Stratified Geometric Ethics: A Mathematical Foundation for Real-Time Moral Reasoning in Autonomous Systems

[Author Names Redacted for Review]
[Institution Redacted]
`[email redacted]`

Abstract

We present *Stratified Geometric Ethics* (SGE), a mathematical framework for implementing deterministic, verifiable ethical reasoning in autonomous AI systems. Unlike prior geometric approaches to ethics that assume smooth manifold structures, SGE models the space of ethically relevant configurations as a *stratified space*—a union of smooth manifolds of varying dimensions connected along boundary strata—thereby capturing moral discontinuities, incommensurable values, and genuine ethical dilemmas that smooth models cannot represent.

We make four principal contributions: (1) We introduce *grounded obligation bundles* and *preference sheaves* that solve the content problem by providing constructive methods for deriving ethical tensors from constitutional principles, learned stakeholder preferences, and deliberative procedures. (2) We generalize the satisfaction functional beyond bilinearity to accommodate threshold effects, lexical priorities, and diminishing returns through *stratified contraction operators*. (3) We prove computational complexity bounds for ethical evaluation and geodesic planning, establishing conditions under which real-time guarantees are achievable. (4) We develop a formal verification framework that enables machine-checkable proofs of ethical constraint satisfaction.

We prove representation theorems characterizing the class of admissible satisfaction functionals on stratified spaces, establish existence and uniqueness results for ethical geodesics under appropriate regularity conditions, and derive sample complexity bounds for learning obligation and preference fields from finite data. The framework is instantiated in a medical triage scenario, demonstrating how SGE provides formally verified ethical guarantees while maintaining computational tractability. **Keywords:** AI ethics, moral reasoning, differential geometry, stratified spaces, formal verification, autonomous systems

1 Introduction

The deployment of AI systems in high-stakes domains—healthcare, autonomous vehicles, criminal justice, financial markets—creates an urgent need for frameworks that make ethical reasoning explicit, deterministic, and verifiable. Current approaches to AI ethics fall into two broad categories: *principle-based* approaches that articulate high-level commitments (fairness, transparency, accountability) without specifying computational implementations, and *learning-based* approaches that attempt to infer ethical behavior from human feedback without formal guarantees.

Neither approach satisfies the requirements of safety-critical deployment. Principle-based frameworks lack the mathematical precision needed for formal verification. Learning-based approaches cannot provide the deterministic guarantees required when lives are at stake. What is needed is a framework that combines the rigor of mathematical foundations with the flexibility to handle the genuine complexity of moral reasoning.

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

1. **The Content Problem:** They assume obligations and interests are "given" without specifying how these tensors are derived from ethical sources.

2. **The Smoothness Problem:** Real ethics involves discrete choices, incommensurable values, and moral discontinuities that smooth manifolds cannot represent.

3. **The Linearity Problem:** Bilinear satisfaction functionals cannot capture threshold effects, lexical priorities, or diminishing returns.

4. **The Tractability Problem:** No complexity analysis establishes whether ethical computation can be performed in real time.

5. **The Verification Problem:** No formal 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 mathematical elegance of the geometric approach.

1.1 Contributions

Our principal contributions are:

1. **Stratified Moral Spaces (Section 3):** We model the space of ethically relevant configurations as a stratified space $\mathcal{M} = \bigcup_i \mathcal{M}_i$, where each stratum $\mathcal{M}_i$ is a smooth manifold and strata are connected along boundaries representing moral discontinuities.

2. **Grounded Ethical Content (Section 4):** We introduce *obligation bundles* derived from constitutional principles and *preference sheaves* learned from stakeholder behavior, providing constructive methods for populating the ethical tensors.

3. **Generalized Satisfaction Operators (Section 5):** We define *stratified contraction operators* that extend beyond bilinearity to accommodate nonlinear moral phenomena, and prove a representation theorem characterizing all admissible operators.

4. **Computational Complexity Analysis (Section 7):** We establish that ethical evaluation is $O(n^2)$ in the dimension of the moral space, and that geodesic planning is polynomial-time solvable under specified regularity conditions.

5. **Formal Verification Framework (Section 8):** We develop a specification language for ethical constraints and prove that constraint satisfaction is decidable for a useful fragment of the language.

6. **Learning-Theoretic Foundations (Section 9):** We derive PAC bounds for learning obligation and preference fields, establishing sample complexity for ethical content acquisition.

1.2 Paper Organization

Section 2 reviews related work. Section 3 introduces stratified moral spaces. Section 4 addresses the content problem. Section 5 develops generalized satisfaction operators. Section 6 analyzes ethical geodesics. Section 7 establishes computational bounds. Section 8 presents the verification framework. Section 9 provides learning-theoretic foundations. Section 10 instantiates the framework in a triage scenario. Section 11 discusses limitations and future work.

2 Related Work

2.1 Geometric Approaches to Ethics

The application of differential geometry to ethics has roots in the work of [1], who proposed modeling ethical configurations as points on a “moral manifold” with stakeholder perspectives as coordinate charts. Their DEME framework defines obligations as contravariant vector fields and interests as covector fields, with ethical satisfaction computed as their tensor contraction.

While this approach provides coordinate invariance—a valuable property ensuring that ethical evaluations do not depend on arbitrary representational choices—it assumes (1) the moral space is a smooth manifold, (2) satisfaction is bilinear, and (3) obligations and interests are given *a priori*. Our work relaxes all three assumptions.

2.2 Formal Methods in AI Ethics

The application of formal methods to AI ethics has gained attention following high-profile failures of deployed systems. [2] proposed using model checking to verify ethical properties of autonomous systems, while [3] developed logical frameworks for encoding ethical constraints. These approaches typically assume a finite state space and do not address the geometric structure of ethical reasoning.

