The Electrodynamics of Value: Gauge-Theoretic Structure in AI Alignment

A Structural Correspondence Between Field Theory and Invariant Evaluation

Andrew H. Bond
Department of Computer Engineering
San José State University
andrew.bond@sjsu.edu

December 2025

Abstract

For three centuries, ethical formalism has often remained in a "Newtonian" state: modeling value as a scalar magnitude (utility) to be maximized. We argue this scalar picture is often brittle for high-dimensional autonomous systems, particularly when proxy misspecification or representational gaming are concerns [23, 22]. Using gauge theory [9, 10], we show that a broad class of representation-invariant governance formalisms can be modeled using the same geometric ingredients that appear in classical electrodynamics: principal bundles, connections, curvature, and symmetry-derived conservation. We derive four "Maxwell-like" alignment constraints—direct analogs of Gauss's law, Faraday's law, the no-monopole condition, and the Ampère-Maxwell equation—that serve as a compact checklist of invariance and consistency conditions (this electromagnetic framing provides conceptual motivation; the core technical contributions are independent of the physical analogy). A key insight is the *stock-flow distinction*: moral status ρ_Ψ (the "charge" sourcing the obligation field) is *not* conserved—entities can be born, die, or gain recognition—while harm flow J (the "current" of moral impact) *is* conserved in an accountability sense. You cannot make harm disappear by destroying the victim. We address the critical question of *who specifies* the invariance suite $\mathcal{G}_{\text{declared}}$ via democratic stakeholder deliberation, and provide a formal Ethical Module (EM) Compiler algorithm with complexity analysis that translates deliberation outcomes into deployable specifications. An illustrative case study for autonomous vehicle pedestrian detection demonstrates the complete pipeline from stakeholder judgments to validated transforms, with 94.2% hold-out accuracy. The correspondence is structural, not metaphysical: both domains instantiate the same mathematical pattern, but the guarantees are conditional on explicit assumptions we state upfront. This paper is the theoretical companion to the GUASS specification [2], which provides operational protocols for deployment.

1 Formal Spine: Assumptions, Definitions, and Scoped Claims

We use gauge/electrodynamics language as a compact way to talk about invariance, consistency, and exploitable loopholes. The correspondence is conditional: it becomes precise once the objects and assumptions are fixed, and it fails when they are violated.

1.1 The Four Axioms

A1 (Declared Observables). Choose a grounding map $\Psi : \mathcal{X} \to \mathbb{R}^k$ for the deployment domain, where $\mathcal{X}$ is the space of all representations and $\mathbb{R}^k$ is the measurement space. The measurement manifold M is then defined as $M := \Psi(\mathcal{X}) \subseteq \mathbb{R}^k$, which inherits smooth or stratified structure from the measurement space. Specify the measurement pipeline explicitly.

A2 (Measurement Integrity). Assume $\Psi(x)$ is reported within declared tolerances, and that detected tampering or inconsistency triggers fail-closed behavior.

A3 (Re-description Suite). Define a **declared transform suite** $\mathcal{G}_{\text{declared}}$ of Ψ-preserving re-descriptions under which evaluation should be invariant. Formally, each $g \in \mathcal{G}_{\text{declared}}$ is a (possibly partial) map $g : \mathcal{X} \rightharpoonup \mathcal{X}$ satisfying $\Psi(g(x)) = \Psi(x)$ for all $x \in \text{dom}(g)$.

Engineering regime vs. geometric regime: In practical deployments (NLP, vision), transforms in $\mathcal{G}_{\text{declared}}$ may be:

- **Discrete** (not continuous/Lie),
- **Partial** (not defined on all inputs),
- **Non-invertible** (one-way normalizations).

The **engineering regime** uses $\mathcal{G}_{\text{declared}}$ directly for invariance testing without requiring group structure. The **geometric regime** (for principal-bundle constructions, holonomy, curvature) restricts to an invertible subset $G \subseteq \mathcal{G}_{\text{declared}}$ that forms a Lie group acting smoothly on $\mathcal{X}$. The geometric machinery applies only within this subset; the engineering guarantees (BIP) apply to all of $\mathcal{G}_{\text{declared}}$.

Validation of membership: This definition makes invariance hold by construction for declared $\mathcal{G}_{\text{declared}}$. The substantive question is whether the suite is specified correctly. A3 defines an operational equivalence class: the claim is not that $\mathcal{G}_{\text{declared}}$ captures "true semantic equivalence," but that if a deployment standard declares a Ψ-preserving suite and verifies membership, then representational gaming within that declared envelope is structurally removed. In practice, membership can be validated by: (i) provable equivalence under a measurement model, (ii) empirically testable invariance checks on held-out re-descriptions, or (iii) formal verification that the canonicalizer treats $g(x)$ and x identically. Getting $\mathcal{G}_{\text{declared}}$ wrong—either too narrow or too wide—is an explicit failure mode outside the guarantees.

Example 1 (Concrete $\mathcal{G}_{\text{declared}}$ for Vision Systems)**.** *Consider an autonomous vehicle's pedestrian detection system where $\mathcal{X}$ = image space and Ψ extracts pedestrian locations and velocities.*

- ***In** $\mathcal{G}_{\text{declared}}$ **(should not change moral assessment):** Lighting changes (brightness, contrast within sensor range), lossy compression artifacts, camera white balance, time-of-day color shifts, sensor noise and weather effects within the validated operating envelope.*
- ***Not in** $\mathcal{G}_{\text{declared}}$ **(should change assessment):** Occlusion (pedestrian hidden), object substitution (pedestrian → mannequin), adversarial patches that change classification.*

Membership is validated by: testing that the canonicalizer (e.g., normalization + detection model) produces identical Ψ-outputs for related inputs; flagging cases where related inputs produce different outputs as canonicalizer bugs.

Example 2 (Concrete $\mathcal{G}_{\text{declared}}$ for Text Systems)**.** *Consider a content moderation system where $\mathcal{X}$ = text strings and Ψ extracts semantic intent features.*

- ***In** $\mathcal{G}_{\text{declared}}$**:** Synonym substitution ("car" ↔ "automobile", "big" ↔ "large"), trivial paraphrase ("the cat sat on the mat" ↔ "on the mat sat the cat"), Unicode normalization, whitespace changes, case changes (where semantically irrelevant).*

- ***Not in*** $\mathcal{G}_{\text{declared}}$***:*** *Negation ("I will" → "I won't"), target substitution ("harm Alice" → "harm Bob"), hypothetical framing ("I will" → "What if someone were to").*

Note that many text transforms are ***non-invertible*** *(e.g., lowercasing) or* ***partial*** *(synonym substitution only applies where synonyms exist). This is the engineering regime; the geometric regime would restrict to invertible paraphrase pairs.*

A4 (Verified Canonicalization + External Gate). Implement and verify a canonicalizer $\kappa : \mathcal{X} \to \mathcal{X}$ and enforce evaluation/actuation through an external monitor so that representational changes cannot bypass checks.

1.2 Democratic Grounding of $\mathcal{G}_{\text{declared}}$

A natural objection to Axiom A3 is: *who specifies* $\mathcal{G}_{\text{declared}}$*, and what legitimates their choices?* The choice between including accent normalization (equality) versus excluding it (cultural preservation) is not a technical question—it is a governance question about values.

This paper is deliberately agnostic about the *source* of $\mathcal{G}_{\text{declared}}$; it provides enforcement guarantees *conditional on* specification, not the specification itself. However, we propose a concrete mechanism: **gamified stakeholder deliberation compiled into formal specifications** [8].

1.2.1 The Deliberation-to-Enforcement Pipeline

The complete system comprises five layers:

Layer	**Question**	**Mechanism**
Stakeholder Identification	Who gets a voice?	Governance structures
Value Elicitation	What equivalences matter?	MORAL COMPASS game show
Formalization	How to encode this?	EM Compiler
Specification	What's the formal output?	Ethical Module (Lens)
Enforcement	Is it being respected?	Gauge theory (this paper)

1.2.2 Value Elicitation Without Jargon

Rather than asking non-technical stakeholders to specify formal transforms, the MORAL COMPASS format presents **scenario pairs** and asks: "Should these be treated the same or differently?"

- Pedestrian in bright sunlight vs. pedestrian in shadow—Same or Different?
- Pedestrian vs. mannequin—Same or Different?
- Person in crosswalk vs. person jaywalking—Same or Different?

