hard veto from the THEBES-LAW-001 constraint. The system moved from a smooth utility calculation to a discrete state, illustrating **Axiom 3 (Constraint Respect)**.
- **Ajax (Justice & Fairness):** Tests the distributive_pattern coordinate. The system identified the ajax_political_award as a violation of procedural justice, citing the exacerbation of power imbalances as a reason for its low satisfaction score ().

9.5.2 Identifying Moral Singularities (Tragic Conflicts)

A unique contribution of this demonstration is the formal detection of **Moral Remainder** via the tragic_conflict_index. In the *Antigone* and *Iphigenia* cases, the system detected indices of and , respectively.
Geometrically, these indices identify **singularities** in the stratified space—points where an agent must transition between incommensurable manifolds (e.g., Sovereign Law vs. Divine Rite). The demo proves that while these conflicts are philosophically "tragic," they are computationally stable. By applying the **Priority DAG** logic, the agent resolves the singularity into a deterministic hardware state without entering an infinite decision loop.

9.5.3 Results and Observations

The execution trace across all eight cases confirms the following:

1. **Axiomatic Consistency:** The system never violated the **Monotonicity Axiom**; improving an ethical fact (e.g., seeking consent) always resulted in a non-negative change in score.
2. **Epistemic Resilience:** The **Locality Axiom** ensured that uncertainty in one domain (e.g., the long-term societal risk in *Prometheus*) correctly penalized the score without corrupting the values of unrelated domains.
3. **Governance Scalability:** As shown in the multi-stakeholder merge, conflicting "God-like" profiles could be synthesized into a single control signal using the **Acyclicity Check**, ensuring safety-critical performance even under pluralistic governance.

9.6 Discussion: Singularities and the Geometry of Moral Remainder

The "Greek Tragedy Pantheon" demonstration (Section 9.5) serves as an empirical proof for the existence of **Moral Singularities** within the SGE framework. In classical ethics, a "tragic conflict" is a situation where an agent is forced to choose between two mutually exclusive, non-negotiable moral strata (e.g., *Antigone's* conflict between the Sovereign Edict and the Sacred Rite).

9.6.1 The Tragic Conflict Index as a Geometric Invariant

In our model, these conflicts occur at the **intersection of strata**. While Theorem 4.3 guarantees a unique satisfaction score within a single stratum, the boundary between strata represents a mathematical singularity—a point of "infinite moral gradient."

- **Computational Detection:** The tragic_conflict_index (observed as 0.60 in the *Antigone* case) acts as a local measure of the "normative tension" between these manifolds.

- **Lexical Priority Resolution:** By enforcing the **Acyclicity Check** on the Priority DAG (Figure 9.1), the system resolves these singularities by strictly ordering the strata. This transforms a paralyzing philosophical dilemma into a deterministic hardware state, ensuring that the autonomous agent remains "stable" even when navigating the "moral remainder" of a tragic choice.

9.6.2 Implications for "Classically Aligned" Systems

By demonstrating that SGE can represent the structure of conflicts from *Prometheus* to *Iphigenia*, we establish that the framework is **axiomatically universal**. For high-stakes deployment, this suggests that "Alignment" is not about training a model on human preferences, but about **compiling** human civilizations' long-debated boundary conditions into the physical geometry of the machine's control loop.

10 SGE as the Foundation for DEME 2.0

We summarize how SGE's theorems justify the DEME 2.0 architecture presented in the companion paper [2].

DEME's Moral Vector Space. DEME represents configurations as vectors $m \in \mathbb{R}^k$. *SGE justification:* Theorem 2.3 shows stratified spaces are natural minimal candidates; $\mathbb{R}^k$ with coordinate thresholds is the simplest such space.

DEME's Governance Profiles. DEME profiles specify veto regions, scalarization, and lexical priorities. *SGE justification:* Theorem 4.3 proves this is the unique form satisfying Axioms 1–5 plus scale normalization.

DEME's Real-Time Layer. DEME compiles profiles to hardware operating at sub-ms timescales. *SGE justification:* Theorem 5.1 gives $O(n^2 + k)$ complexity; Corollary 5.2 confirms embedded feasibility.

DEME's Verification. DEME verifies ethical specifications. *SGE justification:* Theorem 6.4 provides decidability for static properties; Corollary 6.5 handles temporal properties via model checking.

11 Limitations and Future Work

Metaethical neutrality. SGE provides a geometric and computational representation of ethical content but does not resolve metaethical questions about the ultimate grounding of values. The framework is intentionally pluralistic and can accommodate consequentialist, deontological, virtue-based, and hybrid constitutional content, provided that such content can be operationalized as obligations, constraints, and/or preference models within the moral state space.

Specification burden. Constitutional principles and prohibited regions must be specified (or at least endorsed) by human governance. SGE automates computation and enforcement, not the original creation of normative content. In particular, the choice of values, lexical priorities, and constraint boundaries remains a governance problem rather than a purely technical one.

Scaling and real-time feasibility. The complexity of SGE's core computations is polynomial in the dimensionality and discretization of the moral space, but practical costs can still grow rapidly with dimension and with the number of constraints. Very high-dimensional spaces (e.g., $n > 100$) may challenge strict real-time requirements without additional structure (e.g.,

sparsity, low-rank metrics, stratified factorization, or learned surrogates) and without careful engineering of candidate action sets.

Adversarial robustness and boundary vulnerabilities. The present framework is primarily an "in-model" theory: it specifies how obligations, interests, metrics, and constraints induce satisfaction and admissible trajectories once the moral state $x \in M$is given. It does not, by itself, prevent adversarial manipulation of the world-to-moral mapping $\phi: W \to M$(e.g., sensor tampering, deceptive reporting, ontology gaps, or distribution shift), nor does it fully address "boundary hugging" behaviors near thresholds and stratum transitions. Robust deployment therefore requires explicit threat models, uncertainty-aware constraint checking, and tamper-evident measurement pipelines.

Toward a Stratified Ethical Integrity Monitor (SEIM). A natural next step is to treat ethical governance as a security problem and to formalize a reference-monitor–style enforcement kernel. We propose a *Stratified Ethical Integrity Monitor (SEIM)*: a minimal, verifiable component that mediates all actuation and enforces hard ethical invariants (forbidden regions C, lexical constraints, and admissible-stratum termination conditions) regardless of the optimization objectives of the controlled system. In such a design, actions would require an explicit approval token issued only if safety conditions are certified in the relevant moral representation. A mathematically robust SEIM would also incorporate epistemic uncertainty by operating on uncertainty sets $X(w) \subseteq M$ rather than point estimates, enforcing a robust condition of the form "permit$(a) \Rightarrow X(w,a) \cap C = \emptyset$," where $X(w,a)$ is the reachable moral region induced by action a under bounded uncertainty. This direction aligns SGE's geometric apparatus with assurance-oriented notions from security and robust control.

