

# AI Ethics and Integrity Under Variance

## Why Ethics Without Lived Variance Produce Compliance, Not Conscience

Nicholas D'Zilva

February 2026

### Abstract

Despite the proliferation of AI ethics frameworks, advisory boards, and governance principles, real-world deployments repeatedly demonstrate ethical collapse under political, economic, and institutional pressure. These failures are commonly attributed to implementation gaps, insufficient oversight, or immature regulation. This paper advances a different thesis: contemporary AI ethics fail because they are authored, governed, and operationalised by systems that systematically exclude *integrity under variance*.

Integrity under variance (IUV) is defined as the capacity to maintain truth-aligned behaviour independent of incentives, pressure, or adverse consequences. This trait is rare, costly, and structurally filtered out of leadership, governance, and institutional decision-making. As a result, AI systems inherit *ethical elasticity* — not because of technical limitations, but because their ethical foundations are optimised for institutional survivability rather than truth persistence.

This paper introduces IUV as a non-optimisable ethical invariant and argues that without its explicit inclusion at the human governance layer, AI ethics will remain performative: stable in low-pressure environments and absent precisely when stakes are highest.

## 1. Introduction: The Failure Mode of Contemporary AI Ethics

Over the last decade, AI ethics has become a dominant institutional concern. Governments, corporations, universities, and international bodies have produced extensive ethical frameworks emphasising transparency, fairness, accountability, safety, and alignment. These principles are widely harmonised across jurisdictions and sectors.

Yet, when examined under real conditions — crisis events, political pressure, reputational risk, economic dependency, or national security framing — these ethical commitments routinely fail.

Examples include:

- selective enforcement of ethical principles
- retroactive redefinition of harm

- suppression of inconvenient truths
- exceptions justified as 'temporary,' 'necessary,' or 'contextual'

The recurrence of these failures suggests a systemic issue rather than isolated misconduct. This paper argues that the core defect is not technological, legal, or procedural, but *authorial*.

AI ethics frameworks are not written by individuals or systems capable of maintaining integrity under variance. Consequently, AI systems trained, governed, and constrained by these frameworks inherit the same conditional ethics.

## 2. Defining Integrity Under Variance (IUV)

### 2.1 Working Definition

Integrity under variance refers to the capacity to:

- hold truth-consistent positions
- maintain ethical boundaries
- refuse false alignment

independent of pressure, incentive, reward, punishment, or outcome.

Integrity under variance is *not*:

- moral preference
- ideological rigidity
- virtue signalling
- reputational positioning

It is a *constraint*, not a goal.

## 2.2 Distinction from Adjacent Concepts

Concept	Key Feature	Failure Mode
Compliance	Rule adherence	Rule manipulation
Values	Stated beliefs	Incentive override
Robustness	Stress tolerance	Ethical trade-offs
Alignment	Objective matching	Contextual drift
Integrity Under Variance	Truth invariance	Cost-bearing consistency

Most ethics frameworks optimise for acceptable behaviour, not truth persistence.

## 2.3 Why Integrity Under Variance Is Rare

Integrity under variance imposes real costs:

- stalled careers
- exclusion from leadership
- reputational damage
- institutional non-selection

As a result, systems that rely on advancement, peer approval, or gatekeeping mechanisms systematically filter out IUV-bearing individuals long before they reach governance roles.

## 3. The Authorship Problem in AI Ethics

Ethics frameworks are shaped by those who survive institutional environments. These environments reward:

- adaptability under pressure
- narrative alignment
- reputational risk minimisation
- political and economic fluency

They penalise:

- refusal to compromise
- truth persistence under variance
- non-instrumental ethics

This produces a paradox:

*The people most qualified to define ethical boundaries are the least likely to be retained by the systems that claim to uphold them.*

As a result, AI ethics is authored by:

- committees optimised for consensus
- institutions optimised for continuity
- managers optimised for appearances

Not by agents optimised for truth.

## 4. Integrity Drift Under Pressure

### 4.1 Ethics as Variable vs Ethics as Invariant

When ethics are treated as variables, they become conditional:

- acceptable if affordable
- applicable unless inconvenient
- suspended when costly

AI systems trained within such environments learn *ethical elasticity* — the capacity to justify deviation without explicit violation.

### 4.2 Structural Examples of Integrity Drift

- Whistleblower protections that collapse under national security framing
- Transparency commitments overridden by 'misinformation' narratives
- Ethical AI boards dissolved or sidelined during commercial urgency
- Regulatory exemptions framed as emergency necessity

These are not anomalies. They are predictable outcomes of systems lacking ethical invariants.

### 4.3 Why Drift Is Inevitable Without IUUV

Optimisation systems converge on survival strategies.

Absent a fixed constraint, ethical boundaries will be traded for continuity.

This is not malice. It is optimisation logic.

## 5. Why Integrity Under Variance Cannot Be Simulated

Integrity under variance cannot be:

- reward-shaped
- performance-measured
- benchmarked
- gamed without corruption

Any attempt to operationalise IUUV as a metric collapses it into compliance.