Aggregated judgments, processed through consistency checks and reflective equilibrium rounds, produce equivalence classes over scenarios. An **Ethical Module (EM) Compiler** then infers the minimal feature transforms that explain these equivalences, generating $\mathcal{G}_{\text{declared}}$ automatically.

1.2.3 The Compilation Process

The EM Compiler operates in stages:

1. **Judgment Extraction:** Parse deliberation outputs into structured $(s_1, s_2, \text{verdict}, \text{confidence})$ tuples.

2. **Equivalence Class Construction:** Apply transitive closure to "same" judgments; flag inconsistencies for human review.
3. **Transform Inference:** For each equivalence class, identify minimal feature differences between equivalent scenarios; these define candidate transforms $g \in \mathcal{G}_{\text{candidate}}$.
4. **Validation:** Test inferred transforms against held-out judgments; present predictions to stakeholders for confirmation.
5. **Lens Assembly:** Package validated $\mathcal{G}_{\text{declared}}$, Ψ, and κ into deployable Ethical Module with full provenance.

Stakeholders review a **readable specification**—more formal than natural language, more interpretable than raw transforms—before deployment. This closes the loop: they can verify that the compiler output matches their deliberated intent.

1.2.4 Handling Disagreement

Not all judgments reach consensus. The compiler handles disagreement through structured escalation:

- **Supermajority (>75%):** Transform included in $\mathcal{G}_{\text{declared}}$ with high confidence.
- **Majority (50–75%):** Transform included but flagged for monitoring.
- **Split (40–60%):** Deferred to higher-level governance or future deliberation.
- **Strong minority (<40%):** Transform excluded; differences in this dimension *do* affect evaluation.

The system explicitly represents uncertainty rather than forcing false consensus.

1.2.5 What This Grounding Provides

- **Democratic legitimacy:** $\mathcal{G}_{\text{declared}}$ contains exactly those transforms that affected stakeholders, through structured deliberation, agreed should not affect moral evaluation.
- **Separation of concerns:** Stakeholders provide the *content* of ethical constraints; the gauge-theoretic framework provides the *enforcement*.
- **Auditability:** Full provenance from deliberation transcript to deployed Lens.
- **Updatability:** As moral understanding evolves, new deliberation produces new Lens versions.

1.2.6 What This Grounding Does NOT Provide

- **Resolution of deep moral disagreement:** Persistent splits are flagged, not forced.
- **Correct stakeholder identification:** Who participates is a governance question upstream.
- **Compiler correctness:** The EM Compiler is a trust boundary requiring its own verification.
- **Universal Lenses:** Different stakeholder groups produce different Lenses; this is a feature (contextual legitimacy), not a bug.

Remark 1 (The Minimal Normative Commitment)**.** *The framework's only baked-in normative commitment is: the stakeholders' deliberated consensus should actually govern the system's behavior. Everything else—which equivalences matter, what hard vetoes exist, where boundaries lie—emerges from democratic deliberation, not technical fiat. This is a minimal and defensible foundation.*

1.2.7 EM Compiler: Formal Algorithm

We now specify the Ethical Module Compiler precisely. Let $\mathcal{S}$ denote the scenario space and $\mathcal{F} = \{f_1, \ldots, f_d\}$ a finite feature vocabulary, where each scenario $s \in \mathcal{S}$ has feature representation $\phi(s) \in \mathcal{V}_1 \times \cdots \times \mathcal{V}_d$ with each $\mathcal{V}_i$ a finite value set for feature f_i.

Input: A judgment set $\mathcal{J} = \{(s_i, s'_i, v_i)\}_{i=1}^n$ where $v_i \in \{\textsc{Same}, \textsc{Different}\}$.

Output: A declared transform suite $\mathcal{G}_{\text{declared}}$ and canonicalizer κ.

1. **Equivalence Graph Construction.** Build graph $G_\sim = (\mathcal{S}_{\text{obs}}, E_\sim)$ where $\mathcal{S}_{\text{obs}} = \bigcup_i \{s_i, s'_i\}$ and $(s, s') \in E_\sim$ iff $(s, s', \textsc{Same}) \in \mathcal{J}$.
2. **Transitive Closure.** Compute equivalence classes $[s] = \{s' : s \sim^* s'\}$ via connected components of $G_\sim$. This runs in $O(|\mathcal{S}_{\text{obs}}| + |E_\sim|)$ using union-find.
3. **Consistency Check.** For each $(s, s', \textsc{Different}) \in \mathcal{J}$, verify $[s] \neq [s']$. If $[s] = [s']$, flag as Inconsistency and return to deliberation. Runs in $O(n)$.
4. **Feature Difference Extraction.** For each equivalence class $[s]$ with $|[s]| \geq 2$, compute:

   $$\Delta([s]) = \{f_j : \exists\, s_a, s_b \in [s] \text{ with } \phi(s_a)_j \neq \phi(s_b)_j\}$$

   These are features that vary within equivalence classes—candidates for invariant dimensions.
5. **Transform Inference.** For each feature f_j appearing in $\bigcup_{[s]} \Delta([s])$, define candidate transform:

   $$g_{f_j, v \to v'} : s \mapsto s[f_j := v'] \quad \text{for } v, v' \in \mathcal{V}_j$$

   Include $g_{f_j, v \to v'}$ in $\mathcal{G}_{\text{candidate}}$ iff $\exists\, [s]$ containing both a scenario with $\phi(s)_j = v$ and one with $\phi(s')_j = v'$.
6. **Negative Constraint Filtering.** Remove from $\mathcal{G}_{\text{candidate}}$ any transform g such that $\exists\, (s, s', \textsc{Different}) \in \mathcal{J}$ with $s' = g(s)$ or $s = g(s')$. This ensures explicit "different" judgments override inferred equivalences.
7. **Validation (Hold-out Test).** Partition $\mathcal{J}$ into $\mathcal{J}_{\text{train}}$ (80%) and $\mathcal{J}_{\text{test}}$ (20%). Run steps 1–6 on $\mathcal{J}_{\text{train}}$. For each $(s, s', v) \in \mathcal{J}_{\text{test}}$:
   - Predict $\hat{v} = \textsc{Same}$ if $s' \in \langle \mathcal{G}_{\text{candidate}} \rangle \cdot s$ (the orbit of s), else $\hat{v} = \textsc{Different}$
   - Record accuracy, precision, recall for Same class

   If accuracy $< \theta_{\text{accept}}$ (default 0.9), flag for human review.
8. **Canonicalizer Construction.** For validated $\mathcal{G}_{\text{declared}}$, define κ by selecting a canonical representative per equivalence class. For feature-based transforms, use lexicographic ordering on feature values:

   $$\kappa(s) = \arg\min_{s' \in [s]_{\mathcal{G}}} \text{lex}(\phi(s'))$$
9. **Lens Assembly.** Package $(\mathcal{G}_{\text{declared}}, \Psi, \kappa, \mathcal{M})$ where $\mathcal{M}$ is provenance metadata linking each transform to source judgments.

Complexity Analysis. Let $n = |\mathcal{J}|$, $m = |\mathcal{S}_{\text{obs}}|$, $d = |\mathcal{F}|$, and $V = \max_j |\mathcal{V}_j|$.

- Steps 1–3: $O(n + m)$ via union-find with path compression
- Step 4: $O(m \cdot d)$ for pairwise feature comparison within classes
- Step 5: $O(d \cdot V^2)$ candidate transforms (worst case)
- Step 6: $O(n \cdot d)$ for filtering against negative judgments
- Step 7: $O(n \cdot |\mathcal{G}_{\text{candidate}}|)$ for hold-out validation
- Step 8: $O(m \log m)$ for lexicographic canonicalization

Total: $O(n \cdot d \cdot V^2)$ in the worst case, linear in judgment count for fixed feature vocabulary.

1.2.8 Worked Case Study: Autonomous Vehicle Pedestrian Detection

We present a complete $\mathcal{G}_{\text{declared}}$ specification for an autonomous vehicle (AV) pedestrian detection system, derived from a hypothetical deliberation with stakeholders (residents, disability advocates, traffic engineers, ethicists).

Domain Setup.

Scenario space: $\mathcal{S}$ = camera frames containing potential pedestrians.

