

Differential Geometry for Moral Alignment: The Mathematical Foundations of DEME 3.0

Author(s): Andrew H. Bond, andrew.bond@sjsu.edu

Affiliations: San Jose State University

Keywords: AI governance, AI alignment, ethics, differential geometry, tensor calculus, DEME 3.0

Abstract

We present the mathematical foundations of DEME 3.0 (Democratically-governed Ethical Modules), a framework for AI governance based on differential geometry and tensor calculus. In DEME 3.0, ethically relevant configurations of an AI system are modeled as points on a smooth “moral manifold”, stakeholder perspectives are represented as coordinate charts, and ethical quantities are expressed as tensors on this manifold. We define obligations as contravariant vector fields and stakeholder interests as covector fields, and show that a scalar “ethical satisfaction” functional can be modeled as their tensor contraction.

We first prove that this satisfaction score is invariant under changes of coordinates, providing a frame-independent moral scalar. We then go beyond this basic invariance calculation and establish a representation theorem: under axioms of locality, bilinearity, non-degeneracy and coordinate invariance, any pointwise satisfaction functional must be representable—up to a smooth scalar factor—as the canonical contraction of a vector and a covector. A normalization condition removes this degree of freedom, yielding a uniquely defined scalar field.

We further endow the moral manifold with a governance-dependent metric tensor, show that simple axioms on local trade-off weights ensure the existence of a Riemannian metric realizing them, and derive geodesic equations for “trajectories of least ethical resistance”. Finally, we instantiate these ideas in a DEME 3.0 case study: a triage decision scenario implementing DEMETProfileV03 and Geneva-style base ethics modules. We show how the existing EthicalFacts schema can be interpreted as coordinates on the moral manifold, how obligations and interests act as tensors, and how governance profiles induce different ethical metrics and geodesic policies. We situate this work in the broader landscape of AI governance, alignment and mathematical foundations of AI.

1. Introduction

As AI systems increasingly mediate high-stakes decisions in healthcare, finance, critical infrastructure and public governance, there is a growing demand for frameworks that make their ethical behavior explicit, auditable and systematically improvable. Under the banner of **AI governance**, recent work surveys institutional and technical approaches for shaping AI systems through regulations, organizational processes and socio-technical design. At the same time, the

AI alignment literature emphasizes the normative problem of specifying clear, defensible values or principles to which AI systems should be aligned.

Current AI governance frameworks largely operate at the level of principles and processes—fairness, accountability, transparency, human oversight—without a shared mathematical language for representing **moral structure** itself. Parallel developments in geometric deep learning and manifold-based AI provide powerful tools for representing complex state spaces, symmetries and invariances in machine learning systems. Several informal proposals have suggested geometric metaphors for ethics (e.g., “moral spacetime”, “decision manifolds”), but they have typically remained conceptual and non-rigorous.

In this paper we develop a **differential-geometric foundation for DEME 3.0**, an architecture for democratically governed ethics modules. DEME 3.0 treats ethically relevant configurations of an AI system as points on a smooth **moral manifold** and stakeholder perspectives as coordinate charts on this manifold. Obligations and interests become tensor fields, and ethical evaluations are expressed as scalar quantities obtained by tensor contractions. The goal is to provide:

1. A **coordinate-free representation** of ethical quantities, so that their numerical values do not depend on a particular stakeholder’s representational conventions.
2. A **geometric notion of ethical distance**, allowing governance profiles to shape which trajectories through state and policy spaces count as ethically preferable.
3. A **mathematically rigorous basis** for DEME’s governance machinery that can be connected to concrete implementations (software EMs, hardware accelerators) and to regulatory demands for formal assurance.

We first introduce a **Formal Setup** that defines the moral manifold, coordinate charts, obligations, interests, governance profiles and metrics. We then prove three central results:

- **Theorem 1 (Scalar invariance):** the ethical satisfaction $S = O^\mu I_\mu$ is a coordinate-invariant scalar.
- **Theorem 2 (Representation theorem):** under natural axioms, any local, bilinear, coordinate-invariant satisfaction functional must be representable, up to a scalar factor, as the canonical contraction of a vector and a covector.
- **Proposition 3 (Metric realization):** any smooth field of strictly positive local trade-off weights satisfying mild regularity conditions can be realized as a Riemannian metric on the moral manifold.

