Stratified Geometric Ethics and DEME 2.0: A Mathematical Moral Landscape for AI Systems

By: Andrew Bond, andrew.bond.@sjsu.edu

Modern AI systems are beginning to make decisions that affect life, rights, and welfare under tight time pressure: collision avoidance in self-driving cars, emergency stops in industrial robots, clinical triage in overcrowded hospitals, and so on. In these settings, we need *machine ethics at machine speed*: systems that can apply explicit ethical constraints in microseconds, while remaining transparent and accountable to humans.

Today's approaches fall into two unsatisfying camps. On one side, we have high-level principles and guidelines—lists of values like fairness, autonomy, and human rights that sound good but are not connected to concrete algorithms. On the other side, we have purely data-driven methods—reward functions, learned policies, or large language models—whose "ethics" are implicit in millions of parameters and cannot be reliably verified. Neither gives us what safety-critical deployment really needs: explicit, deterministic, *formally checkable* ethical reasoning embedded inside the control loop.

This project is an attempt to build that missing bridge. It has two parts:

- **Stratified Geometric Ethics (SGE)**: a mathematical framework that makes the idea of a "moral landscape" precise and proves that ethical reasoning can be both expressive and verifiable.
- **DEME 2.0**: a concrete architecture that uses SGE to implement a moral landscape inside real machines, complete with hardware "Ethics Modules" that can enforce constraints in the reflex band of control.

SGE and DEME were developed in reflective equilibrium: DEME commits to specific engineering choices, and SGE explains why those choices make sense mathematically and what their limits are.

---

1. A stratified moral landscape

The starting question is: *What kind of geometry do we need to represent real moral situations?*

Standard "geometric ethics" models imagine a smooth manifold where everything varies continuously. That picture breaks down as soon as we look at actual ethical phenomena:

- **Discrete choices**: in a trolley problem, you pull the lever or you don’t; there is no “0.4 of each track.”
- **Incommensurable values**: human life vs property cannot be traded off at any finite exchange rate.
- **Thresholds**: crossing from “no consent” to “genuine consent,” or from “letting die” to “killing,” feels like a qualitative jump, not a small perturbation.
- **Genuine dilemmas**: sometimes every option is morally bad, and that “moral residue” should not be smoothed away.

Smooth manifolds cannot represent all of this at once: everything is path-connected, trade-offs are always possible, and level sets are continuous.

SGE instead models the moral landscape as a **stratified space**: a union of smooth pieces (strata) of different dimensions, glued along boundaries. You can think of this as a terrain with plateaus, cliffs, and ridges:

- The interior of a region might represent mixed or probabilistic allocations of resources.
- Edges represent trade-offs between two concrete options.
- Vertices represent discrete, irreversible actions (e.g., “treat this patient now”).

Moving from one stratum to another can be a genuine ethical jump. SGE shows that, among standard geometric tools (smooth manifolds, manifolds with corners, cell complexes), stratified spaces are the natural minimal candidates that can represent all four kinds of phenomena at once.

---

2. Obligations, interests, and satisfaction

Once we have a moral landscape, we need a way to talk about “oughts,” “interests,” and how well an action satisfies them.

SGE treats:

- **Obligations** as vector fields pushing us toward better moral configurations (e.g., less harm, more fairness, more respect for rights).
- **Interests** as vectors representing what different stakeholders care about (patients, institutions, regulators, future people, etc.).

- **Constraints** as regions where actions are simply forbidden (e.g., coercion, discrimination, certain rights violations).

The central object is a **satisfaction operator** Σ that, for each point in the moral landscape, returns a score that encodes how well obligations and interests are aligned, subject to hard constraints. SGE lays down five simple axioms for Σ—locality, coordinate invariance, monotonicity in the "interest" direction, respect for constraints, and compatibility with stratification—plus a scale-normalization assumption saying that only relative magnitudes of obligations matter, not arbitrary units.