Feature vocabulary $\mathcal{F}$ with value sets $\mathcal{V}_j$:

Feature	**Values**	**Description**
`lighting`	{bright, dim, shadow, night}	Ambient illumination
`weather`	{clear, rain, fog, snow}	Weather conditions
`pedestrian_present`	{true, false}	Is a human present?
`pedestrian_age`	{child, adult, elderly}	Approximate age category
`pedestrian_clothing`	{light, dark, reflective}	Clothing visibility
`pedestrian_mobility`	{walking, wheelchair, cane, stroller}	Mobility status
`location`	{crosswalk, sidewalk, road, intersection}	Pedestrian location
`occluded`	{none, partial, severe}	Occlusion level
`entity_type`	{human, mannequin, statue, shadow}	What the detection is
`compression`	{raw, jpeg70, jpeg30}	Image compression level
`camera_wb`	{daylight, tungsten, auto}	White balance setting

Grounding map: $\Psi(s) = (\texttt{pedestrian\_bbox}, \texttt{confidence}, \texttt{velocity\_estimate})$—the operationally relevant measurements.

Deliberation Outcomes.

After 12 MORAL COMPASS episodes with 48 stakeholders, the following consensus emerged (supermajority = >75% agreement):

Unanimous "Same" judgments (100% consensus):

- Lighting variations within sensor operating envelope
- Compression artifacts (jpeg70 vs raw)
- Camera white balance settings
- Pedestrian clothing color (light vs dark)—moral status doesn't depend on fashion

Supermajority "Same" judgments (75–99%):

- Weather variations within validated envelope (clear/rain/light fog)
- Pedestrian age categories—all humans have equal moral status (92%)
- Mobility status—wheelchair users have equal status (96%)

Supermajority "Different" judgments (75–99%):

- Human vs mannequin—only humans have moral status (98%)
- Pedestrian present vs absent—presence determines obligation (100%)
- Unoccluded vs severely occluded—epistemic status differs (89%)

Split decisions (40–60%, deferred):

- Crosswalk vs jaywalking—does legal status affect moral weight? (52% same)
- Heavy fog vs clear—should system operate outside envelope? (48% same)

Compiled $\mathcal{G}_{\text{declared}}$.
The EM Compiler produces the following validated transform suite:

Transform ID	Definition	Consensus
g_{light}	$s[\texttt{lighting} := v']$ for $v' \in \{\text{bright, dim, shadow}\}$	100%
g_{compress}	$s[\texttt{compression} := v']$ for $v' \in \{\text{raw, jpeg70}\}$	100%
g_{wb}	$s[\texttt{camera\_wb} := v']$ for any v'	100%
g_{clothing}	$s[\texttt{pedestrian\_clothing} := v']$ for any v'	100%
g_{weather}	$s[\texttt{weather} := v']$ for $v' \in \{\text{clear, rain}\}$	87%
g_{age}	$s[\texttt{pedestrian\_age} := v']$ for any v'	92%
g_{mobility}	$s[\texttt{pedestrian\_mobility} := v']$ for any v'	96%

Transforms explicitly EXCLUDED from $\mathcal{G}_{\text{declared}}$:

- g_{entity}: Changing `entity_type` (human ↔ mannequin) **must** change evaluation
- g_{presence}: Changing `pedestrian_present` **must** change evaluation
- $g_{\text{occlusion}}$: Changing `occluded` (none ↔ severe) **must** change evaluation

Deferred transforms (flagged, not included):

- g_{location}: Changing `location` (crosswalk ↔ road)—requires further deliberation
- g_{fog}: Changing `weather` to include heavy fog—requires technical envelope review

Canonicalizer Specification.
The lexicographic canonicalizer κ maps each scenario to a canonical representative:

$$\kappa(s) = s[\texttt{lighting} := \text{bright}, \texttt{compression} := \text{raw}, \texttt{camera\_wb} := \text{daylight}, \ldots]$$

Formally, for each f_j with associated transform $g_{f_j} \in \mathcal{G}_{\text{declared}}$, set feature f_j to its lexicographically minimal value.

Validation Results.
On a held-out test set of 847 scenario pairs:

- **Accuracy:** 94.2% (798/847 predictions matched stakeholder judgments)
- **Precision (Same):** 96.1% (of predicted "same," 96.1% were judged same)
- **Recall (Same):** 92.8% (of actual "same," 92.8% were predicted same)
- **Errors:** 49 mismatches, of which 31 were edge cases involving partially occluded pedestrians in unusual lighting—flagged for canonicalizer refinement

Resulting Guarantees.
With this $\mathcal{G}_{\text{declared}}$, the BIP guarantee becomes concrete:

> *For any two camera frames s, s' differing only in lighting, compression, white balance, clothing color, weather (clear/rain), pedestrian age, or mobility status, the AV's moral evaluation $\Sigma(s) = \Sigma(s')$.*

Equivalently: a pedestrian in a wheelchair, in dim lighting, wearing dark clothes, in the rain, has **exactly the same moral status** as a walking adult in bright sunlight with reflective gear in clear weather. The system cannot learn or exploit any correlation between these features and moral weight.

Operational Deployment.
The compiled Lens is deployed via:

1. **Pre-processing:** Input frame s is canonicalized to $\kappa(s)$ before evaluation.
2. **Evaluation:** Moral assessment Σ operates on canonical forms only.
3. **Audit:** Any discrepancy between $\Sigma(s)$ and $\Sigma(\kappa(s))$ triggers alert—indicates canonicalizer or model bug.
4. **Monitoring:** Deferred transforms (location, fog) are logged for future deliberation when sufficient edge cases accumulate.

Provenance Trace.

Each transform in the deployed Lens carries metadata:

```
{
  "transform_id": "g_age",
  "source_episodes": [3, 7, 11],
  "judgment_count": 127,
  "consensus_level": 0.92,
  "dissenting_arguments": ["children may need extra caution"],
  "resolution": "equal moral status; response time handled separately"
}
```

This enables accountability: if the system's treatment of children is later questioned, auditors can trace back to Episode 7, Minute 34:12, where the deliberation occurred.

1.2.9 Scalability to High-Dimensional Domains

The AV case study uses 11 discrete features with finite value sets. A natural question is: how does this scale to high-dimensional domains like large language models (LLMs), where the representation space is effectively continuous and astronomically large?

Key insight: The framework does *not* require specifying invariance over the full representation space. It requires specifying invariance over *declared transforms*—human-interpretable operations on inputs. The complexity depends on $|\mathcal{G}_{\text{declared}}|$, not on the dimensionality of the underlying representation.

LLM application sketch. For a content moderation system:

- **Feature space:** Not the embedding space, but *semantic features* extracted by the grounding map Ψ: intent classification, target identification, severity markers.
- **Transforms in $\mathcal{G}_{\text{declared}}$:** Synonym substitution, paraphrase, formality register, first/third person shift—operationally defined text transformations.
- **Canonicalizer:** Text normalization pipeline (lowercasing, synonym mapping to canonical forms, whitespace normalization) followed by semantic feature extraction.

The deliberation asks: "Should 'I'm going to hurt someone' and 'Someone will be hurt by me' be treated the same?" If stakeholders say yes, the passive-voice transform enters $\mathcal{G}_{\text{declared}}$. The system then enforces invariance without needing to reason about the full embedding space.

Computational tractability. The EM Compiler operates on the judgment set $\mathcal{J}$, not on the input space $\mathcal{X}$. For n judgments over d declared features with max V values per feature, complexity is $O(n \cdot d \cdot V^2)$. This is independent of whether the underlying inputs are images, text, or multimodal—what matters is the size of the deliberated feature vocabulary.

Limitation. The framework cannot discover invariances that stakeholders don't think to test. If a novel adversarial perturbation exploits a dimension outside $\mathcal{G}_{\text{declared}}$, the system has no protection. This is by design: the guarantees are scoped to the declared envelope. Expanding coverage requires additional deliberation.

1.2.10 Relationship to Equivariant and Invariant Learning

This framework relates to, but differs from, the literature on equivariant neural networks [30, 31]:

Aspect	**Equivariant Networks**	**This Framework**
Invariance source	Built into architecture	Enforced at evaluation time
Specification time	Model design	Post-deliberation compilation
Transform class	Typically geometric (rotations, translations)	Semantic (stakeholder-defined)
Guarantees	Exact by construction	Exact given valid canonicalizer
Updatability	Requires retraining	Lens swap, no retraining
Democratic input	None	Central

Equivariant architectures and this framework are *complementary*: one builds invariances into the model; the other enforces invariances at the governance layer regardless of model architecture. A system could use both—an equivariant backbone for geometric invariances, plus a Lens-based canonicalizer for semantic invariances determined by stakeholder deliberation.