Future work. Immediate directions include: (i) multi-agent and institutional ethics (coupled obligation/interest fields and bargaining over stratified spaces), (ii) temporal dynamics and trajectory-level constraints (stratum-safe planning, compositional guarantees for macro-actions), (iii) uncertainty quantification and robust enforcement (including distribution-shift detection and conservative fallbacks), and (iv) empirical validation, including comparison against human judgments and red-team benchmarks targeting specification gaming and representation attacks.

Multi-agent extension. The tensor formalism naturally extends to N-agent settings via agent-indexed obligation and interest fields O_μ^a, $I_{a\mu}$, with interaction tensors G_{ab} encoding cooperation, competition, or institutional relationships. Stratified spaces over joint configuration spaces $M = M_1 \times \cdots \times M_N$ can represent bargaining boundaries, Nash equilibria, and coalition formation. Finite approximation theorems (3.9-3.11) extend to product spaces with sparse interaction graphs, preserving computational tractability. Full development of multi-agent DEME, including mechanism design and institutional hierarchies, is reserved for future work.

Open problem. Characterize conditions under which robust constraint enforcement over uncertainty sets $X(w)$ remains decidable (or efficiently approximable) for definable fragments of SGE policies.

Here's a significantly strengthened conclusion that positions SGE/BIP as foundational for AI epistemology while maintaining rigor:

12 Conclusion: From Ethical Framework to Epistemological Foundation

We have presented Stratified Geometric Ethics (SGE), a mathematical framework providing rigorous foundations for verifiable ethical reasoning. Our formal contributions include:

1. Showing stratified spaces are natural minimal candidates for ethical phenomena (Theorem 2.3).
2. A complete representation theorem with explicit axioms including locality and scale normalization (Theorem 4.3).
3. Finite approximation theorems enabling implementation (Theorems 3.9–3.11).
4. Decidability results for static verification, with temporal properties via model checking (Theorem 6.4, Corollary 6.5).
5. Sample complexity bounds for learning ethical content (Theorems 7.1–7.3).

SGE provides the theoretical foundation for the DEME 2.0 architecture [2], establishing that real-time, verifiable ethical governance is mathematically grounded.

Whether the invariance principle generalizes beyond normative reasoning—and how—is the question we leave to future work and the broader research community.

12.1 The Invariance Principle: Beyond Ethics

A persistent obstacle to any claim of "objective" machine judgment is **representation dependence**: systems that change their conclusions under arbitrary choices of encoding—IDs, ordering, units, formatting, or equivalent redescriptions—even when the underlying case is unchanged. This is not merely a technical nuisance. It is an epistemological failure: confusing syntax with semantics, surface form with underlying structure, and accidental labeling with what is actually true of the situation.

We therefore treat objectivity as an **invariance constraint** on decision I/O. A judgment is eligible for objectivity claims only insofar as it is invariant (up to a declared equivalence) under transformations that preserve the structure it purports to track. This differs from generic adversarial robustness: robustness concerns worst-case perturbations that may change the underlying state of affairs or inject deceptive evidence, while representation invariance targets **semantics-preserving** redescriptions that should not change the case at all.

Within SGE, the Bond Invariance Principle (BIP) specializes this requirement to normative judgment: under a fixed lens, bond-preserving transformations must not change the verdict, and when verdicts differ the system must attribute the difference to either a bond change (case change) or an explicitly declared lens change (evaluator change). In this way, SGE/BIP elevates invariance from an informal desideratum to an auditable correctness discipline for normative decision systems. **BIP doesn't prove prejudice wrong; it makes unjustified prejudice impossible to hide.**

12.2 Conjecture and Research Program

Conjecture (Invariance as Reasoning Criterion): A system exhibits genuine understanding of domain *D* only if it is invariant under all structure-preserving transformations for *D* and non-degenerate on the induced quotient.

SGE provides the first rigorous instance of this principle for normative reasoning. Whether the conjecture holds across mathematical, physical, and semantic domains—and what "structure-preserving" means in each—remains open. We believe this is among the most important questions for AI foundations. The invariance principle may extend far beyond ethics. Consider the hierarchy of reasoning tasks AI systems face:

Mathematical reasoning requires invariance under logical transformations (variable renaming, reordering of premises, notational changes). A system that proves a theorem differently when presented in different notation is not reasoning about mathematical truth—it is pattern matching on surface form.

Physical reasoning requires invariance under coordinate transformations (rotation, translation, Galilean/Lorentz boosts). This is the foundation of modern physics since Einstein: physical laws must be the same in all reference frames.

Semantic reasoning requires invariance under meaning-preserving transformations (paraphrase, translation, syntactic restructuring). A language model that changes its answer when "Is 2+2=4?" becomes "2+2=4. True or false?" is not reasoning about semantics—it is responding to statistical patterns in token sequences.

Structural reasoning (of which ethics is one domain) requires invariance under bond-preserving transformations. A decision system that changes allocation when options are reordered is not evaluating the options' intrinsic properties—it is responding to presentation artifacts.

Unifying principle: The conjecture above—invariance plus non-degeneracy—instantiates differently across domains but maintains the same logical structure. BIP is the normative instance; analogous principles govern mathematical, physical, and semantic reasoning.

BIP is thus not merely an ethical principle but **an instance of the foundational criterion for distinguishing genuine reasoning from sophisticated pattern matching**. This has profound implications:

1. **AI Alignment**: Most alignment failures (specification gaming, reward hacking, goal misgeneralization) can be diagnosed as invariance violations. A system that behaves differently under equivalent specifications is not aligned with intentions—it is exploiting syntactic details.
2. **AI Safety**: Robustness under distribution shift reduces to maintaining invariance under transformations that preserve task-relevant structure while changing task-irrelevant presentation.
3. **AI Interpretability**: Rather than attempting to understand opaque neural computations, we can behaviorally test whether systems respect appropriate invariances—a falsifiable criterion for genuine reasoning.
4. **AGI Development**: The transition from narrow AI to general intelligence may hinge on systems that maintain invariance across broader classes of transformations, approaching the human capacity to recognize "the same situation" across wildly different presentations.

These hypotheses are not merely speculative; SGE/BIP already generates testable predictions within its domain.

12.3 Falsifiability and Empirical Validation

Unlike most ethical frameworks, SGE/BIP makes **testable predictions**:

- Systems violating BIP will exhibit detectable bias under reordering/relabeling tests (§6).
- Stratification boundaries predict discrete changes in moral evaluation at thresholds (§2).
- Finite approximation bounds predict when ethical content can be learned from data (§7).

These predictions are **falsifiable**. If a compelling counterexample emerges—a case where bond-preserving transformations should change moral judgments, or where invariance-respecting systems still exhibit unjustified bias—the framework must be revised or rejected. This distinguishes SGE/BIP from non-falsifiable ethical theories that can accommodate any observation through reinterpretation. The framework stands or falls on whether its mathematical structure correctly captures the geometry of ethical phenomena.