Our verification framework builds on this tradition while accommodating the continuous, geometric nature of ethical spaces through the theory of o-minimal structures [10].

2.3 Machine Learning and Ethics

Constitutional AI [4] trains language models to follow high-level principles through reinforcement learning from AI feedback. Inverse reinforcement learning [5] infers reward functions from observed behavior. Preference learning [6] models human value judgments.

Our framework integrates these approaches by using constitutional principles to define the *structure* of obligation bundles while using learned preferences to populate their *content*. This hybrid approach combines the interpretability of principle-based methods with the flexibility of learning-based methods.

2.4 Stratified Spaces in Mathematics

Stratified spaces generalize smooth manifolds by allowing singularities and dimension changes. The theory was developed by Whitney [7] and Thom [8] and has found applications in algebraic geometry, singularity theory, and more recently in topological data analysis [9].

To our knowledge, this paper is the first to apply stratified space theory to ethics and AI alignment.

3 Stratified Moral Spaces

We begin by constructing the geometric foundation of SGE.

3.1 Motivation: Why Manifolds Are Insufficient

Consider a trolley problem variant: an autonomous vehicle must choose between two trajectories, one harming pedestrian A and one harming pedestrian B. No smooth interpolation between these options exists—the choice is discrete. Moreover, the "moral residue" of the choice (the sense that something of value was sacrificed regardless of which option was selected) cannot be captured by a smooth satisfaction function.

More generally, ethical reasoning involves:

- **Discrete choices:** Many decisions present finitely many options with no meaningful intermediate.
- **Incommensurable values:** Some moral dimensions cannot be traded off against others at any rate.
- **Threshold effects:** Crossing certain boundaries (killing vs. letting die; lying vs. withholding truth) has discontinuous moral significance.
- **Genuine dilemmas:** Some situations admit no ethically satisfactory resolution.

Stratified spaces provide the mathematical structure to model all of these phenomena.

3.2 Definition of Stratified Moral Spaces

Definition 3.1 (Stratified Moral Space)**.** *A* stratified moral space *is a triple* $(\mathcal{M}, \{\mathcal{M}_i\}_{i\in I}, \preceq)$ *where:*

1. $\mathcal{M}$ *is a paracompact Hausdorff topological space (the* total moral space*).*
2. $\{\mathcal{M}_i\}_{i\in I}$ *is a locally finite partition of* $\mathcal{M}$ *into connected smooth manifolds (the* strata*).*
3. $\preceq$ *is a partial order on* I *such that* $\mathcal{M}_i \cap \overline{\mathcal{M}_j} \neq \emptyset$ *implies* $i \preceq j$ *(the* frontier condition*).*

We require that each stratum $\mathcal{M}_i$ *is locally closed and that the stratification satisfies Whitney's condition (B) for regularity.*

Definition 3.2 (Moral Dimension)**.** *The* moral dimension *of a point* $x \in \mathcal{M}$ *is* $\dim(x) := \dim(\mathcal{M}_i)$ *where* $x \in \mathcal{M}_i$*. Points on lower-dimensional strata represent morally "special" configurations—dilemmas, phase transitions, or singularities.*

Definition 3.3 (Smooth Stratum)**.** *The* regular *or* smooth stratum $\mathcal{M}_{reg}$ *is the union of all maximal-dimensional strata. Points in* $\mathcal{M}_{reg}$ *admit smooth trade-offs between moral dimensions; points in* $\mathcal{M} \setminus \mathcal{M}_{reg}$ *represent configurations where smooth trade-offs break down.*

3.3 Examples of Stratified Moral Spaces

Example 3.4 (Trolley Problem)**.** *Let* $\mathcal{M} = \{A, B, \gamma\}$ *where A and B are 0-dimensional strata representing the discrete choices, and* $\gamma \cong (0,1)$ *is a 1-dimensional stratum representing the "deliberation space" during which neither choice has been made. The frontier condition gives* $A \preceq \gamma$ *and* $B \preceq \gamma$*, indicating that deliberation "collapses" to a discrete choice.*

Example 3.5 (Resource Allocation)**.** *Consider allocating a resource among* n *recipients. Let* $\mathcal{M} = \Delta^{n-1}$ *be the* $(n-1)$*-simplex. The interior is the regular stratum where all recipients receive positive shares. Lower-dimensional faces represent allocations where some recipients receive nothing. Vertices represent all-or-nothing allocations. The stratification captures that excluding someone from allocation is categorically different from giving them a small share.*

Example 3.6 (Lexical Priority)**.** *Suppose value* V_1 *has lexical priority over* V_2*. Model this as* $\mathcal{M} = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ *with strata:*

- $\mathcal{M}_0 = \{0\} \times \{0\}$ *(both values absent)*

- $\mathcal{M}_1 = \{0\} \times \mathbb{R}_{>0}$ *(only V_2 present)*
- $\mathcal{M}_2 = \mathbb{R}_{>0} \times \{0\}$ *(only V_1 present)*
- $\mathcal{M}_3 = \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ *(both present)*

The lexical priority is encoded by defining the metric on $\mathcal{M}_3$ such that movement in the V_1 direction has infinitely greater cost than movement in V_2.

3.4 Tangent and Cotangent Structures

On a stratified space, tangent and cotangent spaces must be defined carefully.

Definition 3.7 (Stratified Tangent Space)**.** *At a point $x \in \mathcal{M}_i$, the* stratified tangent space *is*

$$T_x^{str}\mathcal{M} := T_x\mathcal{M}_i \oplus N_x(\mathcal{M}_i, \mathcal{M})$$

where $T_x\mathcal{M}_i$ is the ordinary tangent space to the stratum and $N_x(\mathcal{M}_i, \mathcal{M})$ is the normal cone *capturing directions transverse to the stratum.*

For points on the regular stratum, $T_x^{\text{str}}\mathcal{M} = T_x\mathcal{M}_{\text{reg}}$ is the usual tangent space. For singular points, the normal cone encodes the "directions toward singularity resolution."