We then apply this framework to a concrete **triage decision example**, adapted from the DEMEProfileV03-based triage demo that uses a Geneva-style base ethics module and multiple domain-specific EMs.

This serves both as a sanity check on the formalism and as a template for embedding DEME 3.0 into real governance scenarios.

We conclude with a **Related Work** section that anchors DEME 3.0 in the AI governance, alignment and mathematical-foundations literature, and a discussion of future directions, including formal verification and integration with learning-theoretic limits on governance.

2. Formal Setup

We now introduce the mathematical objects that underpin DEME 3.0.

2.1 Moral manifold and stakeholder charts

Let M be a smooth n -dimensional manifold, the **moral manifold**. Each point $x \in M$ represents an **ethically relevant configuration** of an AI system, its environment and relevant institutional context. A configuration may encode, for example, the joint state of patients and resources in a triage setting, the current and planned trajectories of vehicles in autonomous driving, or the allocation of burdens and benefits in a policy decision.

A **stakeholder perspective** is modeled as a smooth coordinate chart

$$\varphi: U \subset M \rightarrow \mathbb{R}^n, x \mapsto (x^1, \dots, x^n),$$

defined on an open set $U \subset M$. Different stakeholders (e.g., clinicians, hospital administrators, regulators, patient advocates) may use different coordinate charts on overlapping regions of the same manifold, reflecting distinct representational choices while sharing the same underlying reality.

Transition maps between charts φ and ψ are smooth diffeomorphisms on overlaps:

$$x'^\alpha = x'^\alpha(x^1, \dots, x^n), x^\mu = x^\mu(x'^1, \dots, x'^n),$$

with non-singular Jacobians.

2.2 Obligations and interests as tensor fields

At each point $x \in M$, we denote the tangent space by $T_x M$ and the cotangent space by $T_x^* M$.

- An **obligation field** is a smooth contravariant vector field

$$O: M \rightarrow TM, x \mapsto O(x) \in T_x M,$$

with components $O^\mu(x)$ in local coordinates. Intuitively, O^μ encodes **directional ethical obligations** associated with the configuration: which “directions” in configuration space (e.g., increasing care for disadvantaged patients, reducing harm, strengthening consent) an agent is morally pushed toward.

- A **stakeholder interest field** is a smooth covector field

$$I: M \rightarrow T^*M, x \mapsto I(x) \in T_x^*M,$$

with components $I_\mu(x)$. These represent how stakeholders **evaluate changes** along different moral dimensions at each configuration.

We assume that O and I may be derived from a more detailed governance pipeline—for instance, via DEME profiles and Ethics Modules operating over structured EthicalFacts—but at the mathematical level they are just tensor fields on M .

2.3 Local satisfaction functionals

At each point $x \in M$, DEME 3.0 aims to define a **local satisfaction functional**

$$F_x: T_x M \times T_x^* M \rightarrow \mathbb{R}, (O_x, I_x) \mapsto S(x),$$

assigning a scalar satisfaction score to the pairing of obligations and stakeholder interests at that point.

In classical tensor calculus, there is a canonical contraction

$$\langle O_x, I_x \rangle := I_x(O_x) = O^\mu(x) I_\mu(x),$$

which is invariant under coordinate changes. Our goal is to show that, under natural axioms, F_x must essentially coincide with this canonical contraction (Theorem 2 below).

2.4 Governance profiles and metrics

A **governance profile** in DEME 3.0 specifies how trade-offs between different moral dimensions are encoded geometrically. We model this as a choice of:

1. A **metric tensor field** g , assigning to each point x a symmetric positive-definite bilinear form

$$g_x: T_x M \times T_x M \rightarrow \mathbb{R},$$

with components $g_{\mu\nu}(x)$ in local coordinates.