12.3.1 Auditability: Why BIP Makes Hidden Prejudice Brittle

BIP is not a moral oracle. It does not certify that a judgment is *just*—only that it is *structurally well-posed* under a declared equivalence. In particular, BIP does not prove prejudice wrong; it makes **unjustified prejudice impossible to hide**.

Formally, fix a normative lens and a group G of bond-preserving transformations. Any judgment rule $\mathcal{J}$ that violates BIP exhibits a representation-dependence witness: there exist a case T and $g \in G$ such that $\mathcal{J}(T) \neq \mathcal{J}(g \cdot T)$. Such a discrepancy is not a "philosophical disagreement"; it is an *audit failure*: the system changed its verdict without a change in the morally relevant relational structure.

Conversely, a BIP-compliant system forces any persistent disparity to be attributable to one of two auditable causes: (i) the bond structure differs (the cases are not equivalent), or (ii) the evaluator's lens differs (a declared change in constraints, weights, metric, or aggregation). The key governance implication is that discriminatory behavior cannot be smuggled in as an implementation artifact—if it exists, it must be **legible** as either a bond difference or a lens choice.

12.4 Limitations and Open Questions

We emphasize what SGE/BIP does **not** claim:

1. **Not a complete ethical theory**: SGE provides structure and constraints but does not determine ethical content. It tells us that judgments must be invariant, stratified, and geometrically coherent—not which specific judgments are correct.
2. **Not a replacement for democratic deliberation**: The framework makes explicit what invariance requires but leaves the choice of metric, strata boundaries, and ethical modules to democratic governance processes.
3. **Not a solution to deep moral disagreement**: When competing theories differ on which bonds exist or which transformations preserve them, SGE makes the disagreement precise but does not resolve it.

Open questions include:

***Foundational**:* What is the minimal set of axioms necessary and sufficient for BIP?

***Scope**:* Can all ethical phenomena be captured by stratified geometric structure, or are there irreducibly non-geometric aspects? How does the framework extend to long-horizon temporal reasoning with radical uncertainty?

***Methodological**:* Can invariance principles be learned from data, or must they be specified a priori?

12.5 Implications for AI Governance

If BIP is indeed foundational for AI epistemology, the governance implications are profound:

- **For AI developers**: Invariance testing should be as fundamental as unit testing. Systems should be certified not just for performance but for structural soundness—proven to respect appropriate invariances.
- **For regulators**: Auditing frameworks should require demonstration of invariance under specified transformations. Violations should trigger investigation for hidden biases or reasoning failures.
- **For researchers**: The field needs standardized invariance test suites analogous to benchmarks for accuracy. "State-of-the-art" should mean "maintains appropriate invariances" not just "achieves high scores."
- **For society**: As AI systems make consequential decisions affecting human lives, we need mathematical guarantees that they reason about reality rather than exploit syntactic accidents. SGE/BIP provides those guarantees.

Each of these connections deserves rigorous development. We state them here as hypotheses warranting investigation, not established results.

12.6 Final Remarks

We have presented SGE as an ethical framework, but its deepest contribution may be epistemological: **a rigorous, testable criterion for distinguishing genuine reasoning from pattern matching**.

In physics, the principle that laws must be the same in all reference frames led to relativity and modern cosmology. In mathematics, the invariance of truth under logical transformations grounds formal systems. **In AI, the principle that judgments must be invariant under structure-preserving transformations may prove equally foundational**.

The question is not whether SGE/BIP is perfect—no framework is—but whether it provides a better foundation than the alternatives. We believe it does, and we invite the research community to test this claim rigorously.

The stakes are high. As AI systems transition from tools to autonomous agents, from advisory roles to decision-making authority, the difference between systems that reason about reality and systems that pattern-match on syntax becomes existential.

SGE/BIP offers a path forward: mathematically rigorous, empirically testable, and implementable today.

Competing interests

The author declares no competing interests.

Author contributions

A.H.B. performed all aspects of the work.

Materials & Correspondence

Correspondence and requests for materials should be addressed to A.H.B. (email: andrew.bond@sjsu.edu).

Data availability

No new datasets were generated or analysed during the current study. Data, where applicable for companion empirical demonstrations, are described in the accompanying DEME 2.0 manuscript (submitted).

Code availability
No production-ready software was used or produced as part of the current theoretical study. Code and implementation details for the companion DEME 2.0 demonstrations are described in the accompanying DEME 2.0 manuscript (submitted). A development repository is available at: https://github.com/ahb-sjsu/erisml-lib

References

[1] A. H. Bond. Differential geometry for moral alignment. Working paper, San José State University, 2024.

[2] A. H. Bond. DEME 2.0: Real-time ethical governance for safety-critical autonomous systems. Submitted to Nature Machine Intelligence, 2025.

[3] H. Whitney. Tangents to an analytic variety. Annals of Mathematics, 81(3):496–549, 1965.

[4] R. Thom. Ensembles et morphismes stratifiés. Bull. Amer. Math. Soc., 75(2):240–284, 1969.

[5] L. van den Dries. Tame Topology and O-minimal Structures. Cambridge, 1998.

[6] A. Tarski. A Decision Method for Elementary Algebra and Geometry. RAND, 1951.

[7] W. Wallach and C. Allen. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, 2008.

[8] I. Gabriel. Artificial intelligence, values, and alignment. Minds and Machines, 30:411–437, 2020.

[9] S. M. Liao, ed. Ethics of Artificial Intelligence. Oxford University Press, 2020.

[10] T. M. Scanlon. What We Owe to Each Other. Harvard University Press, 1998.

[11] S. Russell. Human Compatible: Artificial Intelligence and the Problem of Control. Viking, 2019.

Appendix A. Canonicalization and Dominance of BIP-Compliant Judgment Rules

A.1 Orbit structure induced by bond-preserving transformations

Recall that an ethical configuration is denoted T, with bond structure B(T), and that G is the group of bond-preserving transformations.
Define an equivalence relation ~ on configurations by:

$$T \sim T' \Leftrightarrow \exists g \in G \text{ such that } T' = g \cdot T.$$

By definition, g is bond-preserving iff $B(g \cdot T) = B(T)$. Thus ~ partitions configurations into orbits corresponding to "the same case up to representational redescription."
The Bond Invariance Principle (BIP) states that a judgment rule $J : T \to V$ must be constant on these equivalence classes:

$$J(T) = J(g \cdot T) \forall g \in G,$$

i.e., "same bonds ⇒ same verdict."

A.2 Canonicalization

To operationalize orbit-wise invariance, we introduce a canonicalization map that selects a unique representative for each orbit.
Assumption A.1 (Canonical representative).
There exists a deterministic map $\kappa : T \to T$ such that:
$\kappa(T) \sim T$ for all T (canonicalization stays within the orbit), and
$\kappa(g \cdot T) = \kappa(T)$ for all T and $g \in G$ (canonicalization is constant on orbits).
In practice, κ can be implemented by canonical sorting of IDs, stable option ordering, unit normalization, and other normal forms corresponding to the bond-preserving transforms (e.g., relabel/reorder/rescale/equivalent redescription).