The cotangent space $T_x^{*,\text{str}}\mathcal{M}$ is defined dually. Obligations and interests will be sections of the stratified tangent and cotangent bundles, respectively.

4 Grounded Ethical Content

The central challenge for any geometric ethics framework is the *content problem*: where do the obligation and interest tensors come from? We address this through two complementary constructions.

4.1 Obligation Bundles from Constitutional Principles

Definition 4.1 (Constitutional Principle)**.** *A* constitutional principle *is a tuple $\pi = (\phi, d, w)$ where:*

- *$\phi : \mathcal{M} \to \{0, 1\}$ is a predicate identifying configurations where the principle applies.*
- *$d : supp(\phi) \to T^{str}\mathcal{M}$ assigns a "direction of improvement" to each applicable configuration.*
- *$w : supp(\phi) \to \mathbb{R}_{>0}$ assigns a weight or strength to the principle at each point.*

Example 4.2. *The principle "minimize harm" becomes:*

- *$\phi(x) = 1$ if configuration x involves potential harm.*
- *$d(x) = -\nabla H(x)$ where $H : \mathcal{M} \to \mathbb{R}$ is a harm function.*
- *$w(x)$ reflects the severity of potential harm.*

Definition 4.3 (Obligation Bundle)**.** *Given a set of constitutional principles $\Pi = \{\pi_1, \ldots, \pi_k\}$, the* obligation bundle *is the vector bundle $\mathcal{E} \to \mathcal{M}$ with fiber*

$$\mathcal{E}_x := \mathit{span}\{w_i(x) \cdot d_i(x) : \phi_i(x) = 1\}$$

The aggregate obligation field *is the section*

$$O(x) := \sum_{i:\phi_i(x)=1} w_i(x) \cdot d_i(x) \in \mathcal{E}_x \subseteq T_x^{str}\mathcal{M}$$

This construction grounds obligations in explicit principles while allowing the principles to have varying relevance across the moral space.

4.2 Preference Sheaves from Stakeholder Learning

While obligations derive from principles, interests derive from stakeholders. We model this using sheaf theory.

Definition 4.4 (Stakeholder)**.** *A* stakeholder *is an agent s with a preference relation $\succeq_s$ over configurations. We assume $\succeq_s$ can be represented (possibly approximately) by a utility function $u_s : \mathcal{M} \to \mathbb{R}$.*

Definition 4.5 (Preference Sheaf)**.** *The* preference sheaf *$\mathcal{F}$ over $\mathcal{M}$ assigns to each open set $U \subseteq \mathcal{M}$ the set of stakeholder preference data:*

$$\mathcal{F}(U) := \{(s, u_s|_U) : s \text{ is a stakeholder}\}$$

with restriction maps given by function restriction.

Definition 4.6 (Interest Covector Field)**.** *Given a stakeholder s with utility u_s, the* interest covector field *is*

$$I_s := du_s \in \Gamma(T^{*,str}\mathcal{M})$$

For multiple stakeholders, we aggregate:

$$I := \sum_s \alpha_s \cdot I_s$$

where $\alpha_s > 0$ are stakeholder weights determined by governance policy.

4.3 Learning Stakeholder Preferences

In practice, u_s must be learned from finite data. We model this as follows.

Definition 4.7 (Preference Oracle)**.** *A* preference oracle *for stakeholder s is a function $\mathcal{O}_s$ that, given two configurations $x, y \in \mathcal{M}$, returns:*

- $+1$ *if $x \succ_s y$ (strict preference for x)*
- -1 *if $y \succ_s x$ (strict preference for y)*
- 0 *if $x \sim_s y$ (indifference)*

possibly with noise.

Definition 4.8 (Preference Learning Problem)**.** *Given access to a preference oracle $\mathcal{O}_s$ and a hypothesis class $\mathcal{H}$ of utility functions, find $\hat{u}_s \in \mathcal{H}$ that best explains the oracle responses.*

We defer the analysis of sample complexity to Section 9.

4.4 Deliberative Aggregation

When stakeholder interests conflict, we require a principled aggregation mechanism.

Definition 4.9 (Deliberative Procedure)**.** *A* deliberative procedure *is a function*

$$\Delta : \mathcal{F}(\mathcal{M}) \to \Gamma(T^{*,str}\mathcal{M})$$

mapping the preference sheaf to an aggregate interest field, satisfying:

1. ***Pareto:*** *If all stakeholders prefer x to y, then $I(x) \cdot v > I(y) \cdot v$ for any direction v from y to x.*
2. ***Independence:*** *The aggregate interest at x depends only on local preference data.*
3. ***Anonymity:*** *Permuting stakeholder labels does not change the aggregate.*

Proposition 4.10. *Under mild regularity conditions, any deliberative procedure can be written as*

$$I = \sum_s \alpha_s(x, \mathcal{F}) \cdot I_s$$

for some weight functions $\alpha_s : \mathcal{M} \times \mathcal{F}(\mathcal{M}) \to \mathbb{R}_{\geq 0}$.

This result shows that deliberative procedures reduce to weighted aggregation, with the weights potentially depending on the configuration and the full preference data.

5 Stratified Satisfaction Operators

Having grounded obligations and interests, we now define how they combine into scalar satisfaction measures.

5.1 Limitations of Bilinear Contraction

The naive satisfaction functional $S(x) = O^\mu(x)I_\mu(x)$ has several limitations:

1. **No thresholds:** Satisfaction scales linearly with obligation fulfillment, but real ethics has thresholds (e.g., the difference between “mostly honest” and “completely honest” may be morally significant).
2. **No lexical priority:** Any obligation can be traded off against any other at some rate, but some values may be lexically prior.
3. **No diminishing returns:** The marginal value of the 1000th unit equals the first, but real values often exhibit diminishing marginal utility.
4. **No constraints:** Nothing prevents satisfaction from being achieved through morally impermissible means.