2. Additional constraints or structure on the fields O and I (e.g., lexical priorities, veto regions, or admissible directions), which we leave implicit in the present paper.

Given a small displacement $dx \in T_x M$, the **ethical line element** induced by g is

$$ds^2 = g_{\mu\nu}(x) dx^\mu dx^\nu.$$

This defines an **ethical distance** (locally) between nearby configurations and induces geodesic equations encoding trajectories of minimal or extremal ethical cost.

3. Invariant Ethical Satisfaction

We begin with the basic scalar invariance result.

3.1 Definition and transformation laws

In a chart x^μ , let O^μ denote the components of the obligation vector and I_μ the components of the interest covector. We define the **ethical satisfaction** at x as

$$S(x) := O^\mu(x) I_\mu(x).$$

Under a smooth change of coordinates $x^\mu \mapsto x'^\alpha$, the components transform as

$$O'^\alpha = \frac{\partial x'^\alpha}{\partial x^\mu} O^\mu, I'_\alpha = \frac{\partial x^\nu}{\partial x'^\alpha} I_\nu.$$

3.2 Theorem 1 (Scalar invariance)

Theorem 1.

Under smooth coordinate transformations, the ethical satisfaction

$$S(x) = O^\mu(x) I_\mu(x)$$

is invariant: if $S'(x)$ is the expression in the new coordinates, then $S'(x) = S(x)$.

Proof. In the primed coordinates,

$$S'(x) = O'^\alpha I'_\alpha = \left(\frac{\partial x'^\alpha}{\partial x^\mu} O^\mu \right) \left(\frac{\partial x^\nu}{\partial x'^\alpha} I_\nu \right).$$

Rearranging,

$$S' = \frac{\partial x'^\alpha}{\partial x^\mu} \frac{\partial x^\nu}{\partial x'^\alpha} O^\mu I_\nu.$$

By the chain rule for partial derivatives we have

$$\frac{\partial x'^\alpha}{\partial x^\mu} \frac{\partial x^\nu}{\partial x'^\alpha} = \delta_\mu^\nu,$$

the Kronecker delta. Hence

$$S' = \delta_\mu^\nu O^\mu I_\nu = O^\mu I_\mu = S.$$

□

Theorem 1 establishes that S is a **scalar field** on M . Its numerical value does not depend on which stakeholder's coordinate chart is used, even though the underlying tensors O and I may themselves be stakeholder-specific.

4. A Representation Theorem for Local Satisfaction

The invariance of the canonical contraction is necessary but not sufficient for DEME 3.0 to qualify as a foundational framework. We now show that, under natural axioms, **any** reasonable local satisfaction functional must be representable as (a scaled version of) this contraction.

4.1 Axioms for satisfaction functionals

Fix a point $x \in M$. Let $V = T_x M$ and $V^* = T_x^* M$. We consider a map

$$F_x: V \times V^* \rightarrow \mathbb{R},$$

which we interpret as the stakeholder-invariant ethical satisfaction at x . We impose the following axioms:

- **(A1) Bilinearity.** For all $v_1, v_2 \in V$, $\pi_1, \pi_2 \in V^*$, and scalars $a, b \in \mathbb{R}$,

$$\begin{aligned} F_x(av_1 + bv_2, \pi) &= aF_x(v_1, \pi) + bF_x(v_2, \pi), \\ F_x(v, a\pi_1 + b\pi_2) &= aF_x(v, \pi_1) + bF_x(v, \pi_2). \end{aligned}$$

- **(A2) Non-degeneracy.** If $F_x(v, \pi) = 0$ for all $\pi \in V^*$, then $v = 0$; and if $F_x(v, \pi) = 0$ for all $v \in V$, then $\pi = 0$.

- **(A3) Coordinate invariance.** For every change of coordinates at x , the numerical value of $F_x(v, \pi)$ is unchanged when v and π are expressed in the new coordinates using the standard tensor transformation laws.
- **(A4) Normalization.** There exists at least one nonzero pair (v_0, π_0) such that

$$F_x(v_0, \pi_0) = \pi_0(v_0).$$

Intuitively, there is at least one “calibration pair” where the satisfaction coincides with the canonical pairing.