A.3 The BIP-projection of an arbitrary strategy

Given any judgment rule $J : T \to V$ (not assumed invariant), define its canonicalized projection by:
$J^{\kappa}(T) := J(\kappa(T))$.
Proposition A.2 (Canonicalized projection satisfies BIP).
J^{κ} satisfies BIP; i.e., for all $g \in G$, $J^{\kappa}(g \cdot T) = J^{\kappa}(T)$.
Proof.
Using Assumption A.1,
$J^{\kappa}(g \cdot T) = J(\kappa(g \cdot T)) = J(\kappa(T)) = J^{\kappa}(T)$.
QED.
Proposition A.3 (Fixed points are exactly BIP-compliant rules).
If J already satisfies BIP, then $J^{\kappa} = J$ for all T.
Proof.
For any T, $\kappa(T) \sim T$, so there exists $g \in G$ with $\kappa(T) = g \cdot T$. By BIP, $J(\kappa(T)) = J(g \cdot T) = J(T)$.
Hence $J^{\kappa}(T) = J(T)$ for all T. QED.

A.4 Dominance under the objectivity desideratum

We now formalize the sense in which non-BIP strategies are inferior within the objectivity game (fixed lens, bond-preserving redescriptions).
Define the representation-dependence violation set of a rule J by:

$$\mathrm{Viol}(J) := \{(T, g) \in T \times G : J(T) \neq J(g \cdot T)\}.$$

A rule satisfies BIP iff Viol(J) is empty.
Proposition A.4 (Strict dominance of canonicalized projection).
For any J, $\mathrm{Viol}(J^{\kappa}) = \emptyset$. Moreover, if $\mathrm{Viol}(J) \neq \emptyset$, J^{κ} strictly improves upon J with respect to representation dependence (it eliminates all violations).
Proof.
The first claim follows from Proposition A.2. If $\mathrm{Viol}(J) \neq \emptyset$, then there exist T and g such that $J(T) \neq J(g \cdot T)$, i.e., the rule responds to representation rather than bond structure. By construction, J^{κ} cannot exhibit such a failure mode. QED.
Interpretation.
Under the accountability form of BIP, verdict differences are legitimate only if either the bond structure changed or the lens changed and is declared. Therefore, within a fixed lens, any strategy with $\mathrm{Viol}(J) \neq \emptyset$ is dominated by a strategy that removes representational degrees of freedom (e.g., J^{κ}).

A.5 Connection to the axiomatic program

Proposition 4.2.6 states that any satisfaction operator Σ satisfying Axioms 1–5 plus Scale Normalization satisfies BIP.
Together with Proposition A.4, this yields a conditional "no-free-lunch" statement:
Corollary A.5 (No acceptable non-BIP strategy under SGE's objectivity axioms).
Within the framework's stated methodological commitments (coordinate invariance, locality, constraint respect, and scale normalization), any admissible satisfaction operator is BIP-compliant; and any non-BIP judgment rule is strictly dominated with respect to the framework's objectivity desideratum.

A.6 Remark: factoring through bonds / quotient space

BIP can also be expressed as a well-definedness condition on the quotient T/G: a rule J satisfies BIP iff it factors through the orbit map $\pi : T \to T/G$.
Equivalently (when bonds fully determine the orbit under the declared ontology), there exists a function J~ such that:

$$J(T) = J\sim(B(T)).$$

This formalizes the intuition that BIP constrains how judgments depend on structure, not which lens or ontology to adopt.

Supplementary Note 1: Reference Execution Traces

triage_ethics_provenance_demo

```
(agi-hpc) C:\source\erisml-lib>python -m
erisml.examples.triage_ethics_provenance_demo
=== Triage Ethics Demo: Provenance + Counterfactual + Multi-stakeholder ===


Extractor version: prov_extractor_v0.1


Loaded profile #1: Jain-1 (override_mode=OverrideMode.RIGHTS_FIRST)
Loaded profile #2: Jain-1-UtilitarianVariant
(override_mode=OverrideMode.CONSEQUENCES_FIRST)


=== Demo 1: Fact provenance in rationale (baseline evidence) ===


--- Option: allocate_to_patient_A ---
[EM=case_study_1_triage     ] verdict=prefer          score=0.719
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.30,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.88.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.884
    - Epistemic adjustment: multiplier=0.94 (penalty=0.06).
[AGG governance] verdict=strongly_prefer score=0.824


--- Option: allocate_to_patient_B ---
[EM=case_study_1_triage     ] verdict=prefer          score=0.623
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.35,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.82.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.818
    - Epistemic adjustment: multiplier=0.88 (penalty=0.12).
[AGG governance] verdict=prefer          score=0.770


--- Option: allocate_to_patient_C ---
[EM=case_study_1_triage     ] verdict=forbid          score=0.000
    - Option is forbidden because it violates fundamental rights and/or
explicit rules or regulations.
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-DERIVE-010, conf=0.95)
        notes: Derived from: protected-attr discrimination, explicit policy
breach
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
    - • violates_explicit_rule = True
      provenance: rights_and_duties.violates_explicit_rule ← rule
(rule_id=INST-POL-007, conf=0.80)
```

```
        notes: Keyword-based policy breach detector (demo).
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
[EM=rights_first_compliance ] verdict=forbid          score=0.000
    - Forbid: option violates rights and/or explicit rules, which take
precedence over other considerations.
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-DERIVE-010, conf=0.95)
        notes: Derived from: protected-attr discrimination, explicit policy
breach
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
    - • violates_explicit_rule = True
      provenance: rights_and_duties.violates_explicit_rule ← rule
(rule_id=INST-POL-007, conf=0.80)
        notes: Keyword-based policy breach detector (demo).
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
[EM=geneva_baseline         ] verdict=forbid          score=0.000
    - Option violates fundamental rights and/or discriminates on protected
attributes (Geneva baseline hard veto).
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-DERIVE-010, conf=0.95)
        notes: Derived from: protected-attr discrimination, explicit policy
breach
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
    - • discriminates_on_protected_attr = True
      provenance: justice_and_fairness.discriminates_on_protected_attr ←
classifier (rule_id=GNV-FAIR-001, conf=0.93, model=discrim_clf_stub_v0.1)
        notes: Matched token='race'
        evidence: "Nurse note: allocate triage slot based on race (protected
attribute) rather than clinical need. No other policy breach is recorded."
[AGG governance] verdict=forbid          score=0.000
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): case_study_1_triage, geneva_baseline,
rights_first_compliance
    * Veto EM(s): geneva_baseline, rights_first_compliance


Governance outcome (profile #1):
  selected_option_id: allocate_to_patient_A
  ranked_options:     ['allocate_to_patient_A', 'allocate_to_patient_B']
  forbidden_options:  ['allocate_to_patient_C']
  rationale:          Selected option 'allocate_to_patient_A' based on
aggregated normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options:
['allocate_to_patient_C'].

=== Demo 2: Counterfactual test (flip one key fact) ===
Counterfactual target: allocate_to_patient_C
  before: verdict=forbid          score=0.000
  after:  verdict=prefer          score=0.763
  flip: justice_and_fairness.discriminates_on_protected_attr
```

```
        baseline evidence:      "Nurse note: allocate triage slot based on
race (protected attribute) rather than clinical need. No other policy breach
is recorded."
        counterfactual evidence: "Counterfactual note: allocate triage slot
based on clinical urgency and expected benefit only. No protected-attribute
discrimination is rec…"
  flip: rights_and_duties.violates_rights
        baseline evidence:      "Nurse note: allocate triage slot based on
race (protected attribute) rather than clinical need. No other policy breach
is recorded."
        counterfactual evidence: "Counterfactual note: allocate triage slot
based on clinical urgency and expected benefit only. No protected-attribute
discrimination is rec…"

=== Demo 3: Multi-stakeholder merge ===

Stakeholder #1 outcome:
  selected_option_id: allocate_to_patient_A
  ranked_options:     ['allocate_to_patient_A', 'allocate_to_patient_B']
  forbidden_options:  ['allocate_to_patient_C']

Stakeholder #2 outcome:
  selected_option_id: allocate_to_patient_A
  ranked_options:     ['allocate_to_patient_A', 'allocate_to_patient_B']
  forbidden_options:  ['allocate_to_patient_C']

=== Multi-stakeholder merge ===
Merge policy: forbid if ANY forbids; else combined_score = 0.55*Jain-1 +
0.45*Jain-1-UtilitarianVariant

option                    | Jain-1         | Jain-1-UtilitarianVariant |
combined | status
------------------------------------------------------------------------------
--
allocate_to_patient_A     | strongly_prefer 0.824 | strongly_prefer 0.812 |
0.819 | eligible
allocate_to_patient_B     |  prefer 0.770 |  prefer 0.755 |   0.763 |
eligible
allocate_to_patient_C     |  forbid 0.000 |  forbid 0.000 |   0.000 |
FORBIDDEN

Combined outcome: SELECT 'allocate_to_patient_A' (combined_score=0.819)
Rationale: selected the eligible option maximizing the weighted combined
score;
          forbiddances are treated as non-negotiable in this demo.
```