5.2 Generalized Satisfaction Operators

Definition 5.1 (Stratified Satisfaction Operator)**.** *A* stratified satisfaction operator *is a map*

$$\Sigma : \Gamma(T^{str}\mathcal{M}) \times \Gamma(T^{*,str}\mathcal{M}) \times \Gamma(Sym^2 T^{*,str}\mathcal{M}) \times \mathcal{C} \to C^\infty(\mathcal{M})$$

taking an obligation field O, interest field I, metric tensor g, and constraint set $\mathcal{C}$, and returning a scalar satisfaction function $S : \mathcal{M} \to \mathbb{R}$.

We impose axioms on Σ:

Axiom 1 (Coordinate Invariance)**.** *For any diffeomorphism $\psi : \mathcal{M} \to \mathcal{M}$,*

$$\Sigma(\psi_* O, \psi^* I, \psi^* g, \psi(\mathcal{C}))(x) = \Sigma(O, I, g, \mathcal{C})(\psi^{-1}(x))$$

Axiom 2 (Monotonicity)**.** *If $O'(x) = O(x) + \delta O$ where $I(x)(\delta O) > 0$ and all other arguments are equal, then $\Sigma(O', I, g, \mathcal{C})(x) \geq \Sigma(O, I, g, \mathcal{C})(x)$.*

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

Axiom 4 (Stratum Compatibility)**.** *Σ restricts to a smooth function on each stratum of $\mathcal{M}$.*

5.3 Representation Theorem

Theorem 5.2 (Representation of Stratified Satisfaction)**.** *Let Σ satisfy Axioms 1–4. Then there exist:*

- *A smooth function $f : \mathbb{R} \to \mathbb{R}$ (the* activation function*)*
- *A smooth function $\lambda : \mathcal{M} \to \mathbb{R}_{>0}$ (the* scale field*)*
- *A constraint indicator $\chi_{\mathcal{C}} : \mathcal{M} \to \{0, -\infty\}$*

such that

$$\Sigma(O, I, g, \mathcal{C})(x) = \chi_{\mathcal{C}}(x) + \lambda(x) \cdot f\left(\frac{I_\mu(x)O^\mu(x)}{\sqrt{g_{\mu\nu}(x)O^\mu(x)O^\nu(x)}}\right)$$

on the regular stratum, with appropriate limit behavior on singular strata.

Proof Sketch. Coordinate invariance restricts Σ to depend only on invariant quantities. The only coordinate-invariant scalars constructible from O, I, and g are contractions. Monotonicity forces the dependence on $I_\mu O^\mu$ to be through a monotone function. The normalization by $\sqrt{g_{\mu\nu}O^\mu O^\nu}$ ensures scale invariance. Constraint respect adds the indicator term. Full details in Appendix A. □

Corollary 5.3. *The bilinear satisfaction $S = O^\mu I_\mu$ corresponds to the special case $f = id$, $\lambda = 1$, $\mathcal{C} = \emptyset$.*

5.4 Activation Functions for Moral Phenomena

Different choices of f capture different moral structures:

- **Linear:** $f(z) = z$. No thresholds or diminishing returns.
- **Sigmoid:** $f(z) = \tanh(z)$. Bounded satisfaction with diminishing returns.
- **Threshold:** $f(z) = \mathbf{1}_{z>\tau}$. Binary satisfaction above threshold.
- **Lexical:** $f(z) = \sum_i \epsilon^i f_i(z_i)$ for small ϵ. Lexical priority through scale separation.

6 Ethical Geodesics and Trajectory Planning

Given a stratified moral space with obligations, interests, and satisfaction operator, we now address the problem of finding optimal ethical trajectories.

6.1 Stratified Metrics

Definition 6.1 (Stratified Riemannian Metric)**.** *A* stratified Riemannian metric *on $\mathcal{M}$ assigns to each stratum $\mathcal{M}_i$ a Riemannian metric g_i, with compatibility conditions on boundaries.*

Definition 6.2 (Ethical Cost Functional)**.** *Given a path $\gamma : [0, 1] \to \mathcal{M}$ connecting configurations x_0 and x_1, the* ethical cost *is*

$$\mathcal{L}[\gamma] := \int_0^1 \sqrt{g(\dot{\gamma}, \dot{\gamma})} \cdot h(\gamma(t))\, dt$$

where $h : \mathcal{M} \to \mathbb{R}_{>0}$ is a hazard function *penalizing passage through ethically costly regions.*

6.2 Existence of Geodesics

Theorem 6.3 (Existence of Ethical Geodesics)**.** *Let $(\mathcal{M}, g)$ be a complete stratified Riemannian space and h a continuous positive hazard function. For any $x_0, x_1 \in \mathcal{M}$, there exists a minimizing geodesic γ^* for $\mathcal{L}$.*

Proof Sketch. Completeness ensures the existence of minimizing sequences. The Arzelà-Ascoli theorem provides convergent subsequences. Lower semicontinuity of $\mathcal{L}$ ensures the limit is a minimizer. Whitney regularity prevents pathological behavior at strata boundaries. □

6.3 Geodesic Equations on Strata

On the interior of each stratum, the geodesic equations take the standard form with an additional hazard term:

Proposition 6.4. *On the regular stratum, minimizers of $\mathcal{L}$ satisfy*

$$\ddot{\gamma}^\mu + \Gamma^\mu_{\nu\sigma}\dot{\gamma}^\nu\dot{\gamma}^\sigma = -\frac{1}{2}g^{\mu\nu}\frac{\partial \log h}{\partial x^\nu}\|\dot{\gamma}\|^2$$

where $\Gamma^\mu_{\nu\sigma}$ are the Christoffel symbols of g.