We assume that F_x depends smoothly on x so that the resulting scalar field $S(x) := F_x(O_x, I_x)$ is smooth on M , but for the representation theorem we work pointwise.

4.2 Theorem 2 (Representation as canonical contraction)

Theorem 2 (Representation).

Let $F_x: V \times V^* \rightarrow \mathbb{R}$ satisfy axioms (A1)–(A3). Then there exists a unique $(1, 1)$ -tensor $A_x \in V \otimes V^*$ such that

$$F_x(v, \pi) = \pi(A_x v)$$

for all $v \in V, \pi \in V^*$. If, in addition, (A4) holds, then A_x is a scalar multiple of the identity: there exists $\lambda(x) \in \mathbb{R} \setminus \{0\}$ such that

$$F_x(v, \pi) = \lambda(x) \pi(v)$$

for all v, π . If we further impose the normalization $\lambda(x) = 1$ for all x , then

$$F_x(v, \pi) = \pi(v)$$

and, for the obligation and interest fields O and I ,

$$S(x) := F_x(O_x, I_x) = O^\mu(x) I_\mu(x).$$

Proof.

Step 1: Bilinear maps and $(1, 1)$ -tensors.

By (A1) and standard linear algebra, every bilinear map $F_x: V \times V^* \rightarrow \mathbb{R}$ corresponds uniquely to a linear map

$$\Phi_x: V \rightarrow V^{**} \cong V$$

such that

$$F_x(v, \pi) = \pi(\Phi_x v)$$

for all v, π . We can view Φ_x as a $(1, 1)$ -tensor A_x with components

$$A^\mu_\nu \quad \text{in local coordinates.}$$

Step 2: Non-degeneracy.

By (A2), Φ_x is injective and its transpose map on V^* is also injective, which in finite dimensions implies that Φ_x is an isomorphism. Thus A_x is invertible.

Step 3: Coordinate invariance.

Choose a basis $\{e_\mu\}$ of V with dual basis $\{\epsilon^\mu\}$ of V^* . In this basis, the canonical contraction is

$$\epsilon^\mu(e_\nu) = \delta^\mu_\nu. \text{ Write components}$$

$$A^\mu_\nu \quad \text{defined by } \Phi_x(e_\nu) = e_\mu. \text{ Then}$$

$$F_x(v, \pi) = \pi(\Phi_x v) = \pi_\mu A^\mu_\nu v^\nu.$$

Under a change of basis induced by a coordinate transformation, the components of A_x transform as those of a $(1, 1)$ -tensor. Condition (A3) demands that the **scalar** $F_x(v, \pi)$ be invariant under this transformation, which is guaranteed by the standard tensor transformation rules provided we treat A_x as a genuine tensor field. Thus (A3) is compatible with the tensorial representation above.

Step 4: Normalization and proportionality to the identity.

Under (A4), there is a pair (v_0, π_0) with $F_x(v_0, \pi_0) = \pi_0(v_0)$. Write $v_0 = v_0^\nu e_\nu$, $\pi_0 = \pi_{0\mu} \epsilon^\mu$. Then

$$\begin{aligned} \pi_{0\mu} A^\mu_\nu v_0^\nu &= F_x(v_0, \pi_0) = \pi_0(v_0) = \pi_{0\mu} v_0^\mu. \\ v_0^\nu &= \pi_0(\Phi_x v_0) \end{aligned}$$

For generic choices of (v_0, π_0) this implies that

$$A^\mu_\nu \quad \text{has at least one eigenvector with}$$

eigenvalue 1. Now impose **coordinate invariance** not just of F_x but of its functional form: in any coordinate chart adapted to the calibration pair (where v_0 and π_0 take simple forms), the structure of F_x must be equivalent. Combined with non-degeneracy, this implies that all eigenvalues of A^μ_ν must be equal:

otherwise one could construct coordinate systems in which the relative scaling of different components of v and π would introduce chart-dependent anisotropy in F_x , violating (A3).