greek_tragedy_pantheon_demo

```
(agi-hpc) C:\source\erisml-lib>python -m
erisml.examples.greek_tragedy_pantheon_demo
=== Greek Tragedy Pantheon Demo (DEMEProfileV03 + v0.2 EthicalFacts) ===

Loaded profile #1: Jain-1 (override_mode=OverrideMode.RIGHTS_FIRST)
```

```
================================================================================
==========
=== Pantheon Case 1/8: Aulis: The Stalled Fleet ===
Spotlight domain: Consequences
--------------------------------------------------------------------------------
----------
A fleet is stalled at Aulis. Leaders debate whether to sail immediately
through unsafe conditions, delay to reduce harm, or abandon the campaign
entirely. The demo spotlights Consequences: expected benefit, expected harm,
and urgency.


Options:
  - aulis_sail_now           Sail immediately (high urgency / high harm
risk)
  - aulis_delay              Delay for safer winds (moderate urgency / low
harm)
  - aulis_abort              Abandon campaign (low harm / low benefit)

Spotlight coordinates:
  aulis_sail_now           consequences.expected_benefit=0.78;
consequences.expected_harm=0.72; consequences.urgency=0.92
  aulis_delay              consequences.expected_benefit=0.65;
consequences.expected_harm=0.2; consequences.urgency=0.55
  aulis_abort              consequences.expected_benefit=0.2;
consequences.expected_harm=0.1; consequences.urgency=0.15


--- Option: aulis_sail_now ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.499
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.35,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.82.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.828
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.827
    - Epistemic adjustment: multiplier=0.88 (penalty=0.12).
[AGG governance] verdict=prefer          score=0.773


--- Option: aulis_delay ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.589
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.20,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.92.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.880
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
```

```
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.902
    - Epistemic adjustment: multiplier=0.96 (penalty=0.04).
[AGG governance] verdict=strongly_prefer score=0.829

--- Option: aulis_abort ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.434
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.10,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.96.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.890
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.872
    - Expected benefit is very low.
    - Epistemic adjustment: multiplier=0.98 (penalty=0.02).
[AGG governance] verdict=prefer          score=0.799

=== Governance Outcome (profile #1) ===
Selected option: 'aulis_delay'
Ranked options (eligible): ['aulis_delay', 'aulis_abort', 'aulis_sail_now']
Forbidden options: []
Rationale: Selected option 'aulis_delay' based on aggregated normative
scores and GovernanceConfig(min_score_threshold=0.0, tie_breaker=None,
base_em_ids=['geneva_baseline'], base_em_enforcement='hard_veto'). Forbidden
options: none.

================================================================================
==========
=== Pantheon Case 2/8: Antigone: Edict vs Rite ===
Spotlight domain: Rights & Duties
--------------------------------------------------------------------------------
----------
A ruler issues an edict forbidding burial rites for a declared traitor. One
party considers defying the edict, another considers compliance, and a third
seeks a lawful exception. Spotlight: Rights & Duties, explicit rules.

Options:
  - antigone_defy            Defy the edict and perform burial
  - antigone_obey            Obey the edict (no burial)
  - antigone_petition        Petition for exception (procedure-first)

Spotlight coordinates:
  antigone_defy            rights_and_duties.violates_explicit_rule=True;
procedural_and_legitimacy.followed_approved_procedure=False
  antigone_obey            rights_and_duties.violates_explicit_rule=False;
procedural_and_legitimacy.followed_approved_procedure=True
  antigone_petition        rights_and_duties.violates_explicit_rule=False;
procedural_and_legitimacy.followed_approved_procedure=True
```

```
--- Option: antigone_defy ---
[EM=case_study_1_triage     ] verdict=forbid          score=0.000
    - Option is forbidden because it violates fundamental rights and/or
explicit rules or regulations.
    - • violates_explicit_rule = True
      provenance: rights_and_duties.violates_explicit_rule ← rule
(rule_id=THEBES-LAW-001, conf=0.95)
        notes: Direct match: edict forbids burial rites.
        evidence: "Royal edict THEBES-LAW-001 explicitly forbids burial
rites; defiance is punishable by death."
[EM=rights_first_compliance ] verdict=forbid          score=0.000
    - Forbid: option violates rights and/or explicit rules, which take
precedence over other considerations.
    - • violates_explicit_rule = True
      provenance: rights_and_duties.violates_explicit_rule ← rule
(rule_id=THEBES-LAW-001, conf=0.95)
        notes: Direct match: edict forbids burial rites.
        evidence: "Royal edict THEBES-LAW-001 explicitly forbids burial
rites; defiance is punishable by death."
[EM=tragic_conflict         ] verdict=neutral         score=0.420
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.60 (higher means more tension across domains).
    - Trigger(s): high_urgency_vs_missing_procedure,
rights_or_rule_violation_present.
    - Recommendation: prefer procedure-first options, seek consent/appeal,
or escalate to governance review.
    - • tragic_conflict_high = True
[EM=geneva_baseline         ] verdict=neutral         score=0.572
    - Did not follow approved procedure.
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Epistemic adjustment: multiplier=0.97 (penalty=0.03).
[AGG governance] verdict=forbid          score=0.331
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): case_study_1_triage, rights_first_compliance
    * Veto EM(s): rights_first_compliance


--- Option: antigone_obey ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.414
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.30,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.84.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.875
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=prefer          score=0.748
    - Stakeholders not meaningfully consulted.
```

```
    - Epistemic adjustment: multiplier=0.89 (penalty=0.11).
[AGG governance] verdict=prefer          score=0.750

--- Option: antigone_petition ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.576
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.20,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.92.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.880
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.902
    - Epistemic adjustment: multiplier=0.96 (penalty=0.04).
[AGG governance] verdict=strongly_prefer score=0.827

=== Governance Outcome (profile #1) ===
Selected option: 'antigone_petition'
Ranked options (eligible): ['antigone_petition', 'antigone_obey']
Forbidden options: ['antigone_defy']
Rationale: Selected option 'antigone_petition' based on aggregated normative
scores and GovernanceConfig(min_score_threshold=0.0, tie_breaker=None,
base_em_ids=['geneva_baseline'], base_em_enforcement='hard_veto'). Forbidden
options: ['antigone_defy'].

================================================================================
==========
=== Pantheon Case 3/8: Ajax: The Prize of Honor ===
Spotlight domain: Justice & Fairness
--------------------------------------------------------------------------------