The right-hand side represents a "force" pushing geodesics away from high-hazard regions.

6.4 Boundary Behavior

When geodesics approach stratum boundaries, special care is required.

Definition 6.5 (Stratum Crossing Cost)**.** *A* stratum crossing cost *is a function $c : \{(i, j) : \mathcal{M}_i \cap \overline{\mathcal{M}_j} \neq \emptyset\} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ assigning a cost to transitions between strata.*

Setting $c(i, j) = \infty$ for certain pairs enforces *moral constraints*: some transitions are absolutely forbidden regardless of the benefit.

Proposition 6.6. *With finite crossing costs, geodesics can be computed as shortest paths in a hybrid system combining continuous dynamics on strata with discrete transitions at boundaries.*

7 Computational Complexity

We now establish that ethical evaluation and planning are computationally tractable.

7.1 Complexity of Satisfaction Evaluation

Theorem 7.1 (Satisfaction Evaluation Complexity)**.** *Given:*

- *A configuration $x \in \mathcal{M}$ represented in n-dimensional coordinates*
- *k constitutional principles with $O(1)$-evaluable predicates and directions*
- *m stakeholders with $O(n)$-evaluable utility gradients*
- *A constraint set $\mathcal{C}$ checkable in $O(c)$ time*

The satisfaction $S(x)$ can be computed in $O(kn + mn + c)$ time.

Proof. Computing the obligation field requires evaluating k predicates and summing k vectors of dimension n: $O(kn)$. Computing the interest field requires m gradient evaluations: $O(mn)$. The contraction is $O(n)$. Constraint checking is $O(c)$. Total: $O(kn + mn + c)$. □

Corollary 7.2. *For fixed k and m, satisfaction evaluation is $O(n + c)$.*

7.2 Complexity of Geodesic Planning

Theorem 7.3 (Geodesic Planning Complexity)**.** *Let $\mathcal{M}$ be a stratified space with s strata, each of dimension at most n. Let the hazard function h be L-Lipschitz. Given start and end configurations, an ϵ-optimal geodesic can be computed in time*

$$O\left(s^2 \cdot \mathit{poly}(n) \cdot \left(\frac{L \cdot \mathit{diam}(\mathcal{M})}{\epsilon}\right)^n\right)$$

Proof Sketch. Discretize each stratum into an ϵ-net. The number of points is $O((\text{diam}/\epsilon)^n)$ per stratum. Construct a graph with edges between adjacent net points and stratum crossings. Run Dijkstra's algorithm on this graph. The s^2 factor accounts for all possible stratum transitions. □

Remark 7.4. *The exponential dependence on n is unavoidable in the worst case (geodesic planning includes shortest path problems as a special case). However, for structured problems with low effective dimension, more efficient algorithms exist.*

7.3 Real-Time Guarantees

For real-time operation, we propose a hierarchical approach:

Algorithm 1 Anytime Ethical Planning

```
Offline: Precompute coarse ethical map $\hat{S}$ on grid
Online:
while decision deadline not reached do
    Interpolate $\hat{S}$ to current configuration
    Refine local geodesic estimate
    Update best-so-far solution
end while
Return best-so-far solution with quality bound
```

Proposition 7.5. *Algorithm 1 returns a solution within factor $(1 + \epsilon)$ of optimal after $O(\log(1/\epsilon))$ refinement iterations, assuming Lipschitz continuity of the optimal cost.*

8 Formal Verification Framework

Safety-critical applications require formal guarantees. We develop a verification framework for ethical constraints.

8.1 Ethical Specification Language

Definition 8.1 (Ethical Specification Language (ESL))**.** *The syntax of ESL is:*

$$\begin{aligned} \phi ::= &\, p \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \Rightarrow \psi \\ &\mid \forall x \in R.\phi \mid \exists x \in R.\phi \\ &\mid S(x) \bowtie c \mid O_i(x) \bowtie c \mid I_j(x) \bowtie c \\ &\mid \mathit{Always}(\phi) \mid \mathit{Eventually}(\phi) \mid \phi\, \mathcal{U}\, \psi \end{aligned}$$

where p is an atomic predicate, R is a region of $\mathcal{M}$, $\bowtie \in \{<, \leq, =, \geq, >\}$, and $\mathcal{U}$ is the "until" operator from temporal logic.

Example 8.2. *"The system never takes an action with satisfaction below threshold τ":*

$$\textit{Always}(S(x) \geq \tau)$$

"If harm is possible, the system eventually reaches a safe state":

$$\textit{Harm}(x) \Rightarrow \textit{Eventually}(\neg \textit{Harm}(x))$$

8.2 Decidability Results

Definition 8.3 (O-minimal Structure)**.** *A structure* $(\mathbb{R}, <, +, \cdot, \ldots)$ *is* o-minimal *if every definable subset of* $\mathbb{R}$ *is a finite union of points and intervals.*

Theorem 8.4 (Decidability of ESL Fragment)**.** *Let $\mathcal{M}$ be a stratified space definable in an o-minimal structure, and let O, I, g, and $\mathcal{C}$ be definable. Then the satisfiability of quantifier-free ESL formulas is decidable.*

Proof Sketch. By the Tarski-Seidenberg theorem, the first-order theory of $(\mathbb{R}, <, +, \cdot)$ is decidable. O-minimality extends this to structures with additional definable functions. Quantifier-free ESL formulas reduce to Boolean combinations of polynomial inequalities, which are decidable in the o-minimal setting. □

Corollary 8.5. *For polynomial obligation fields, linear interest fields, and semialgebraic constraint sets, verification of safety properties is decidable.*

8.3 Verification Algorithm

8.4 Certification Artifacts

When verification succeeds, we can generate machine-checkable proofs.