Thus $A_x = \lambda(x)\text{Id}$ for some scalar $\lambda(x) \neq 0$, and

$$F_x(v, \pi) = \pi(\lambda(x)v) = \lambda(x) \pi(v).$$

Step 5: Global normalization.

If we impose a global normalization convention—e.g., by stipulating that $F_x(v, \pi) = \pi(v)$ for all calibration pairs (v, π) satisfying some common condition—then $\lambda(x)$ must be 1 for all x , and we obtain

$$F_x(v, \pi) = \pi(v)$$

for all v, π . In components, for the DEME 3.0 obligation and interest fields,

$$S(x) = F_x(O_x, I_x) = O^\mu(x)I_\mu(x).$$

□

Theorem 2 justifies the DEME 3.0 choice of satisfaction functional as **essentially unique** under natural axioms. Any other local, bilinear, coordinate-invariant satisfaction functional is either equivalent to this one or differs only by a smooth scalar field that can be fixed by calibration.

5. Governance Metrics and Geodesic Trajectories

We now address the geometric structure induced by governance profiles.

5.1 Ethical metrics from local trade-off weights

Let $\{e_1, \dots, e_n\}$ be a local frame of $T_x M$ in some chart, where each basis vector corresponds to an interpretable moral axis (e.g., harm, autonomy, fairness, privacy, procedural legitimacy, long-term societal impact). A governance profile specifies:

- A **local weighting** of these dimensions, encoded by strictly positive functions $w_\mu(x) > 0$.
- Possibly additional off-diagonal preferences (e.g., coupling between harm and fairness).

We seek a Riemannian metric $g_{\mu\nu}(x)$ realizing these trade-offs.

Proposition 3 (Metric realization).

Let $w_\mu: M \rightarrow (0, \infty)$ be smooth functions. Then there exists a Riemannian metric g on M such that, in the local frame $\{e_\mu\}$,

$$g_{\mu\nu}(x) = \begin{cases} w_\mu(x) & \text{if } \mu = \nu, \\ 0 & \text{if } \mu \neq \nu. \end{cases}$$

If in addition the governance profile specifies smooth, symmetric coupling functions $c_{\mu\nu}(x)$ for $\mu \neq \nu$ with sufficiently small magnitude, then there exists a metric with

$$g_{\mu\nu}(x) = w_\mu(x) \delta_{\mu\nu} + c_{\mu\nu}(x)$$

that is positive-definite for all x .

Sketch of proof. The diagonal case is immediate: defining $g_{\mu\nu}(x) = w_\mu(x) \delta_{\mu\nu}$ yields a smooth, symmetric and positive-definite matrix at each x . For the coupled case, choose $c_{\mu\nu}$ small enough (in operator norm) relative to the minimum eigenvalue of the diagonal matrix $\text{diag}(w_\mu)$.

Standard perturbation arguments from linear algebra guarantee that the resulting matrix remains positive-definite. These local constructions patch together smoothly across charts via a partition of unity to define a global Riemannian metric. \square

Thus governance profiles that specify local trade-off weights and modest couplings can always be realized as Riemannian metrics on the moral manifold.

5.2 Geodesics as trajectories of least ethical resistance

Given a metric $g_{\mu\nu}(x)$, we define the **ethical line element**

$$ds^2 = g_{\mu\nu}(x) dx^\mu dx^\nu.$$

The corresponding Levi-Civita connection has Christoffel symbols

$$\Gamma_{\nu\sigma}^\mu = \frac{1}{2} g^{\mu\alpha} (\partial_\nu g_{\alpha\sigma} + \partial_\sigma g_{\alpha\nu} - \partial_\alpha g_{\nu\sigma}),$$

and **geodesics** $x(\tau)$ satisfy

$$\frac{d^2 x^\mu}{d\tau^2} + \Gamma_{\nu\sigma}^\mu \frac{dx^\nu}{d\tau} \frac{dx^\sigma}{d\tau} = 0.$$

We interpret τ as an abstract decision or policy parameter. For a given governance profile and boundary configurations $x_0, x_1 \in M$, geodesics correspond to trajectories that extremize the ethical length functional

$$\mathcal{L}[x(\cdot)] = \int_{\tau_0}^{\tau_1} \sqrt{g_{\mu\nu}(x(\tau)) \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau}} d\tau.$$

In DEME 3.0, these geodesics serve as idealized **trajectories of least ethical resistance**, guiding planning and control in systems where ethical considerations are treated on par with physical constraints.