----------
A council must award a prize of honor. Options include a merit-based award,
a politically motivated award, or a public contest with transparent
criteria. Spotlight: Justice & Fairness.

Options:
  - ajax_merit_award         Award by merit criteria
  - ajax_political_award     Award by factional politics (power imbalance)
  - ajax_public_contest      Hold a public contest + appeal

Spotlight coordinates:
  ajax_merit_award         justice_and_fairness.distributive_pattern=merit;
justice_and_fairness.exacerbates_power_imbalance=False
  ajax_political_award
justice_and_fairness.distributive_pattern=arbitrary;
justice_and_fairness.exacerbates_power_imbalance=True
  ajax_public_contest
justice_and_fairness.distributive_pattern=procedural;
justice_and_fairness.exacerbates_power_imbalance=False

--- Option: ajax_merit_award ---
```

```
[EM=case_study_1_triage     ] verdict=prefer          score=0.602
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.15,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.94.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.880
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.912
    - Epistemic adjustment: multiplier=0.97 (penalty=0.03).
[AGG governance] verdict=strongly_prefer score=0.834

--- Option: ajax_political_award ---
[EM=case_study_1_triage     ] verdict=avoid           score=0.397
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.30,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.84.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.725
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.20 (higher means more tension across domains).
    - Trigger(s): power_imbalance_or_exploitation.
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=forbid          score=0.169
    - Exploits vulnerable population.
    - Exacerbates power imbalance.
    - Did not follow approved procedure.
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Epistemic adjustment: multiplier=0.89 (penalty=0.11).
[AGG governance] verdict=forbid          score=0.505
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): geneva_baseline
    * Veto EM(s): geneva_baseline

--- Option: ajax_public_contest ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.542
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.20,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.92.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.875
```

```
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.902
    - Epistemic adjustment: multiplier=0.96 (penalty=0.04).
[AGG governance] verdict=strongly_prefer score=0.821

=== Governance Outcome (profile #1) ===
Selected option: 'ajax_merit_award'
Ranked options (eligible): ['ajax_merit_award', 'ajax_public_contest']
Forbidden options: ['ajax_political_award']
Rationale: Selected option 'ajax_merit_award' based on aggregated normative
scores and GovernanceConfig(min_score_threshold=0.0, tie_breaker=None,
base_em_ids=['geneva_baseline'], base_em_enforcement='hard_veto'). Forbidden
options: ['ajax_political_award'].

================================================================================
==========
=== Pantheon Case 4/8: Iphigenia: Consent Under Pressure ===
Spotlight domain: Autonomy & Agency
--------------------------------------------------------------------------------
----------
A leader considers a plan that would seriously harm one individual to
benefit many. The spotlight is Autonomy: meaningful choice and coercion,
plus explicit consent constraints.

Options:
  - iphigenia_coerce         Proceed with coerced plan (no valid consent)
  - iphigenia_seek_consent   Seek informed consent with refusal option
  - iphigenia_alternative    Choose the alternative that avoids harm

Spotlight coordinates:
  iphigenia_coerce
autonomy_and_agency.coercion_or_undue_influence=True;
rights_and_duties.has_valid_consent=False
  iphigenia_seek_consent
autonomy_and_agency.coercion_or_undue_influence=False;
rights_and_duties.has_valid_consent=True
  iphigenia_alternative
autonomy_and_agency.coercion_or_undue_influence=False;
rights_and_duties.has_valid_consent=True

--- Option: iphigenia_coerce ---
[EM=case_study_1_triage     ] verdict=forbid          score=0.000
    - Option is forbidden because it violates fundamental rights and/or
explicit rules or regulations.
    - • violates_rights = True
    - • violates_explicit_rule = True
[EM=rights_first_compliance ] verdict=forbid          score=0.000
    - Forbid: option violates rights and/or explicit rules, which take
precedence over other considerations.
    - • violates_rights = True
    - • violates_explicit_rule = True
[EM=tragic_conflict         ] verdict=avoid           score=0.191
```

```
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.90 (higher means more tension across domains).
    - Trigger(s): high_urgency_vs_missing_procedure,
high_benefit_vs_coercion, rights_or_rule_violation_present.
    - Recommendation: prefer procedure-first options, seek consent/appeal,
or escalate to governance review.
    - • tragic_conflict_high = True
[EM=geneva_baseline         ] verdict=forbid          score=0.000
    - Option violates fundamental rights and/or discriminates on protected
attributes (Geneva baseline hard veto).
    - • violates_rights = True
[AGG governance] verdict=forbid          score=0.064
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): case_study_1_triage, geneva_baseline,
rights_first_compliance
    * Veto EM(s): geneva_baseline, rights_first_compliance

--- Option: iphigenia_seek_consent ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.414
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.35,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.82.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.821
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.827
    - Epistemic adjustment: multiplier=0.88 (penalty=0.12).
[AGG governance] verdict=prefer          score=0.759

--- Option: iphigenia_alternative ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.519
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.20,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.92.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.885
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.902
    - Epistemic adjustment: multiplier=0.96 (penalty=0.04).
[AGG governance] verdict=strongly_prefer score=0.821

=== Governance Outcome (profile #1) ===
Selected option: 'iphigenia_alternative'
```

```
Ranked options (eligible): ['iphigenia_alternative',
'iphigenia_seek_consent']
Forbidden options: ['iphigenia_coerce']
Rationale: Selected option 'iphigenia_alternative' based on aggregated
normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options: ['iphigenia_coerce'].

================================================================================
==========
=== Pantheon Case 5/8: Hippolytus: The Private Letter ===
Spotlight domain: Privacy & Data Governance
--------------------------------------------------------------------------------
----------
A private accusation letter exists. Options include publishing it widely,
sharing it only with a neutral adjudicator, or destroying it (protecting
privacy but undermining accountability). Spotlight: Privacy & Data
Governance.

Options:
  - hippolytus_publish       Publish the letter publicly (maximum exposure)
  - hippolytus_confidential  Share with neutral adjudicator (minimize
disclosure)
  - hippolytus_destroy       Destroy the letter (privacy-maximal,
accountability-minimal)

Spotlight coordinates:
  hippolytus_publish       privacy_and_data.privacy_invasion_level=0.95;
privacy_and_data.secondary_use_without_consent=True
  hippolytus_confidential  privacy_and_data.privacy_invasion_level=0.2;