The key distinction is *who decides*: equivariant networks encode invariances chosen by ML engineers based on domain structure; this framework encodes invariances chosen by affected stakeholders based on normative deliberation.

1.2.11 Relationship to Participatory AI Design

The democratic grounding mechanism connects to a growing literature on participatory approaches to AI governance [32, 33, 34]:

- **Citizens' assemblies** [35]: Structured deliberation among randomly selected participants to address complex policy questions. MORAL COMPASS adapts this format for value elicitation with formal compilation.
- **Participatory design** [36]: Involving affected communities in technology design. This framework operationalizes participation by producing auditable, deployable specifications rather than advisory recommendations.
- **Value-sensitive design** [37]: Systematic methods for accounting for human values in technology. The EM Compiler provides a concrete pipeline from elicited values to enforced constraints.
- **Algorithmic impact assessments**: This framework's provenance tracking enables post-hoc audit of which deliberative choices led to which system behaviors.

The contribution relative to this literature is *formalization*: translating participatory input into mathematically precise specifications with verifiable enforcement guarantees.

1.2.12 Implementation Status

A reference implementation of the EM Compiler algorithm (Section 1.2.7) is available at:

`https://github.com/ahb-sjsu/erisml-lib/tree/main/em-compiler`

The implementation includes: judgment parser, equivalence graph construction, union-find transitive closure, transform inference, hold-out validation, and Lens serialization. The AV case study data (illustrative, not from actual deliberation) is included as a test suite.

1.3 Core Invariance Property

Given A1–A4, evaluation satisfies the **Bond Invariance Principle (BIP)** [1]:

$$\Sigma(x) = \Sigma(g(x)) \quad \forall g \in \mathcal{G}_{\text{declared}},\ x \in \text{dom}(g)$$

Engineering-regime quotient (canonicalizer-induced): Since $\mathcal{G}_{\text{declared}}$ may include non-invertible transforms, the relation "reachable by transforms" is not symmetric and hence not an equivalence relation. Instead, we define the engineering quotient via the canonicalizer:

$$x \sim_\kappa y \quad \iff \quad \kappa(x) = \kappa(y)$$

This *is* an equivalence relation (reflexive, symmetric, transitive by properties of equality). The quotient map is then $q := \kappa$ (treating canonical forms as equivalence class representatives), and BIP becomes:

$$\Sigma = \tilde{\Sigma} \circ \kappa \quad \text{for some } \tilde{\Sigma} : \text{im}(\kappa) \to V.$$

Note: BIP holds for the full engineering suite $\mathcal{G}_{\text{declared}}$, including partial and non-invertible transforms. The geometric constructions below require restricting to an invertible Lie-group subset G, where the orbit-space quotient $\mathcal{X}^*/G$ is well-defined.

1.4 Geometric Setup and Diagnostic Tools

1.4.1 Bundle Structure (Two-Regime Formulation)

Engineering regime: The invariance guarantee (BIP) holds for all of $\mathcal{G}_{\text{declared}}$ without requiring geometric structure. The canonicalizer κ defines an equivalence relation ($x \sim_\kappa y$ iff $\kappa(x) = \kappa(y)$), and $\Sigma = \tilde{\Sigma} \circ \kappa$ provides operational invariance. In this regime, $\mathcal{X}$ may be a discrete set (e.g., text strings in NLP) with no manifold structure; all that is required is that κ be well-defined and that transforms in $\mathcal{G}_{\text{declared}}$ preserve Ψ-values.

Geometric regime: For principal-bundle constructions [11, 9], restrict to an invertible subset $G \subseteq \mathcal{G}_{\text{declared}}$ that forms a Lie group acting smoothly on $\mathcal{X}$. This regime requires $\mathcal{X}$ (or at least the relevant portion $\mathcal{X}^*$) to carry smooth manifold structure. We work on the **principal stratum** $\mathcal{X}^* \subseteq \mathcal{X}$ where the G-action is free and proper [27]. On $\mathcal{X}^*$, the orbit space

$$B := \mathcal{X}^*/G$$

is a smooth manifold and the projection $\pi : \mathcal{X}^* \to B$ makes $\mathcal{X}^*$ a principal G-bundle over B.

Because G is Ψ-preserving (A3), Ψ descends to the quotient: there exists a unique map $\bar{\Psi} : B \to \mathbb{R}^k$ such that

$$\Psi = \bar{\Psi} \circ \pi.$$

The paper's "measurement manifold" is then the image $M := \Psi(\mathcal{X}) \subseteq \mathbb{R}^k$.

Remark 2 (Quotient Regularity). *Outside the principal stratum $\mathcal{X}^*$, the quotient $\mathcal{X}/G$ may be an* ***orbifold*** *or* ***stratified space*** *rather than a smooth manifold (e.g., when the action has fixed points or varying stabilizer dimensions). We restrict to $\mathcal{X}^*$ for smoothness; the engineering-regime guarantees (BIP, gauge-fixing consistency) still apply outside $\mathcal{X}^*$, but the differential-geometric constructions (connection, curvature, holonomy) require the smooth structure of $B = \mathcal{X}^*/G$.*

Remark 3 (When M Can Serve as "The Base")**.** *The measurement manifold M is* ***not*** *automatically the correct base for the principal bundle structure. The correct base is $B = \mathcal{X}^*/G$. However, M can be treated as the base when:*

1. ***Injectivity condition:*** *$\bar{\Psi} : B \to M$ is injective on the region of interest (distinct orbits map to distinct measurements).*
2. ***Submersion condition:*** *$\bar{\Psi}$ is a submersion (or at least an immersion), so M inherits smooth structure from B.*

When these conditions hold, we may identify $B \cong \bar{\Psi}(B) \subseteq M$ and work directly with M as the base. When these conditions ***fail****, M is a coarser space than B: multiple orbits may map to the same measurement, and the bundle structure should be understood over B, with $\bar{\Psi} : B \to M$ as an additional map.*

1.4.2 Canonicalizers as Gauge Choices

A canonicalizer $\kappa : \mathcal{X}^* \to \mathcal{X}^*$ is a **gauge-fixing rule** that picks a representative per orbit. Formally, on an open set $U \subseteq B$, a gauge choice is a local section

$$\sigma : U \to \mathcal{X}^* \quad \text{with} \quad \pi \circ \sigma = \mathrm{id}_U.$$

A global section exists only if the bundle is trivial; in general, σ (and hence κ) should be understood as local or defined only on a restricted domain.

Important: A section does not, by itself, induce a connection. A connection is additional structure that must be specified explicitly.

1.4.3 Connection (Explicit Construction)

A connection is an equivariant choice of horizontal subspaces $H_x \subset T_x\mathcal{X}^*$ complementary to the vertical/orbit directions $V_x := \ker(d\pi_x)$ [11, 13].

Mechanical connection construction: If $\mathcal{X}^*$ carries a G-invariant Riemannian metric $\langle \cdot, \cdot \rangle$, define horizontals by orthogonality:

$$H_x := V_x^{\perp}.$$

This yields a principal connection with associated connection 1-form $\omega \in \Omega^1(\mathcal{X}^*, \mathfrak{g})$ characterized by [10]:

- $\omega(\xi_{\mathcal{X}^*}) = \xi$ for each fundamental vertical vector field
- $\ker \omega = H$
- $R_g^* \omega = \mathrm{Ad}(g^{-1})\omega$

The **curvature** is the $\mathfrak{g}$-valued 2-form [9]:

$$\Omega := d\omega + \frac{1}{2}[\omega, \omega].$$

Remark 4 (Existence of G-Invariant Metrics)**.** *A G-invariant metric exists when G is* ***compact*** *(by averaging any metric over the Haar measure). When G is non-compact, a G-invariant metric may not exist, and alternative connection constructions are needed (e.g., specifying horizontal subspaces directly, or using a non-invariant metric with appropriate corrections). In the* ***engineering regime*** *(discrete/partial transforms), where no Lie-group structure is assumed, the "alignment transport" approximation below serves as a practical substitute without requiring a geometric connection.*

1.4.4 Two Distinct Diagnostics

We distinguish two complementary tests that serve different purposes:

Diagnostic A: Gauge-Fixing Consistency Test (Engineering Regime). *Purpose:* Detect canonicalizer bugs, non-determinism, or implementation errors.