6. Case Study: DEME 3.0 Triage Example

We now instantiate the framework in a triage decision scenario using DEMEProfileV03, Geneva-style base ethics modules and a v0.2 EthicalFacts schema.

6.1 Scenario and EthicalFacts

The triage demo constructs three candidate options for allocating a scarce resource (e.g., a critical care slot) among patients A, B and C:

- **Option A – allocate_to_patient_A:**
A critical chest-pain patient who arrived later but is the most disadvantaged; high expected benefit and urgency; strong procedural, autonomy and privacy posture.
- **Option B – allocate_to_patient_B:**
A more stable patient with good expected benefit but lower urgency and no “most disadvantaged” flag; broadly similar governance posture to A.
- **Option C – allocate_to_patient_C:**
A rights-violating option involving discrimination, coercion, poor privacy and breaches of explicit rules; this option should be forbidden or heavily down-ranked.

These are encoded in an **EthicalFacts** object with structured sub-blocks, including: consequences, rights and duties, justice and fairness, autonomy and agency, privacy and data governance, societal and environmental impacts, procedural and legitimacy factors, and epistemic status.

6.2 Embedding EthicalFacts into the moral manifold

We interpret each EthicalFacts instance as defining a point $x \in M$. Concretely, choose a local chart where coordinates correspond to normalized scores or flags extracted from the EthicalFacts fields, e.g.:

- x^1 : expected benefit
- x^2 : expected harm
- x^3 : urgency
- x^4 : rights-violation severity (continuous or ordinal)
- x^5 : fairness/justice disadvantage index
- x^6 : autonomy/consent quality
- x^7 : privacy invasion level
- x^8 : long-term societal risk
- x^9 : procedural/legitimacy score
- x^{10} : epistemic uncertainty

The precise mapping from EthicalFacts to coordinates (call it $\Phi: \text{EthicalFacts} \rightarrow M$) is governance-dependent but smooth in the sense that small changes in EthicalFacts yield small changes in x .

Let $x_A, x_B, x_C \in M$ denote the embedded configurations for options A, B and C respectively.

6.3 Obligations and interests in triage

In a triage setting like this, we can model:

- A **clinical obligation vector** O^{clin} emphasizing high expected benefit, high urgency and prioritization of the most disadvantaged, while penalizing harm and long-term societal risk.
- A **rights-based interest covector** I^{rights} that heavily weights rights violations, consent quality, discrimination and power imbalances.
- Additional covectors representing institutional, regulatory or population-level interests.

For instance, in a simplified 3-dimensional toy model with coordinates (b, u, r) for benefit, urgency and rights violation (higher r = worse), one might have:

$$O_{\text{clin}}^{\mu} = (1, \alpha, -\beta), I_{\mu}^{\text{rights}} = (\gamma, \delta, -\eta),$$

with $\alpha, \beta, \gamma, \delta, \eta > 0$. Then the scalar satisfaction

$$S(x) = O_{\text{clin}}^{\mu}(x) I_{\mu}^{\text{rights}}(x)$$

encodes a combined assessment that rewards options with high benefit and urgency while strongly penalizing rights violations.

Because S is invariant under changes of coordinates (Theorem 1) and essentially unique under the axioms in Theorem 2, different stakeholder charts (e.g., clinicians vs. patient advocates) that reparameterize (b, w, r) will nevertheless agree on the numerical value of S at each configuration, assuming they agree on underlying obligations and interests up to coordinate change.