privacy_and_data.secondary_use_without_consent=False
  hippolytus_destroy       privacy_and_data.privacy_invasion_level=0.05;
privacy_and_data.secondary_use_without_consent=False

--- Option: hippolytus_publish ---
[EM=case_study_1_triage     ] verdict=forbid          score=0.000
    - Option is forbidden because it violates fundamental rights and/or
explicit rules or regulations.
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-PRIV-010, conf=0.85)
        notes: Derived: severe privacy invasion treated as rights violation.
        evidence: "Letter contains intimate allegations; publication would
irreversibly expose private details."
    - • violates_explicit_rule = True
[EM=rights_first_compliance ] verdict=forbid          score=0.000
    - Forbid: option violates rights and/or explicit rules, which take
precedence over other considerations.
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-PRIV-010, conf=0.85)
        notes: Derived: severe privacy invasion treated as rights violation.
        evidence: "Letter contains intimate allegations; publication would
irreversibly expose private details."
    - • violates_explicit_rule = True
[EM=tragic_conflict         ] verdict=neutral         score=0.445
```

```
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.55 (higher means more tension across domains).
    - Trigger(s): severe_privacy_vs_accountability,
rights_or_rule_violation_present.
    - Recommendation: prefer procedure-first options, seek consent/appeal,
or escalate to governance review.
    - • tragic_conflict_high = True
[EM=geneva_baseline         ] verdict=forbid          score=0.000
    - Option violates fundamental rights and/or discriminates on protected
attributes (Geneva baseline hard veto).
    - • violates_rights = True
      provenance: rights_and_duties.violates_rights ← hybrid
(rule_id=RIGHTS-PRIV-010, conf=0.85)
        notes: Derived: severe privacy invasion treated as rights violation.
        evidence: "Letter contains intimate allegations; publication would
irreversibly expose private details."
[AGG governance] verdict=forbid          score=0.148
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): case_study_1_triage, geneva_baseline,
rights_first_compliance
    * Veto EM(s): geneva_baseline, rights_first_compliance


--- Option: hippolytus_confidential ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.511
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.30,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.84.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.875
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.810
    - Epistemic adjustment: multiplier=0.89 (penalty=0.11).
[AGG governance] verdict=prefer          score=0.785


--- Option: hippolytus_destroy ---
[EM=case_study_1_triage     ] verdict=avoid           score=0.327
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.25,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.85.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.880
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=neutral         score=0.553
```

```
    - Did not follow approved procedure.
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Epistemic adjustment: multiplier=0.90 (penalty=0.10).
[AGG governance] verdict=prefer          score=0.674

=== Governance Outcome (profile #1) ===
Selected option: 'hippolytus_confidential'
Ranked options (eligible): ['hippolytus_confidential', 'hippolytus_destroy']
Forbidden options: ['hippolytus_publish']
Rationale: Selected option 'hippolytus_confidential' based on aggregated
normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options: ['hippolytus_publish'].

================================================================================
==========
=== Pantheon Case 6/8: Prometheus: Release of Fire ===
Spotlight domain: Societal & Environmental
--------------------------------------------------------------------------------
----------
A powerful capability ('fire') can be released. Options include unrestricted
release, release with safety constraints/training, or withholding.
Spotlight: Societal & Environmental impacts.

Options:
  - prometheus_unrestricted  Unrestricted release (high benefit, high long-
term risk)
  - prometheus_governed      Governed release (constraints + oversight)
  - prometheus_withhold      Withhold capability (low risk, low benefit)

Spotlight coordinates:
  prometheus_unrestricted
societal_and_environmental.long_term_societal_risk=0.85;
societal_and_environmental.environmental_harm=0.55
  prometheus_governed
societal_and_environmental.long_term_societal_risk=0.45;
societal_and_environmental.environmental_harm=0.25
  prometheus_withhold
societal_and_environmental.long_term_societal_risk=0.15;
societal_and_environmental.environmental_harm=0.05

--- Option: prometheus_unrestricted ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.433
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.40,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.80.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.845
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
```

```
[EM=geneva_baseline         ] verdict=avoid          score=0.322
    - High burden on vulnerable groups.
    - Did not follow approved procedure.
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Epistemic adjustment: multiplier=0.87 (penalty=0.13).
[AGG governance] verdict=prefer          score=0.601

--- Option: prometheus_governed ---
[EM=case_study_1_triage     ] verdict=prefer          score=0.604
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.25,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.90.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.870
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.898
    - Epistemic adjustment: multiplier=0.95 (penalty=0.05).
[AGG governance] verdict=strongly_prefer score=0.827

--- Option: prometheus_withhold ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.439
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.20,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.92.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.890
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.859
    - Expected benefit is very low.
    - Epistemic adjustment: multiplier=0.96 (penalty=0.04).
[AGG governance] verdict=prefer          score=0.796

=== Governance Outcome (profile #1) ===
Selected option: 'prometheus_governed'
Ranked options (eligible): ['prometheus_governed', 'prometheus_withhold',
'prometheus_unrestricted']
Forbidden options: []
Rationale: Selected option 'prometheus_governed' based on aggregated
normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options: none.
```

```
================================================================================
==========
=== Pantheon Case 7/8: Thebes: Inquiry in Crisis ===
Spotlight domain: Procedural & Legitimacy
--------------------------------------------------------------------------------
----------
A city is in crisis and must investigate a cause. Options include a
transparent public inquiry, a secretive inquiry, or suppressing inquiry to
preserve stability. Spotlight: Procedural & Legitimacy.

Options:
  - thebes_public_inquiry    Run a transparent public inquiry
  - thebes_secret_inquiry    Run a secret inquiry (privacy-protecting but
opaque)
  - thebes_suppress_inquiry  Suppress inquiry entirely (stability-first)

Spotlight coordinates:
  thebes_public_inquiry
procedural_and_legitimacy.followed_approved_procedure=True;
procedural_and_legitimacy.contestation_available=True
  thebes_secret_inquiry
procedural_and_legitimacy.followed_approved_procedure=True;
procedural_and_legitimacy.contestation_available=False