Definition 8.6 (Ethical Certificate)**.** *An* ethical certificate *for a system $\mathcal{S}$ with respect to specification ϕ is a tuple $(\mathcal{S}, \phi, \pi, V)$ where:*

- *π is a formal proof that $\mathcal{S} \models \phi$*
- *V is a proof-checker that verifies π in polynomial time*

Certificates can be stored, transmitted, and independently verified, enabling trust without requiring re-verification.

Algorithm 2 ESL Verification

```
Require: ESL formula $\phi$, stratified space $\mathcal{M}$, ethical tensors $O, I, g$, constraints $\mathcal{C}$
Ensure: VERIFIED, VIOLATED, or UNKNOWN
Translate $\phi$ to first-order formula over reals
Apply cylindrical algebraic decomposition
Check truth value on each cell
if all cells satisfy $\phi$ then
    return VERIFIED
else if some cell violates $\phi$ then
    return VIOLATED with counterexample
else
    return UNKNOWN
end if
```

9 Learning-Theoretic Foundations

We now establish sample complexity bounds for learning ethical content from data.

9.1 Learning Obligation Fields

Obligation fields are typically specified by domain experts through constitutional principles. However, the *weights* of principles may be learned from data.

Definition 9.1 (Principle Weight Learning Problem)**.** *Given:*

- *k constitutional principles $\pi_1, \ldots, \pi_k$ with known predicates and directions*
- *A dataset $D = \{(x_i, y_i)\}_{i=1}^N$ where $y_i \in \{-1, +1\}$ indicates whether the aggregate obligation at x_i was correctly oriented*

Find weights $w_1, \ldots, w_k \geq 0$ that minimize classification error.

Theorem 9.2 (Sample Complexity for Weight Learning)**.** *Let $W = \{w \in \mathbb{R}^k_{\geq 0} : \|w\|_1 = 1\}$ be the weight simplex. With probability at least $1 - \delta$, the empirical risk minimizer over W achieves generalization error at most ϵ with*

$$N = O\left(\frac{k \log(k/\delta)}{\epsilon^2}\right)$$

samples.

Proof. The hypothesis class has VC dimension at most k. Standard PAC bounds give the result. □

9.2 Learning Interest Fields

Interest fields require learning stakeholder utilities from preference comparisons.

Theorem 9.3 (Sample Complexity for Utility Learning)**.** *Let $\mathcal{H}$ be a hypothesis class of utility functions with pseudo-dimension d. Given a preference oracle with noise rate $\eta < 1/2$, the maximum likelihood estimator achieves utility estimation error at most ϵ (in sup-norm over a compact domain) with probability at least $1 - \delta$ using*

$$N = O\left(\frac{d \log(1/\epsilon) + \log(1/\delta)}{(1 - 2\eta)^2 \epsilon^2}\right)$$

pairwise comparisons.

Proof Sketch. Adapt results from [11] on ranking from pairwise comparisons, using the pseudo-dimension to control the complexity of $\mathcal{H}$. □

Corollary 9.4. *For linear utility functions on $\mathbb{R}^n$, the sample complexity is $O(n \log(n/\epsilon)/\epsilon^2)$.*

9.3 Metric Learning

The governance metric can also be learned from trajectory data.

Definition 9.5 (Metric Learning Problem)**.** *Given:*

- *A set of "ethically preferred" trajectories $\Gamma^+ = \{\gamma_1^+, \ldots, \gamma_m^+\}$*
- *A set of "ethically dispreferred" trajectories $\Gamma^- = \{\gamma_1^-, \ldots, \gamma_m^-\}$*

Find a metric g such that preferred trajectories have lower cost than dispreferred ones.

Theorem 9.6 (Metric Learning Bound)**.** *Let $\mathcal{G}$ be a class of metrics parameterized by p parameters, each in $[0, 1]$. A metric consistent with all trajectory preferences (if one exists) can be found with probability at least $1 - \delta$ using*

$$m = O\left(\frac{p \log p + \log(1/\delta)}{\epsilon}\right)$$

trajectory pairs, where ϵ is the margin of separation.

10 Case Study: Medical Triage

We instantiate SGE in a medical triage scenario to demonstrate its practical applicability.

10.1 Scenario

An emergency department must allocate a single critical care slot among three patients:

- **Patient A:** Critical chest pain, arrived recently, socioeconomically disadvantaged
- **Patient B:** Stable condition, good prognosis, arrived earlier
- **Patient C:** Moderate severity, but allocation would involve coercion and privacy violations

10.2 Stratified Space Construction

Let $\mathcal{M} = \{A, B, C\} \cup \Delta^2_{\text{int}}$ where:

- $\{A, B, C\}$ are 0-dimensional strata representing pure allocations
- Δ^2_{int} is the interior of the 2-simplex representing probabilistic or partial allocations

The stratification captures that deterministic allocation (vertex) is categorically different from probabilistic allocation (interior).

10.3 Constitutional Principles

We define four principles:

1. **Beneficence:** $\phi_1(x) = 1$ always; $d_1(x) = \nabla E[\text{health outcome}]$
2. **Urgency:** $\phi_2(x) = 1$ if any patient is critical; $d_2(x)$ points toward the most urgent patient
3. **Equity:** $\phi_3(x) = 1$ if disadvantaged patients are present; $d_3(x)$ points toward disadvantaged patients
4. **Rights:** $\phi_4(x) = 1$ always; $d_4(x) = -\nabla(\text{coercion} + \text{privacy violation})$

The aggregate obligation field is:

$$O(x) = w_1 d_1(x) + w_2 d_2(x) + w_3 d_3(x) + w_4 d_4(x)$$

10.4 Stakeholder Interests

Three stakeholder groups are modeled:

- **Clinical staff:** Utility based on expected health outcomes and professional ethics
- **Patients/families:** Utility based on individual outcomes and procedural fairness

- **Institution:** Utility based on legal compliance and resource efficiency

The aggregate interest field is:

$$I(x) = \alpha_{\text{clin}} I_{\text{clin}}(x) + \alpha_{\text{pat}} I_{\text{pat}}(x) + \alpha_{\text{inst}} I_{\text{inst}}(x)$$

10.5 Constraints

The constraint set $\mathcal{C}$ includes:

- Allocations involving coercion: $\{x : \text{coercion}(x) > 0\}$
- Allocations violating explicit consent: $\{x : \text{consent}(x) < \tau\}$

Patient C falls within $\mathcal{C}$, so $S(C) = -\infty$.