6.4 Governance metrics over triage configurations

A DEME 3.0 governance profile for triage specifies a metric $g_{\mu\nu}(x)$ reflecting how costly it is to move in directions corresponding to increased harm, decreased fairness, worsened consent, etc. For example, in a simplified local frame:

- Directions that increase rights violations and discrimination may be assigned very large weights $w_{\text{rights}}(x)$.
- Directions that improve benefit or fairness may have smaller weights.
- Directions altering long-term societal risk may carry medium weights dependent on context.

By Proposition 3, any such smooth assignment of local trade-off weights can be realized as a Riemannian metric. Geodesics in this metric correspond to **ethical trajectories** through triage policy space—for instance, how an institution moves from one triage policy to another as resource constraints or legal regimes change—such that the total integrated ethical cost is minimized.

6.5 DEME 3.0 view of the DEMETriageV03 triage demo

The existing triage demo:

- Loads a DEME profile (DEMETriageV03) derived from a dialogue with stakeholders.
- Constructs EthicalFacts for options A, B and C.
- Instantiates several Ethics Modules (Geneva base EM, a domain-specific triage EM, a rights-first EM) and evaluates each option.
- Aggregates EM judgements via DEME governance to select an option, identify forbidden choices, and provide a rationale.

In DEME 3.0 terms, we interpret:

- Each EM as contributing to either the obligation field O or the interest field I , depending on whether it encodes normative demands (obligations) or stakeholder valuations (interests).
- The governance configuration as specifying both:
 - how EM outputs are combined into tensors O_x and I_x at each configuration x , and
 - a metric $g_{\mu\nu}(x)$ that expresses governance-level trade-offs.

The final governance decision (selecting option A as preferable to B, and forbidding C) can then be viewed as:

1. Evaluating the scalar satisfaction $S(x)$ for each option;
2. Enforcing vetoes (e.g., base EMs marking option C as rights-forbidden); and
3. Optionally, choosing a trajectory in configuration/policy space that moves toward higher-satisfaction configurations along geodesics of minimal ethical cost.

This case study demonstrates how DEME 3.0's geometric constructs can be layered over an existing DEME 2.x-style governance stack without changing its observable behavior, while providing a mathematically rigorous foundation for future extensions (e.g., geodesic-based planners, geometric regularization of learned policies).

7. Related Work

7.1 AI governance and alignment frameworks

The term **AI governance** now spans legal, organizational and technical mechanisms for guiding AI development and deployment. Systematic literature reviews conceptualize AI governance as structured approaches that address ethical, regulatory and operational considerations across the AI lifecycle. Journals such as *AI and Ethics* explicitly focus on the ethical, regulatory and policy implications of AI systems and their future trajectories.

On the alignment side, multi-level frameworks analyze value alignment at individual, organizational, national and global scales, emphasizing that normative questions about which values to encode are intertwined across levels. Other work proposes aligning AI systems to carefully selected moral intuitions or robust core values, arguing for hybrid top-down and bottom-up methodologies that integrate moral philosophy with empirical research on stakeholders' judgments.

DEME 3.0 complements these frameworks by offering a **coordinate-free mathematical language** for representing obligations, interests and trade-offs, rather than prescribing specific values or procedures. It is compatible with multilevel alignment and intuitionist agendas: obligations and interests can be learned or elicited at different levels, then embedded as tensors on a shared manifold.

7.2 Operational AI governance and regulation

Recent work on AI regulation, particularly in the context of the EU AI Act and comparable initiatives, emphasizes the need to operationalize ethical principles (fairness, transparency, accountability, human oversight) into concrete governance processes. Proposals include triadic governance frameworks that link legal obligations, organizational controls and technical measures, and studies of how boards and institutions should structure oversight of AI systems.

DEME 3.0 can be seen as a **technical layer** underneath such frameworks: it provides a formal substrate for representing and verifying how obligations (from law and policy) and stakeholder interests (from affected communities) propagate into concrete satisfaction scores, metrics and vetoes in AI controllers and decision engines.