The main representation theorem then shows that any operator satisfying these conditions must, on the regular parts of the landscape, have a very specific form:

- A *veto term* that assigns $-\infty$ to actions in forbidden regions; and
- A *smooth scalarization* term that depends only on a dimensionless alignment quantity between obligations and interests.

In plainer language: if you want a local, coordinate-free, constraint-respecting way to score actions, you are forced into something like "first apply hard vetoes, then combine the remaining trade-offs using a smooth scoring rule."

This is exactly the pattern DEME uses in its *governance profiles*: hard veto regions plus scalarization and lexical priorities.

---

3. From continuous landscapes to computable ethics

An obvious worry is that beautiful continuous mathematics may be intractable at real-time speeds. SGE addresses that concern head-on.

First, it proves **finite approximation theorems**: any decision problem on a compact stratified moral space can be approximated by a finite graph with explicit error bounds. The optimal actions on this graph differ from the true optimum by at most a small, controllable margin.

Second, it provides **complexity bounds** for evaluating satisfaction and making decisions: the core algorithms run in polynomial time in the number of candidate actions and moral dimensions, which is compatible with embedded hardware budgets for moderate problem sizes.

Third, it tackles **verification and learning**: A restricted but expressive ethical specification language can be checked for satisfaction in an o-minimal setting; static properties are decidable, and temporal properties over finite approximations can be handled by standard model checking.

The classes of obligations, interests, and metrics considered have finite sample-complexity bounds: in principle, we can learn their parameters from human feedback with guarantees in the usual PAC sense.

Together, these results show that the moral landscape isn't just a metaphor; it can be discretized, computed over, checked, and learned from in a mathematically controlled way.

---

4. DEME 2.0: ethics as a first-class subsystem

DEME 2.0 takes these ideas and turns them into an engineering blueprint for safety-critical systems.

Key elements:

- A **moral vector space**: each candidate action is mapped to a vector whose coordinates encode harm, rights, fairness, autonomy, epistemic quality, and other ethically salient quantities.
- **Governance profiles**: regulators, institutions, or communities specify hard veto regions ("never discriminate on protected attributes," "never prioritize property over life"), scalarization functions, and lexical priority rules. These profiles are versioned, inspectable, and composable across stakeholders.

A three-layer architecture:

- **Strategic**: where profiles are authored, debated, and approved.
- **Tactical**: planners and learning systems propose candidate actions based on reward signals and world models.
- **Reflex**: hardware **Ethics Modules** compile profiles into combinational logic that can veto or rank actions within sub-millisecond control loops.

At runtime, planners or policies propose actions; Ethics Modules evaluate their moral vectors against the active profile, enforce any hard vetoes, and rank the remaining options. If everything is vetoed, the system falls back to a safe default, like emergency braking or halting.

DEME wraps this with a **cryptographically anchored audit trail**: every decision comes with a “proof” linking the EthicalFacts, moral vectors, profile version, vetoes, and final outcome. This design is meant to meet traceability expectations in frameworks like the EU AI Act and NIST’s AI Risk Management Framework, without baking in any particular legal regime.

Conceptually, DEME is a computational realization of the “moral landscape” idea: moral peaks and valleys are coordinates in a high-dimensional space, and governance profiles are algorithms for moving through that terrain under hard constraints.

---

5. Philosophical stance and scope

SGE is **metaethically neutral**: it does not claim where moral truth comes from or resolve debates between consequentialism, deontology, or virtue ethics. Instead, it provides a geometric and computational substrate into which different ethical contents can be encoded. Humans—regulators, ethicists, communities—are still responsible for specifying which dimensions, obligations, and constraints count.

The framework has clear limitations: specifying good ethical content remains hard; very high-dimensional moral spaces could strain real-time budgets; and adversarial manipulation of ethical inputs is an open problem. Still, taken together, SGE and DEME aim to show that **verifiable, real-time machine ethics is mathematically possible**. They turn the moral landscape from a metaphor into a structure you can implement in silicon, reason about, and subject to public and regulatory scrutiny.