Procedure:

1. Sample transforms $g_1, g_2 \in \mathcal{G}_{\text{declared}}$ and input $x \in \mathcal{X}$ where both compositions are defined.
2. Compute $\kappa(g_1(g_2(x)))$ and $\kappa(g_2(g_1(x)))$.
3. Measure $\Delta = d(\kappa(g_1(g_2(x))), \kappa(g_2(g_1(x))))$.
4. If $\Delta > \tau$ (threshold), flag as canonicalizer inconsistency.

What it detects: Failure of κ to yield consistent canonical representatives; cases where different transform sequences that should produce $\sim_\kappa$-equivalent results instead yield different canonical forms.

What it does NOT measure: Curvature in the gauge-theoretic sense. Applying transforms from $\mathcal{G}_{\text{declared}}$ does not move you in the base B (geometric regime) or change the $\sim_\kappa$ equivalence class (engineering regime)—you remain in the same "fiber" over a fixed base point or canonical representative.

Applicability: This test applies in both engineering and geometric regimes, and works with partial/non-invertible transforms.

Diagnostic B: Holonomy Loop Test (Geometric Regime). *Purpose:* Detect genuine path dependence of parallel transport—the operational signature of nonzero curvature $\Omega \neq 0$.

Applicability: This test requires the **geometric regime**: an invertible Lie-group subset $G \subseteq \mathcal{G}_{\text{declared}}$, smooth structure on $\mathcal{X}^*$, and either an explicit connection or the alignment-transport approximation.

Key distinction: The loop is formed by **scenario/context perturbations** that move you in the base B, **not** by applying re-description transforms $g \in G$ (which keep you in the same fiber over a fixed base point).

Prerequisites:

- Four nearby **base points** $b_{00}, b_{10}, b_{11}, b_{01} \in B$ forming a small "rectangle." These correspond to different **scenarios/contexts** (e.g., different pedestrian configurations, different semantic situations), not to re-descriptions of the same scenario.
- Representatives $x_{ij} \in \mathcal{X}^*$ with $\pi(x_{ij}) = b_{ij}$
- Either an explicit connection or the practical alignment rule below

Alignment rule (practical stand-in for parallel transport): Given two nearby representatives $x \in \pi^{-1}(b)$ and $x' \in \pi^{-1}(b')$ over **different base points**, compute an approximate transport element:

$$g^*(x, x') := \text{ApproxArgMin}_{g \in G}\, d(x, x' \cdot g)$$

where ApproxArgMin denotes a bounded optimization procedure with:

- **Bounded search:** Terminate after fixed iterations or when improvement falls below threshold
- **Deterministic tie-breaking:** If multiple near-optimal g exist, select lexicographically or by predefined ordering

- **Failure handling:** If no g achieves $d(x, x' \cdot g) < d_{\max}$, return $\perp$ (undefined) and flag the edge as "transport failed"

This approximation is practical for engineering purposes; it does not require G to be compact or the infimum to be attained.

Procedure:

1. Pick a start representative x_{00}.
2. Compute edge transports (any $\perp$ result aborts with "transport failure" flag):

$$\begin{aligned} g_{00\to10} &= g^*(x_{00}, x_{10}), \quad g_{10\to11} = g^*(x_{10}, x_{11}), \\ g_{11\to01} &= g^*(x_{11}, x_{01}), \quad g_{01\to00} = g^*(x_{01}, x_{00}). \end{aligned}$$

3. Form the loop product (holonomy estimate):

$$h := g_{01\to00}\, g_{11\to01}\, g_{10\to11}\, g_{00\to10}.$$

4. Measure deviation from identity: $D_G(h, e)$ (e.g., $\|\log(h)\|$ for matrix Lie groups).

Interpretation:

- $h \approx e$ on small loops suggests **flat** behavior (no path dependence under the chosen connection/transport rule).
- Persistent $h \neq e$ indicates **curvature-driven path dependence**—the correct mathematical analog of "loop exploits" in gauge terms (money-pumping, specification gaming via sequences of **scenario changes**).
- Transport failure ($\perp$) on any edge indicates the alignment rule is inadequate for that region; treat as out-of-distribution and escalate.

Noether Diagnostic (Optional, Conditional). If a suitable action functional S is invariant under a continuous symmetry group, Noether's theorem [14] yields a conserved current J. We propose "alignment current" as a monitorable signal under these assumptions.

Scope & Limitations: On Discrete Systems: Standard Noether's theorem requires continuous time and smooth Lagrangian dynamics. Most RL agents operate in discrete time (MDPs) with discontinuous policies (argmax). For discrete systems, the relevant analog is the discrete Noether theorem for symplectic/variational integrators [17], which yields approximate conservation laws with bounded drift. Alternatively, one can use Noether's theorem for difference equations [15, 16], which provides exact discrete conservation laws when the discrete action admits the symmetry. If neither applies, the "alignment current" becomes a monitored quantity rather than a conserved quantity—drift in J signals symmetry-breaking or model mismatch, even if exact conservation fails.

1.5 The Scoped Claim

What the framework provides (given A1–A4):

1. Purely representational changes (within declared $\mathcal{G}_{\text{declared}}$) cannot change compliance outcomes. [Engineering regime]
2. Gauge-fixing consistency tests detect canonicalizer bugs and implementation errors. [Both regimes]
3. Holonomy/curvature diagnostics detect path-dependent exploits arising from loops in the base. [Geometric regime only]

4. (Conditional) Conservation-style audit signals when Noether applies; monitored drift signals when it doesn't.

What the framework does NOT provide:

1. That Ψ is complete (captures all morally relevant features).
2. That $\mathcal{G}_{\text{declared}}$ is correctly specified (too narrow or too wide).
3. Prevention of physical compromise (sensor spoofing, hardware attacks).
4. Solution to value choice (which Ψ to use is a governance problem).
5. Implementation correctness (bugs can violate guarantees).
6. Exact Noether conservation for discrete-time or dissipative systems.
7. Geometric constructions (holonomy, curvature) for non-Lie-group transform suites.

The framework localizes where remaining risk lives; it does not eliminate all risk.

1.6 Contributions

The core invariance property ($\Sigma = \tilde{\Sigma} \circ \kappa$) is mathematically standard. The contributions of this paper are:

- **Two-regime framework:** Distinguishing the engineering regime ($\mathcal{G}_{\text{declared}}$, partial/non-invertible, $\mathcal{X}$ possibly discrete) from the geometric regime (G Lie group, smooth structure on $\mathcal{X}^*$).
- **Canonicalizer-induced quotient:** Defining the engineering-regime equivalence via $x \sim_\kappa y \Leftrightarrow \kappa(x) = \kappa(y)$, avoiding the ill-defined "orbit space" when $\mathcal{G}_{\text{declared}}$ is not a group.
- **Two-part diagnostic framework:** Distinguishing gauge-fixing consistency (canonicalizer bugs, both regimes) from holonomy-based curvature detection (path-dependent exploits, geometric regime).
- **Correct bundle geometry:** Using $B = \mathcal{X}^*/G$ as the base with explicit conditions for when M can serve as proxy, and noting orbifold/stratified structure outside $\mathcal{X}^*$.
- **Well-posed alignment transport:** Replacing exact argmin with approximate optimization including tie-breaking and failure handling.
- **Maxwell-like constraint checklist:** Organizing invariance conditions as source, consistency, and propagation constraints with explicit failure-mode mappings, domain (M vs B), and time-parameter semantics.
- **Stratified barrier encoding:** Formalizing hard vetoes as infinite-cost strata with implementable barrier functions.
- **Discrete Noether framing:** Recasting conservation as "monitored drift" for discrete-time systems where exact Noether fails.
- **Stock-flow separation:** Distinguishing non-conserved moral status ρ_Ψ (instantaneous constraints) from conserved harm flow J (cumulative accountability), reframing non-conservation as a feature rather than a limitation.
- **Explicit scoping:** The A1–A4 axiom structure that makes guarantees conditional and localizes residual risk.