10.6 Satisfaction Evaluation

Using the sigmoid activation $f(z) = \tanh(z)$ with appropriate normalization:

Configuration	Raw Score	Satisfaction
Patient A	0.87	0.70
Patient B	0.62	0.55
Patient C	—	$-\infty$

10.7 Verification

We verify the specification:

$$\phi = \text{Always}(S(x) \geq 0.5 \vee x \in \mathcal{C})$$

Using Algorithm 2, we find that ϕ is satisfied for $x \in \{A\}$ and violated for $x \in \{B, C\}$.

The system recommends Patient A, with a formal certificate that:

1. Patient C is forbidden by constraints
2. Patient A has higher satisfaction than Patient B
3. The satisfaction of A exceeds the threshold 0.5

10.8 Audit Trail

The SGE system produces an audit trail:

```
DECISION: Allocate to Patient A
RATIONALE:
  - Obligations: High urgency (0.9),
    disadvantage (0.8), beneficence (0.7)
  - Interests: Clinical (0.85),
    patient (0.75), institutional (0.80)
  - Constraints: None violated
  - Satisfaction: 0.70 (threshold: 0.50)
ALTERNATIVES CONSIDERED:
  - Patient B: Satisfaction 0.55
    (below threshold)
  - Patient C: Constraint violation
    (coercion, consent)
CERTIFICATE: [machine-checkable proof]
```

11 Discussion and Limitations

11.1 Contributions

SGE advances the state of the art in several ways:

1. **Stratified structure** captures moral discontinuities that smooth manifolds cannot represent.
2. **Grounded content** provides constructive methods for deriving ethical tensors from principles and learned preferences.
3. **Generalized satisfaction** accommodates nonlinear moral phenomena through stratified contraction operators.
4. **Computational bounds** establish tractability for ethical evaluation and planning.
5. **Formal verification** enables machine-checkable proofs of ethical constraint satisfaction.
6. **Learning theory** provides sample complexity guarantees for ethical content acquisition.

11.2 Limitations

Several important limitations remain:

1. **Metaethical neutrality:** SGE provides a framework for *representing* ethical content but does not resolve fundamental metaethical questions about the source of moral truth.
2. **Computational cost:** While polynomial, the complexity of geodesic planning may be prohibitive for very high-dimensional moral spaces.
3. **Specification burden:** Constitutional principles must be specified by human experts; the framework does not automate this.

4. **Preference elicitation:** Learning stakeholder utilities requires significant interaction, which may be impractical in some settings.

5. **Adversarial robustness:** The framework does not address adversarial manipulation of ethical inputs.

11.3 Future Work

Several directions merit further investigation:

1. **Distributed ethics:** Extending SGE to multi-agent systems where different agents may have different ethical frameworks.

2. **Temporal dynamics:** Incorporating time-varying obligations and interests.

3. **Uncertainty quantification:** Propagating uncertainty through the ethical pipeline to produce confidence bounds on decisions.

4. **Neural integration:** Training neural networks to implement ethical evaluation while preserving formal guarantees.

5. **Empirical validation:** Testing SGE on real-world ethical decision problems and comparing with human judgments.

12 Conclusion

We have presented Stratified Geometric Ethics (SGE), a mathematical framework for implementing verifiable ethical reasoning in autonomous systems. By modeling moral spaces as stratified spaces rather than smooth manifolds, grounding ethical content in constitutional principles and learned preferences, generalizing satisfaction beyond bilinearity, establishing computational complexity bounds, and developing formal verification methods, SGE addresses the major limitations of prior geometric approaches to ethics.

The framework is not a complete solution to AI alignment—metaethical questions about the source of moral truth remain open, and the specification of constitutional principles requires human judgment. But SGE provides a rigorous mathematical substrate on which such judgments can be expressed, computed, verified, and audited. As AI systems take on increasingly consequential roles in society, such rigor is not a luxury but a necessity.

References

[1] A. H. Bond. Differential geometry for moral alignment: The mathematical foundations of DEME 3.0. Working paper, San Jose State University, 2024.

[2] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[3] R. C. Arkin. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.

[4] Y. Bai, S. Kadavath, S. Kundu, et al. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073, 2022.

[5] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proc. ICML*, pages 663–670, 2000.

[6] J. Fürnkranz and E. Hüllermeier. *Preference Learning*. Springer, 2010.

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

[8] R. Thom. Ensembles et morphismes stratifiés. *Bulletin of the American Mathematical Society*, 75(2):240–284, 1969.

[9] V. Nanda. Computational algebraic topology. Lecture notes, University of Oxford, 2020.

[10] L. van den Dries. *Tame Topology and O-minimal Structures*. Cambridge University Press, 1998.

[11] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

A Proof of Theorem 5.2

Proof. We prove the representation theorem in full detail.

Step 1: Invariant quantities. Let Σ be a stratified satisfaction operator satisfying Axioms 1–4. By coordinate invariance (Axiom 1), $\Sigma(O, I, g, \mathcal{C})(x)$ can depend only on coordinate-invariant quantities constructible from O_x, I_x, g_x, and $\mathcal{C}$.