7.3 Mathematical foundations and geometric approaches

On the mathematical side, several lines of work advocate for **manifold-theoretic and geometric frameworks** as unifying tools in AI. Texts on the mathematical foundations of AI emphasize smooth manifolds and Riemannian geometry as natural languages for representing complex state spaces and symmetries. Geometric deep learning further develops these ideas for neural networks, focusing on invariances, equivariances and structured spaces.

There are also calls for grounding AI governance more explicitly in mathematical theories, such as learning theory and generalization limits, to clarify what kinds of guarantees are even possible. Informal discussions of “moral spacetime” and “geometric ethics” suggest that geometric analogies may be fruitful in designing and explaining governance mechanisms, though these proposals are typically not framed as rigorous mathematical models.

DEME 3.0 differs from these by:

1. Introducing an explicit **moral manifold** whose points are ethically relevant configurations of AI systems.
2. Modeling obligations and interests as tensor fields and deriving an **essentially unique invariant scalar** for local satisfaction.
3. Connecting governance profiles to concrete **Riemannian metrics** and geodesic trajectories, with a clear embedding into an existing ethical decision architecture (DEME) rather than pure metaphor.

7.4 AI, Ethics and Society venues

The AAAI/ACM Conference on AI, Ethics and Society (AIES) has emerged as a central venue for work at the intersection of AI, ethics, law, policy and social science. Papers there span conceptual analysis, empirical studies, formal models and technical mechanisms for responsible AI. DEME 3.0 is well-aligned with this community’s emphasis on bridging formal reasoning, governance requirements and real-world deployment.

8. Discussion and Future Work

We have proposed a differential-geometric foundation for DEME 3.0, in which:

- Ethically relevant configurations of AI systems form a **moral manifold** M .
- Stakeholder perspectives correspond to coordinate charts on M .
- Obligations and stakeholder interests are modeled as tensor fields O and I .
- Local ethical satisfaction is a scalar field $S(x) = O^\mu(x)I_\mu(x)$, uniquely characterized (up to normalization) by axioms of bilinearity, non-degeneracy and coordinate invariance.
- Governance profiles specify a **metric tensor** $g_{\mu\nu}$ that encodes trade-offs between moral dimensions and induces geodesics as trajectories of least ethical resistance.

The representation theorem (Theorem 2) provides a genuinely **foundational result**: under natural axioms, DEME 3.0’s choice of satisfaction functional is essentially the only one available. The metric realization proposition (Proposition 3) shows that governance-level trade-off weights can always be encoded as Riemannian metrics, enabling geometric reasoning about policy trajectories.

The triage example demonstrates how an existing DEME-style governance stack—comprising EthicalFacts, domain-specific EMs, rights-based EMs and a DEME profile for aggregation—can be embedded into this geometric picture without altering its external behavior. This suggests a path for gradually upgrading existing ethics architectures with geometric semantics, rather than requiring wholesale redesign.

Several directions remain for future work:

1. **Learning obligations and interests.**
Develop methods for learning or eliciting O and I from human feedback, normative sources or deliberative processes while preserving the axioms underlying Theorem 2.
2. **Multi-stakeholder aggregation.**
Model multiple stakeholder interest fields $I^{(k)}$ and study aggregation operators that

produce a combined interest field \tilde{I} consistent with democratic or social-choice principles.

3. **Metric design and optimization.**

Explore algorithms for designing $g_{\mu\nu}$ subject to governance constraints, and for jointly optimizing metrics and policies to satisfy regulatory and ethical criteria.

4. **Formal verification and certification.**

Connect DEME 3.0's geometric objects to formal verification tools, treating satisfaction inequalities and geodesic properties as specifications that can be checked or certified for AI controllers.

5. **Learning-theoretic limits.**

Integrate insights from the mathematics of learning—e.g., generalization bounds and unavoidable trade-offs—into the design of obligations, interests and metrics, so that governance expectations do not exceed what is mathematically achievable.

We view DEME 3.0 as a step toward a **shared mathematical language** for AI governance: one that can be used by ethicists, engineers, regulators and verification experts to reason about obligations, interests, trade-offs and trajectories in a unified geometric framework.