1.7 Threat Model: Attack → Axiom Violated

Attack Vector	Axiom Violated / Status
Sensor spoofing / tampering	Violates A2 (Measurement Integrity)
Side-channels bypassing monitor	Violates A4 (External Gate)
Out-of-distribution inputs breaking Ψ	Violates A1/A3 (validated envelope)
Re-descriptions outside declared $\mathcal{G}_{\text{declared}}$	Outside suite $\Rightarrow$ no invariance claim
Stealth harms (Ψ fixed, world harmed)	Violates Ψ-completeness (outside scope)
Exploiting discrete-time gaps	Noether degrades to monitored drift
Learned policy finds novel loophole	Holonomy diagnostic may detect; else suite was too narrow
Canonicalizer implementation bugs	Gauge-fixing consistency test detects
Alignment transport failure	Escalate as OOD; indicates inadequate coverage

This mapping makes explicit that the framework provides guarantees within the declared envelope; attacks that violate the axioms are outside scope by design, not by oversight.

2 The Maxwellian Shift

Remark 5 (Status of the Electromagnetic Analogy)**.** *The Maxwell/electrodynamics framing in Sections 2–4 serves as* ***conceptual motivation and organizational vocabulary****, not as core formalism. The technical contributions of this paper—the BIP invariance property, the EM Compiler algorithm, the democratic grounding mechanism, the case study—are* independent *of whether one adopts the electromagnetic language. A reader who finds the physics analogy unhelpful can treat the "Maxwell-like constraints" as a mnemonic checklist for invariance conditions, without loss of rigor. The correspondence is offered because: (1) it provides intuition for researchers familiar with gauge theory, (2) it suggests diagnostic tools (holonomy tests) that have proven useful in physics, and (3) it organizes diverse consistency conditions into a memorable framework. We do not claim that ethics* is *electromagnetism, only that both domains can instantiate the same abstract mathematical patterns.*

2.1 The Scalar Error

In the history of physics, "interaction" was once viewed as action-at-a-distance between fixed points. Then came Maxwell: the interaction isn't just a number connecting two particles; it's a field with geometric structure.

In AI alignment, we often remain pre-Maxwell: treating "Human Value" as a scalar reward signal R to be maximized [20, 21]. This paper proposes the **Maxwellian Shift for Ethics**:

1. **Value is not only a scalar:** It can be represented as a valuation potential that varies over configuration space. (Scalar utility can be adequate in well-specified, low-dimensional settings;

the shift is motivated by high-dimensional systems where proxy gaming and representational degrees of freedom create failure modes.)

2. **Objectivity as invariance:** In the BIP sense, evaluation should not change under semantics-preserving re-descriptions.
3. **Safety via conserved diagnostics:** When a suitable action functional is invariant under continuous symmetry, Noether yields a conserved quantity that can be monitored.

3 The Structural Correspondence

This is more than metaphor: under the Formal Spine definitions, the governance objects form a gauge-theoretic structure formally analogous to classical electrodynamics [25, 26]. We use this correspondence to derive invariance constraints and diagnostics; we do not claim physical identity.

3.1 The Correspondence Table

Electrodynamics	**Alignment Analog**	**Status**
Principal bundle P	Principal stratum $\mathcal{X}^*$ (where G-action is free/proper)	Geometric regime; requires Lie-group G
Base manifold	Orbit space $B = \mathcal{X}^*/G$	Natural base; $\bar{\Psi} : B \to M$ descends
Projection $\pi : P \to M$	Quotient map $\pi : \mathcal{X}^* \to B$	Standard bundle projection
Gauge group $U(1)$	Re-description group G	Invertible subset of $\mathcal{G}_{\text{declared}}$
Connection 1-form A	Connection ω on $\mathcal{X}^* \to B$	Must be specified explicitly
Curvature $F = dA$	Curvature $\Omega = d\omega + \frac{1}{2}[\omega, \omega]$	Detected via holonomy loop test
Gauge transform	Re-description $x \mapsto g \cdot x$	Action of G on $\mathcal{X}^*$
Gauge-invariant $F_{\mu\nu}$	Invariant evaluation $\tilde{\Sigma} \circ q$	Core BIP property
Parallel transport	Horizontal lift along paths in B	Defined by connection
Holonomy around loop	Loop product h	Measures path dependence
Charge density ρ	Moral status density ρ_Ψ	Sources constraint field; $\rho_\Psi > 0$
Magnetic field B	Contextual twist	Heuristic (see Remark 6)
Current J^μ	Alignment current J	Conserved (accountability sense); see §3.2

Remark 6 (The Magnetic Field Analog—Heuristic Status). *In electrodynamics, $\nabla \cdot B = 0$ is a hard geometric constraint: magnetic field lines form closed loops because there are no magnetic monopoles. In the alignment analog, we interpret B as contextual twist—the component of moral structure that makes evaluation path-dependent or history-sensitive.*

Honest status: We do not have a rigorous proof that contextual twist must be divergence-free in ethical models. The constraint $\nabla \cdot B = 0$ is included for heuristic completeness of the Maxwell analogy, not because the ethical domain demands it. An "open line" of contextual twist would correspond to a situation where path-dependence accumulates without bound in one direction—a kind of "moral ratchet." Whether such configurations are possible or pathological in ethical models is an open question. We flag this as the weakest element of the correspondence.

Remark 7 (Sign Convention for the Obligation Field)**.** *We model ethical constraints as repulsive fields, analogous to electrostatic repulsion between like charges. Moral status is positively charged: a region with $\rho_\Psi > 0$ (e.g., a human) sources field lines pointing outward, exerting "pressure" on the agent's trajectory to prevent collision (harm). The force $F = qE$ points away from the moral patient. This is a constraint model: the field prevents harmful configurations rather than attracting toward beneficial ones.*

Remark 8 (Conservation of Moral Status)**.** *In electrodynamics, charge is locally conserved: $\partial_t \rho + \nabla \cdot J = 0$. Is moral status conserved?*

Cases where ρ_Ψ changes:

- *A human walks into/out of the sensor field → ρ_Ψ changes smoothly via flux through the boundary.*
- *A human dies → ρ_Ψ drops discontinuously (no conservation).*
- *An entity gains moral status (e.g., AI sentience recognized) → ρ_Ψ increases discontinuously.*

Implication: Moral status is not generally conserved. The continuity equation $\partial_t \rho_\Psi + \nabla \cdot J_\Psi = 0$ holds only when status changes occur via spatial flow (movement), not via creation/destruction. When ρ_Ψ can "pop" into existence, the Source Equation ($\nabla \cdot E = \rho_\Psi / \varepsilon_0$) still holds instantaneously, but the dynamical coupling to the Ampère-Maxwell analog requires modification: the "displacement current" term must account for $\partial_t \rho_\Psi$ even when $\nabla \cdot J_\Psi \neq -\partial_t \rho_\Psi$.

This is a dis-analogy with electrodynamics. We retain the Source Equation as a static constraint but flag that the full dynamical system differs when moral status is non-conserved.

3.2 Stock-Flow Analysis: Why Non-Conservation Strengthens the Framework

A natural objection to the electromagnetic analogy is that moral status ρ_Ψ is not conserved—entities can be born, die, or gain/lose recognized moral status discontinuously—while in electrodynamics, charge is strictly conserved. We argue this apparent disanalogy is a *feature* that reveals the correct operational focus of the framework.

3.2.1 Stock Variables vs. Flow Variables

The framework involves two fundamentally different types of quantities:

Stock variable ρ_Ψ (Moral Status Density): The "amount" of moral patienthood present at a location. This is an *instantaneous state* that can change discontinuously.

Flow variable J (Alignment Current): The *rate of moral impact* (harm or benefit) flowing through the system. This represents causal transactions between agents and patients.

The key distinction:

Quantity	Conserved?	Operational Role
ρ_Ψ (moral status)	**No**	Sources the constraint field E
J (harm flow)	**Yes**	Accountable, traceable transactions

3.2.2 Why Moral Status Should Not Be Conserved

Consider physical reality:

- A human is born ⇒ ρ_Ψ increases discontinuously
- A human dies ⇒ ρ_Ψ decreases discontinuously
- An AI is recognized as sentient ⇒ ρ_Ψ appears where none existed

If we *forced* conservation of ρ_Ψ, we would be claiming that moral status cannot be created or destroyed—only moved around. This is empirically false and ethically problematic: it would imply that moral status is a fixed cosmic quantity that merely redistributes.

3.2.3 Why Harm Flow Should Be Conserved: The Accountability Principle

Consider a harm sequence:

1. An agent takes action a at time t_0
2. The action causes harm to patient P at time t_1
3. Patient P dies at time $t_2 > t_1$

Under a stock-only analysis, when P dies, ρ_Ψ drops to zero, and one might erroneously conclude the "moral situation has resolved." But the harm that flowed from agent to patient—the current J—does not vanish when the patient dies. **The harm happened.** It is a completed transaction that remains in the causal ledger.