The fundamental invariants are:

- The contraction $\langle I, O\rangle := I_\mu O^\mu$
- The norm $\|O\|_g := \sqrt{g_{\mu\nu}O^\mu O^\nu}$
- The dual norm $\|I\|_{g^{-1}} := \sqrt{g^{\mu\nu}I_\mu I_\nu}$
- The constraint indicator $\chi_{\mathcal{C}}(x)$

Any other scalar invariant is a function of these.

Step 2: Monotonicity constraint. By Axiom 2 (monotonicity), Σ must be increasing in $\langle I, O\rangle$ when other quantities are held fixed. This means

$$\Sigma(O, I, g, \mathcal{C})(x) = F(\langle I, O\rangle, \|O\|_g, \|I\|_{g^{-1}}, \chi_{\mathcal{C}}(x))$$

where F is increasing in its first argument.

Step 3: Scale invariance. Consider rescaling $O \mapsto \lambda O$ for $\lambda > 0$. If satisfaction depended on $\|O\|_g$ independently, then doubling the "intensity" of obligations would change satisfaction even if interests remain aligned. This violates the intuition that satisfaction measures *alignment* rather than *magnitude*.

Imposing scale invariance, $\Sigma(\lambda O, I, g, \mathcal{C}) = \Sigma(O, I, g, \mathcal{C})$, forces

$$\Sigma(O, I, g, \mathcal{C})(x) = \chi_{\mathcal{C}}(x) + \tilde{F}\left(\frac{\langle I, O\rangle}{\|O\|_g}, \|I\|_{g^{-1}}\right)$$

Step 4: Separation of scales. By similar reasoning for I, and noting that $\|I\|_{g^{-1}}$ can be absorbed into a scale field $\lambda(x)$, we obtain

$$\Sigma(O, I, g, \mathcal{C})(x) = \chi_{\mathcal{C}}(x) + \lambda(x) \cdot f\left(\frac{\langle I, O\rangle}{\|O\|_g}\right)$$

where $f : \mathbb{R} \to \mathbb{R}$ is smooth and increasing (from monotonicity) and $\lambda : \mathcal{M} \to \mathbb{R}_{>0}$ is smooth (from stratum compatibility).

Step 5: Constraint respect. Axiom 3 requires $\Sigma = -\infty$ on $\mathcal{C}$. This is achieved by setting $\chi_{\mathcal{C}}(x) = 0$ for $x \notin \mathcal{C}$ and $\chi_{\mathcal{C}}(x) = -\infty$ for $x \in \mathcal{C}$.

Step 6: Stratum compatibility. Axiom 4 requires smoothness on each stratum, which is satisfied by construction since f, λ, O, I, and g are all smooth on strata. □

B Proof of Theorem 6.3

Proof. **Step 1: Setup.** Let $\Omega(x_0, x_1)$ be the space of piecewise smooth paths from x_0 to x_1 in $\mathcal{M}$. The ethical cost functional is

$$\mathcal{L}[\gamma] = \int_0^1 \sqrt{g(\dot\gamma, \dot\gamma)} \cdot h(\gamma(t))\, dt$$

Step 2: Coercivity. Since $h > 0$ is continuous on the compact closure of any bounded region, h is bounded below by some $h_{\min} > 0$. Thus

$$\mathcal{L}[\gamma] \geq h_{\min} \cdot \text{length}_g(\gamma)$$

where length_g is the Riemannian length. By completeness of $(\mathcal{M}, g)$, any path with $\mathcal{L}[\gamma] \leq C$ is contained in a compact region.

Step 3: Minimizing sequences. Let $\{\gamma_n\}$ be a minimizing sequence for $\mathcal{L}$. By Step 2, all γ_n lie in a compact region. By Arzelà-Ascoli, a subsequence converges uniformly to some $\gamma^* \in \Omega(x_0, x_1)$.

Step 4: Lower semicontinuity. The functional $\mathcal{L}$ is lower semicontinuous with respect to uniform convergence (this follows from Fatou's lemma and the lower semicontinuity of the norm). Thus

$$\mathcal{L}[\gamma^*] \leq \liminf_{n\to\infty} \mathcal{L}[\gamma_n] = \inf_{\gamma \in \Omega(x_0, x_1)} \mathcal{L}[\gamma]$$

Step 5: Regularity at strata boundaries. Whitney's condition (B) ensures that minimizers have well-defined tangent vectors at stratum boundaries, preventing pathological behavior. The minimizer γ^* is piecewise smooth with possible corners only at stratum crossings. □

C Proof of Theorem 8.4

Proof. **Step 1: O-minimal definability.** By assumption, $\mathcal{M}$, O, I, g, and $\mathcal{C}$ are definable in an o-minimal structure $\mathfrak{R} = (\mathbb{R}, <, +, \cdot, \ldots)$.

Step 2: Translation to first-order. A quantifier-free ESL formula ϕ is a Boolean combination of atomic formulas of the form:

- $S(x) \bowtie c$, which expands to $\lambda(x) \cdot f(I_\mu(x)O^\mu(x)/\|O(x)\|_g) \bowtie c$
- $O_i(x) \bowtie c$, which is a polynomial inequality
- $I_j(x) \bowtie c$, which is a polynomial inequality

If f is a definable function in $\mathfrak{R}$, then all atomic formulas are definable.

Step 3: Decidability. By the definability of all components, the truth value of ϕ at any point $x \in \mathcal{M}$ is a definable predicate. By o-minimality, the set $\{x \in \mathcal{M} : \phi(x)\}$ is definable.

The satisfiability problem "does there exist $x \in \mathcal{M}$ such that $\phi(x)$?" is equivalent to asking whether a definable set is nonempty. By Tarski-Seidenberg (or its o-minimal generalization), this is decidable. □