Autonomous Vehicle Ethical Decision-Making Through Computable Integrity

by Nick Clark | Published March 27, 2026 | PDF

The trolley problem is the wrong frame for autonomous vehicle ethics. Real ethical challenges in autonomous driving are not rare dilemmas but continuous normative consistency problems: maintaining safe following distances while optimizing traffic flow, balancing passenger comfort against pedestrian safety, and ensuring that the vehicle's aggregate behavior over millions of decisions remains within its declared ethical parameters. Computable integrity provides the structural mechanism for this through three-domain normative tracking with real-time deviation detection and self-correction.

The real ethical problem in autonomous driving

Public discourse fixates on dramatic ethical dilemmas: should the car hit one person or five? Actual autonomous driving ethics are mundane and continuous. Should the car maintain a two-second following distance in heavy traffic, making merging difficult for other vehicles? Should it yield preemptively to pedestrians who have not yet entered the crosswalk? Should it accept a slightly longer route to avoid a school zone during dismissal time?

These decisions are made thousands of times per hour. Each individual decision is low-stakes. But the aggregate pattern of decisions defines the vehicle's ethical character. A vehicle that consistently prioritizes passenger time over pedestrian comfort has made an ethical choice, even though no single decision was dramatic. Current autonomous driving systems make these decisions through optimization functions that have no mechanism for tracking their own normative consistency.

Why rule-based and learned approaches are insufficient

Rule-based systems encode specific ethical decisions: always yield to pedestrians in crosswalks, never exceed the speed limit. These rules are necessary but insufficient. They cannot cover the continuous space of ethical trade-offs that driving requires. Two rules may conflict: yield to a pedestrian requires slowing, but maintaining minimum traffic flow speed is also required. The system needs a mechanism to evaluate which normative consideration takes precedence in context.

Learned approaches train on human driving data, inheriting whatever ethical biases exist in the training set. A model trained on aggressive driving data makes aggressive ethical trade-offs. A model trained on overly cautious data impedes traffic. Neither approach tracks whether the vehicle's behavior is consistent with a declared normative standard over time.

How computable integrity addresses this

Computable integrity provides a three-domain model for tracking normative consistency. The vehicle declares its normative parameters: safety margins, pedestrian deference levels, traffic cooperation thresholds. These declarations constitute the vehicle's integrity baseline. At every decision point, the vehicle's action is evaluated against its declared norms, and any deviation is computed using the deviation function $D=(N-T)/(E×S)$.

The deviation function captures how far the vehicle's actual behavior has drifted from its declared norms, weighted by the ethical significance of the context and the vehicle's recent behavioral trajectory. Small deviations in low-significance contexts accumulate slowly. A large deviation in a high-significance context triggers immediate self-correction.

The coherence trifecta ensures that the vehicle's internal state, its external behavior, and its self-assessment remain aligned. A vehicle that is deviating from its norms but does not detect the deviation has a coherence failure. The trifecta mechanism makes this detectable: the vehicle tracks not just its behavior but whether its self-assessment of its behavior is accurate.

Self-correction operates continuously. When deviation accumulates beyond a threshold, the vehicle adjusts its behavior to restore normative consistency. This is not a dramatic intervention. It is a continuous feedback loop that keeps the vehicle's aggregate behavior aligned with its declared ethical parameters across millions of individual decisions.

What implementation looks like

An autonomous vehicle manufacturer deploying computable integrity parameterizes each vehicle with a normative profile: safety priority weights, pedestrian interaction preferences, traffic cooperation parameters. The integrity field tracks deviation from these parameters continuously. Regulators can audit the integrity field to verify that the vehicle's actual behavior matches its certified normative profile.

For fleet operators, integrity tracking provides fleet-level ethical consistency monitoring. A fleet whose aggregate deviation is increasing may have a systematic calibration issue. A single vehicle with anomalous integrity readings may have a sensor or software problem that is causing normatively inconsistent behavior before it manifests as a safety incident.

For regulators, computable integrity provides the auditable ethical record that current autonomous driving systems lack. Instead of inspecting millions of individual driving decisions, regulators examine the integrity trajectory: has the vehicle maintained consistency with its declared normative parameters? The answer is a computable, verifiable state rather than a judgment call.

Integrity & Coherence All 21 steps →

Track normative consistency. Detect deviation. Self-correct.

AQ
deterministic
autonomy

Legal

Subject to one or more pending U.S. and international patent applications, see [Patents](#) for the current list and status. No license, express or implied, is granted. Any use requires a separate written agreement—see [Licensing](#). Patent applications referenced on this site are pending. Claim scope, if any, is subject to examination and may issue in altered form or not at all. See [Legal](#) for terms and conditions.

Last updated: 2026-03-03

-
- [Inventive Steps](#)

-
- nick@qu3ry.net
- 72 28 14 36 01

[Invented by Nick Clark](#) | Founding Investors: Devin Wilkie