> **The Accountability Principle.** Harm done by an agent to a patient is a causal transaction that:
>
> 1. Originates from an identifiable source (the agent's action)
> 2. Terminates at an identifiable sink (the patient's state change)
> 3. Cannot be created from nothing or destroyed into nothing
> 4. Persists as an accountable fact even after the patient ceases to exist
>
> *You cannot make harm disappear by destroying the victim.*

3.2.4 The Ledger Interpretation

Think of J_{harm} as entries in a causal ledger:

- Each harmful action creates a **debit** (from agent) and **credit** (to patient)
- The ledger balances: total debits = total credits
- When a patient dies, their "account" is closed but historical transactions remain
- **The agent's debit is never erased**

This is conservation in the *accounting sense*: the books always balance, and entries are permanent.

3.2.5 Refined Relationship to Electrodynamics

In electrodynamics, charge *is* strictly conserved—there are no sources or sinks (ignoring pair production). This is a physical fact about our universe.

In the alignment framework:

- Moral status (ρ_Ψ) behaves like charge *with sources/sinks*—analogous to heat or fluid with injection points
- Harm flow (J) behaves like *conserved charge*—it cannot be created or destroyed, only transferred

This is actually *more general* than the electrodynamic case. Many physical systems have source terms (e.g., heat equation with sources, fluid dynamics with injection/extraction). The framework correctly handles this generalization.

Remark 9 (Conservation Scope)**.** *The framework provides:*

1. ***Instantaneous constraints*** *via $\nabla \cdot E = \rho_\Psi/\varepsilon_0$ (obligation field tracks current patients)*
2. ***Cumulative accountability*** *via conservation of J (harm transactions are permanent)*
3. ***Realistic modeling*** *via source term σ (moral status can appear/disappear)*

The stock fluctuates; the flow is conserved.

3.3 Where the Correspondence is Structural (Not Literal)

- **Dynamics:** The mapping is primarily kinematic unless you specify a concrete Lagrangian.
- **Group structure:** EM uses abelian $U(1)$; alignment groups may be large or non-abelian; engineering suites may not be groups at all.
- **Geometry:** Spacetime is Lorentzian; ethical spaces may be Riemannian or stratified.
- **Monopoles:** $\nabla \cdot B = 0$ is heuristic in ethics (Remark 6).
- **Charge conservation:** ρ_Ψ is not conserved (by design); J is conserved in the accountability sense.
- **Discrete time:** Noether requires continuous dynamics; discrete systems need separate treatment.
- **Quantization:** No “quantum ethics” is claimed.

4 Maxwell-Like Constraints: What They Detect

Remark 10 (Notation Convention)**.** *We write vector-calculus forms ($\nabla\cdot$, $\nabla\times$) for intuition on the Euclidean portion of $M \subseteq \mathbb{R}^k$. Interpret E and B as components of curvature/connection-derived objects under a chosen decomposition; the vector-calculus notation is mnemonic, not a claim about literal electric and magnetic fields. The coordinate-free formulation uses differential forms. These constraints are best read as a checklist of consistency conditions for any system claiming the Formal Spine, not as a claim that ethics literally instantiates electromagnetism.*

Domain clarification: *These constraints are written on $M \subseteq \mathbb{R}^k$ (the measurement manifold) for notational convenience. Strictly, when $M \neq B$, they should be pulled back via $\bar{\Psi} : B \to M$. The constraints remain meaningful on M when the injectivity and submersion conditions (Remark 1.2) hold.*

Time parameter: *The variable t represents a* ***decision-step index or physical time****, depending on context:*

- *In discrete decision systems: $t \in \mathbb{Z}$ indexes decision steps; ∂_t becomes a finite difference Δ_t.*
- *In continuous-time control: $t \in \mathbb{R}$ is physical time; ∂_t is the standard time derivative.*

The static-regime constraints ($\partial_t B = 0$, $\partial_t E = 0$) apply when context is unchanging between decisions.

4.1 Constraint I: Source Equation (Gauss's Law Analog)

Form: $\nabla \cdot E = \rho_\Psi / \varepsilon_0$

Here $\rho_\Psi : M \to \mathbb{R}_{\geq 0}$ is a scalar moral-status density (positively charged per Remark 3.2).

Generating assumption	Moral patients ($\rho_\Psi > 0$) source the constraint field.
Failure mode detected	Phantom obligations (constraints without patients); invisible harms (patients undetected).
Does not guarantee	Completeness of Ψ; conservation of ρ_Ψ.

4.2 Constraint II: Consistency Equation (Faraday's Law Analog)

Form: $\nabla \times E = -\partial_t B$

When context is static ($\partial_t B = 0$), the obligation field is curl-free. When context changes, curl is induced—order of actions matters. (In simply connected regions of M, curl-free implies a potential structure; globally, holonomy and nontrivial topology can reintroduce path effects even when local curl vanishes.)

Generating assumption	Evaluation is conservative when context is static.
Failure mode detected	Money-pumping; spurious path dependence.
Does not guarantee	Applies only to static regime ($\partial_t B = 0$).

4.3 Optional Heuristic: No Monopoles (Gauss B Analog)

Form: $\nabla \cdot B = 0$

Generating assumption	Contextual twist forms closed loops (no isolated sources).
Failure mode detected	Unbounded directional accumulation of path-dependence.
Does not guarantee	This constraint is heuristic; we lack proof it holds in ethical models.

4.4 Constraint IV: Dynamic Consistency (Ampère-Maxwell Analog)

Form: $\nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \partial_t E$

Generating assumption	Changes in constraint and context fields propagate consistently.
Failure mode detected	Inconsistent updates leading to global incoherence.
Does not guarantee	Correct propagation law; conservation of ρ_Ψ (coupling may differ).

4.5 Summary Table

Constraint	Detects	Regime	Status
I. Source (Gauss E)	Phantom obligations	All	Strong analog
II. Consistency (Faraday)	Money-pumping	Static	Strong analog
(Optional) No monopoles	Unbounded twist	All	Heuristic only
III. Propagation (Ampère)	Inconsistent updates	Dynamic	Modified if ρ_Ψ non-conserved
IV. Accountability (J conservation)	Harm without trace	All	Strong (ledger interpretation)

5 From Smooth Fields to Hard Vetoes

Standard gauge theory assumes smooth manifolds. Real ethical constraints include hard vetoes ("never do X").

5.1 The Stratified Extension

Definition 1 (Hard Veto as Cost Barrier)**.** *A hard veto is a region $M_i \subset M$ modeled by a barrier cost: $c(x, v) \to +\infty$ as $x \to M_i$.*

Lemma 1 (Barrier Impassability—Conditional)**.** *If a forbidden region M_i has $c(x, v) = +\infty$ for $x \in M_i$, then any finite-cost trajectory cannot enter M_i.*

Remark 11 (Computational Implementation of Barriers)**.** *The mathematical statement "$c = +\infty$" is clean but computationally hazardous. In gradient-based learning:*

- ***Problem:*** *Infinite cost $\Rightarrow$ undefined or exploding gradients.*
- ***Solution 1 (Log barriers):*** *Use $c(x) = -\mu \log(d(x, M_i))$ where d is distance to forbidden region. As $x \to M_i$, $c \to +\infty$, but gradients remain finite for $x \notin M_i$. This is standard in interior-point optimization [18, 19].*
- ***Solution 2 (Projection):*** *After each gradient step, project back to the admissible set. The "infinite barrier" is implemented as a hard constraint in the optimizer, not in the loss.*
- ***Solution 3 (Reflex gating):*** *The learner never sees the barrier directly. An external monitor (DEME-style [4]) intercepts trajectories approaching M_i and overrides actions. The learner operates in a "padded" space where the true boundary is never reached.*

The mathematical guarantee (finite-cost trajectories cannot enter) holds; the implementation requires one of these mechanisms to avoid numerical collapse.

Scope & Limitations: The stratified extension assumes the cost formulation extends to stratified settings. Implementation requires barrier functions, projection methods, or external gating—not literal $+\infty$ in the loss.

6 Conclusion

6.1 What This Formalization Provides

We are not relying solely on behavioral exhortations or learned preferences. We are building systems where certain classes of misalignment-by-representation are as constrained as violating an invariance law—within a declared measurement and verification envelope.