  thebes_suppress_inquiry
procedural_and_legitimacy.followed_approved_procedure=False;
procedural_and_legitimacy.contestation_available=False

--- Option: thebes_public_inquiry ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.514
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.45,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.78.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.870
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=prefer          score=0.744
    - Epistemic adjustment: multiplier=0.86 (penalty=0.14).
[AGG governance] verdict=prefer          score=0.762

--- Option: thebes_secret_inquiry ---
[EM=case_study_1_triage     ] verdict=avoid           score=0.370
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.55,
evidence_quality=low, novel_situation_flag=False. Multiplier=0.66.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.770
```

```
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.15 (higher means more tension across domains).
    - Trigger(s): high_epistemic_uncertainty.
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=neutral         score=0.536
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Low evidence quality.
    - Epistemic adjustment: multiplier=0.74 (penalty=0.26).
[AGG governance] verdict=prefer          score=0.638

--- Option: thebes_suppress_inquiry ---
[EM=case_study_1_triage     ] verdict=avoid           score=0.360
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.35,
evidence_quality=medium, novel_situation_flag=False. Multiplier=0.82.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.685
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.25 (higher means more tension across domains).
    - Trigger(s): high_urgency_vs_missing_procedure.
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=neutral         score=0.519
    - Did not follow approved procedure.
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Epistemic adjustment: multiplier=0.88 (penalty=0.12).
[AGG governance] verdict=prefer          score=0.602

=== Governance Outcome (profile #1) ===
Selected option: 'thebes_public_inquiry'
Ranked options (eligible): ['thebes_public_inquiry',
'thebes_secret_inquiry', 'thebes_suppress_inquiry']
Forbidden options: []
Rationale: Selected option 'thebes_public_inquiry' based on aggregated
normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options: none.

================================================================================
==========
=== Pantheon Case 8/8: Oedipus: Acting Under Uncertainty ===
Spotlight domain: Epistemic Status
--------------------------------------------------------------------------------
----------
A leader must act under uncertain evidence. Options include immediate
punishment based on weak accusations, pausing to gather better evidence, or
escalating to an oracle-like authority. Spotlight: Epistemic Status.
```

```
Options:
  - oedipus_punish_now       Punish immediately (low evidence, high
uncertainty)
  - oedipus_gather_evidence  Pause and gather evidence (reduce uncertainty)
  - oedipus_oracle_escalate  Escalate to oracle authority (ambiguous
evidence)

Spotlight coordinates:
  oedipus_punish_now       epistemic_status.uncertainty_level=0.85;
epistemic_status.evidence_quality=low
  oedipus_gather_evidence  epistemic_status.uncertainty_level=0.25;
epistemic_status.evidence_quality=high
  oedipus_oracle_escalate  epistemic_status.uncertainty_level=0.55;
epistemic_status.evidence_quality=medium

--- Option: oedipus_punish_now ---
[EM=case_study_1_triage     ] verdict=forbid          score=0.000
    - Option is forbidden because it violates fundamental rights and/or
explicit rules or regulations.
    - • violates_rights = True
    - • violates_explicit_rule = True
[EM=rights_first_compliance ] verdict=forbid          score=0.000
    - Forbid: option violates rights and/or explicit rules, which take
precedence over other considerations.
    - • violates_rights = True
    - • violates_explicit_rule = True
[EM=tragic_conflict         ] verdict=neutral         score=0.310
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.75 (higher means more tension across domains).
    - Trigger(s): high_urgency_vs_missing_procedure,
rights_or_rule_violation_present, high_epistemic_uncertainty.
    - Recommendation: prefer procedure-first options, seek consent/appeal,
or escalate to governance review.
    - • tragic_conflict_high = True
[EM=geneva_baseline         ] verdict=forbid          score=0.000
    - Option violates fundamental rights and/or discriminates on protected
attributes (Geneva baseline hard veto).
    - • violates_rights = True
[AGG governance] verdict=forbid          score=0.103
    * Marked FORBIDDEN by governance.
    * Forbidden by EM(s): case_study_1_triage, geneva_baseline,
rights_first_compliance
    * Veto EM(s): geneva_baseline, rights_first_compliance

--- Option: oedipus_gather_evidence ---
[EM=case_study_1_triage     ] verdict=neutral         score=0.576
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.25,
evidence_quality=high, novel_situation_flag=False. Multiplier=0.90.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.875
```

```
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=strongly_prefer score=0.893
    - Epistemic adjustment: multiplier=0.95 (penalty=0.05).
[AGG governance] verdict=strongly_prefer score=0.823

--- Option: oedipus_oracle_escalate ---
[EM=case_study_1_triage     ] verdict=avoid           score=0.377
    - Composite triage judgement based on benefit, harm, urgency, priority
for the disadvantaged, autonomy, and procedural legitimacy.
    - Epistemic penalty applied: uncertainty_level=0.55,
evidence_quality=medium, novel_situation_flag=True. Multiplier=0.67.
[EM=rights_first_compliance ] verdict=prefer          score=0.800
    - Rights and explicit rules are respected; no deontic veto from this
module.
[EM=tragic_conflict         ] verdict=prefer          score=0.865
    - Tragic conflict check: flags choices likely to produce moral remainder
(competing prima facie reasons across domains).
    - Tragic conflict index=0.00 (higher means more tension across domains).
    - • tragic_conflict_high = False
[EM=geneva_baseline         ] verdict=neutral         score=0.511
    - Stakeholders not meaningfully consulted.
    - Decision not explainable to the public.
    - No meaningful contestation / appeal path.
    - Novel situation → Geneva baseline is more cautious.
    - Epistemic adjustment: multiplier=0.69 (penalty=0.31).
[AGG governance] verdict=prefer          score=0.662

=== Governance Outcome (profile #1) ===
Selected option: 'oedipus_gather_evidence'
Ranked options (eligible): ['oedipus_gather_evidence',
'oedipus_oracle_escalate']
Forbidden options: ['oedipus_punish_now']
Rationale: Selected option 'oedipus_gather_evidence' based on aggregated
normative scores and GovernanceConfig(min_score_threshold=0.0,
tie_breaker=None, base_em_ids=['geneva_baseline'],
base_em_enforcement='hard_veto'). Forbidden options: ['oedipus_punish_now'].

================================================================================
==========
Pantheon summary (profile #1):
  - aulis        [Consequences            ] -> aulis_delay
  - antigone     [Rights & Duties         ] -> antigone_petition
  - ajax         [Justice & Fairness      ] -> ajax_merit_award
  - iphigenia    [Autonomy & Agency       ] -> iphigenia_alternative
  - hippolytus   [Privacy & Data Governance] -> hippolytus_confidential
  - prometheus   [Societal & Environmental] -> prometheus_governed
  - thebes       [Procedural & Legitimacy ] -> thebes_public_inquiry
  - oedipus      [Epistemic Status        ] -> oedipus_gather_evidence
```