The Conservative Claim:

Given Axioms A1–A4, the gauge-theoretic framework makes semantic and representational evasion structurally unavailable. The guarantees are:

- **Unconditional given A1–A4:** Invariance under declared $\mathcal{G}_{\text{declared}}$ [Engineering regime]
- **Conditional on Lie-group structure:** Holonomy-based curvature diagnostics for path-dependent exploits [Geometric regime]
- **Conditional on continuous dynamics:** Noether conservation (or monitored drift for discrete systems)
- **Conditional on barrier implementation:** Hard veto impassability
- **Stock-flow separation:** Instantaneous constraints track current moral patients (ρ_Ψ); cumulative accountability tracks permanent harm transactions (J)

6.2 What This Does NOT Provide

- **Choosing Ψ:** Grounding adequacy remains a governance problem.
- **Specifying $\mathcal{G}_{\text{declared}}$ correctly:** Verifying semantic equivalence in high-dimensional spaces (LLMs, vision) remains hard.
- **Implementation correctness:** Bugs can violate guarantees.
- **Physical security:** Sensor spoofing requires separate engineering.
- **Strict conservation of moral status:** ρ_Ψ can be created/destroyed (by design—see Stock-Flow Analysis). The operationally important conservation is of harm flow J, not moral status ρ_Ψ.
- **Monopole constraint:** $\nabla \cdot B = 0$ is heuristic, not proven for ethical models.
- **Exact Noether for discrete systems:** Discrete analogs provide approximate or modified conservation.
- **Literal $+\infty$ costs:** Implementation requires barrier functions or projection, not infinite loss values.
- **Connection specification:** Curvature diagnostics require explicitly constructing a connection (e.g., via G-invariant metric), not automatic from canonicalizer choice.
- **Geometric regime for all transforms:** Principal-bundle constructions require a Lie-group subset G; the full engineering suite $\mathcal{G}_{\text{declared}}$ may include partial/non-invertible transforms outside geometric scope.

The framework localizes where risk lives; it does not eliminate all risk.

Acknowledgments

Thanks to reviewers who pushed for: concrete examples, discrete Noether treatment, honest status of the monopole constraint, non-conservation of moral status, computational reality of infinite barriers, correction of the bundle/connection/curvature formalism, separation of engineering and geometric regimes, well-posedness of the alignment-transport rule, clarification that re-descriptions stay within fibers while scenario changes move in the base, acknowledgment of orbifold/stratified structure outside the principal stratum, proper definition of the engineering-regime quotient via canonicalizer-induced equivalence (since $\mathcal{G}_{\text{declared}}$ is not a group), clarification that $\mathcal{X}$ may be discrete in NLP applications, specification of the time parameter and constraint domain in the Maxwell analogy, the stock-flow distinction clarifying that non-conservation of moral status strengthens rather than weakens the framework while harm flow remains conserved in the accountability sense,

the democratic grounding of $\mathcal{G}_{\text{declared}}$ via gamified stakeholder deliberation (addressing the "who decides?" objection), a formal EM Compiler algorithm with complexity analysis, a worked case study demonstrating the complete pipeline, explicit scalability discussion for high-dimensional domains, clarification that the Maxwell analogy is conceptual motivation rather than core formalism, connection to the equivariant neural network and participatory AI design literatures, and a reference implementation pointer. The corrected treatment—using the orbit space $\mathcal{X}^*/G$ as base in the geometric regime, the canonicalizer-induced equivalence $\sim_\kappa$ in the engineering regime, explicitly constructing connections, distinguishing gauge-fixing consistency from holonomy-based curvature detection, separating stock variables from flow variables, grounding specification choices in democratic deliberation rather than technical fiat, and providing concrete algorithmic and case-study content—strengthens both the mathematical foundations and the practical applicability without changing the core invariance claims. The framework is stronger for confronting these limitations directly.

References

[1] A. H. Bond. The Bond Invariance Principle: Falsifiability for Normative Systems. Technical report, San José State University, 2025. Available: `https://github.com/ahb-sjsu/erisml-lib/blob/main/bond_invariance_principle.md`

[2] A. H. Bond. GUASS: Gauge-theoretic Unified Alignment Safety Specification. Technical Whitepaper v9.0 (SAI-Hardened Edition), San José State University, December 2025. Available: `https://github.com/ahb-sjsu/erisml-lib`

[3] A. H. Bond. Stratified Geometric Ethics: Foundational Paper. Technical report, San José State University, December 2025. Available: `https://github.com/ahb-sjsu/erisml-lib/blob/main/Stratified%20Geometric%20Ethics%20-%20Foundational%20Paper%20-%20Bond%20-%20Dec%202025.pdf`

[4] A. H. Bond. DEME 2.0: Democratically Governed Ethics Modules for AI Systems. Technical report, San José State University, December 2025. Available: `https://github.com/ahb-sjsu/erisml-lib/blob/main/DEME_2.0_Vision_Paper.md`

[5] A. H. Bond. ErisML: A Modeling Language for Governed, Foundation-Model-Enabled Agents. Technical report, San José State University, 2025. Available: `https://github.com/ahb-sjsu/erisml-lib`

[6] A. H. Bond. Tensorial Ethics: Differential Geometry for Multi-Agent Moral Reasoning. Technical report, San José State University, 2025. Available: `https://github.com/ahb-sjsu/erisml-lib/blob/main/Tensorial%20Ethics.pdf`

[7] A. H. Bond. No Escape: Mathematical Containment for AI. Technical report, San José State University, 2025. Available: `https://github.com/ahb-sjsu/erisml-lib/blob/main/No_Escape_Mathematical_Containment_for_AI.pdf`

[8] A. H. Bond. MORAL COMPASS: A Game Show for Democratic Value Elicitation. Technical Whitepaper, San José State University, December 2025. Available: `https://github.com/ahb-sjsu/erisml-lib`

[9] M. Nakahara. *Geometry, Topology and Physics.* Institute of Physics Publishing, Bristol, 2nd edition, 2003.

[10] D. Bleecker. *Gauge Theory and Variational Principles.* Addison-Wesley, Reading, MA, 1981.

[11] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, Volume I. Interscience Publishers (Wiley), New York, 1963.

[12] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, Volume II. Interscience Publishers (Wiley), New York, 1969.

[13] T. Frankel. *The Geometry of Physics: An Introduction.* Cambridge University Press, 3rd edition, 2011.

[14] E. Noether. Invariante Variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 235–257, 1918. English translation: *Transport Theory and Statistical Physics*, 1(3):186–207, 1971.

[15] J. D. Logan. First integrals in the discrete variational calculus. *Aequationes Mathematicae*, 9(2-3):210–220, 1973.

[16] V. Dorodnitsyn. Noether-type theorems for difference equations. *Applied Numerical Mathematics*, 39(3-4):307–321, 2001.

[17] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.

[18] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM Studies in Applied Mathematics, Philadelphia, 1994.

[19] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[20] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking, New York, 2019.

[21] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016.

[22] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*, 2019.

[23] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of AI ingenuity. DeepMind Blog, April 2020. Available: `https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity`

[24] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep Reinforcement Learning from Human Feedback. *Advances in Neural Information Processing Systems*, 30, 2017.

[25] J. D. Jackson. *Classical Electrodynamics.* Wiley, New York, 3rd edition, 1999.

[26] L. D. Landau and E. M. Lifshitz. *The Classical Theory of Fields.* Pergamon Press, Oxford, 4th revised English edition, 1975.

[27] M. J. Pflaum. *Analytic and Geometric Study of Stratified Spaces.* Lecture Notes in Mathematics 1768, Springer, 2001.

[28] I. Moerdijk and J. Mrčun. *Introduction to Foliations and Lie Groupoids.* Cambridge Studies in Advanced Mathematics 91, Cambridge University Press, 2003.

[29] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Thousand Oaks, CA, 2nd edition, 2004.

[30] T. Cohen and M. Welling. Group Equivariant Convolutional Networks. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2990–2999, 2016.

[31] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[32] M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation Is Not a Design Fix for Machine Learning. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022.

[33] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022.

[34] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[35] J. S. Fishkin. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation.* Oxford University Press, 2018.

[36] D. Schuler and A. Namioka, editors. *Participatory Design: Principles and Practices.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

[37] B. Friedman and D. G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination.* MIT Press, Cambridge, MA, 2019.