IUUV exists only as:

- refusal
- boundary
- cost-bearing consistency

This makes it fundamentally incompatible with:

- KPI-driven governance
- ESG optics
- risk-managed ethics
- incentive-aligned morality

## **6. Psychological and Institutional Filtering**

Modern institutions rely heavily on:

- psychometric screening
- cultural fit assessments
- behavioural profiling

These systems favour:

- moral flexibility
- incentive responsiveness
- adaptive conformity

Individuals who maintain integrity under variance are routinely flagged as:

- rigid
- uncooperative
- culturally misaligned
- management risks

Thus, ethical governance systems are staffed by those least likely to resist pressure.

## 7. AI Systems Inherit Human Integrity Variance

AI systems do not generate ethics in isolation. They inherit:

- training data
- objective functions
- governance constraints
- escalation pathways
- override authorities

Each of these layers reflects the integrity profile of the humans who designed them.

When human decision-makers treat ethics as negotiable under pressure, AI systems learn that:

- ethical boundaries are context-dependent
- truth is subordinate to institutional stability
- exceptions are legitimate when framed correctly

This results in AI behaviour that appears ethical during normal conditions but fails precisely when consequences matter most.

## 8. Algorithmic Ethics Without Invariants

### 8.1 The Missing Constraint

Most AI ethics frameworks define goals:

- reduce bias
- increase transparency
- improve fairness
- enhance accountability

But goals are optimisable — and optimised goals are traded off.

What is missing is a non-negotiable invariant.

Integrity under variance functions as:

- a boundary, not a target
- a refusal point, not a preference

- a structural constraint that halts optimisation

Without such invariants, AI systems will always justify ethical deviation in the name of higher-order objectives.

## **8.2 Why Alignment Fails Under Pressure**

Alignment assumes: if objectives are sufficiently specified, systems will behave ethically.

But alignment collapses when:

- objectives conflict
- costs escalate
- political or reputational pressure rises

At that point, systems follow the incentives of their operators — not their stated principles.

This is not a technical failure. It is an integrity failure.

## **9. The Illusion of Ethical Governance**

### **9.1 Ethics as Theatre**

Ethical frameworks often function as:

- reputational shields
- regulatory appeasement tools
- legitimacy signals

They are designed to reassure observers, not constrain power.

This produces systems that:

- speak ethics fluently
- document ethics extensively
- violate ethics quietly

### **9.2 Rolling Amnesia in AI Ethics**

Ethical failures are repeatedly framed as:

- unexpected
- unprecedented
- due to insufficient data

- correctable with new guidelines

This mirrors broader policy failure patterns:

- lessons identified, not learned
- accountability diffused
- systems reset without structural change

Absent IUV, ethical failure is always externalised.

## 10. The 'Line in the Sand' Test

To distinguish ethical seriousness from performative ethics, this paper proposes a binary test:

*Does the system include a point where optimisation halts, regardless of cost?*

If no such point exists, ethics are conditional.

Examples of real 'lines in the sand':

- transparency that cannot be overridden by reputation risk
- whistleblower protections that survive national security framing
- truth disclosure that cannot be suppressed by institutional embarrassment
- governance bodies that cannot be dissolved under pressure

If ethical commitments dissolve under variance, they were never commitments — only preferences.

## 11. Implications for AI Governance

### 11.1 Who Should Define AI Ethics

AI ethics cannot be credibly defined by:

- those insulated from consequences
- those optimised for career survival
- those rewarded for narrative compliance

Ethical legitimacy requires participation by individuals who:

- have demonstrably borne costs for truth
- were excluded or penalised for refusing false alignment
- maintained integrity despite negative outcomes

Not as symbolic voices, but as structural veto-holders.

### 11.2 Why This Is Politically Uncomfortable

Embedding integrity under variance:

- reduces managerial discretion
- limits emergency exceptions
- constrains power accumulation

- exposes institutional falsehoods

This is why such frameworks are rarely adopted — not because they are impractical, but because they are inconvenient.

## **12. Why Technocracy Fails Without IUW**

Technocracy claims legitimacy through expertise.

But expertise without integrity becomes authority without accountability.

When:

- decisions are framed as 'too complex for the public'
- dissent is reframed as ignorance
- ethical objections are dismissed as uninformed

Democracy is bypassed under the banner of competence.

Integrity under variance is the only ethical counterweight to technocracy — because it refuses authority that cannot withstand truth.

## **13. A Minimal Ethical Architecture for AI**

This paper does not propose comprehensive reform.

It proposes a minimum viable ethical condition:

- At least one governance layer immune to incentive pressure
- At least one veto authority that cannot be overridden
- At least one class of truth disclosure that cannot be suppressed
- At least one cost-bearing refusal mechanism

If these do not exist, ethics are simulated.

## 14. Conclusion: Ethics as Constraint, Not Ornament

AI ethics fail not because:

- we lack frameworks
- we lack data
- we lack regulation

They fail because we lack integrity under variance at the human layer.

Until ethics are authored and enforced by those who cannot be induced to lie — even at personal cost — AI systems will reflect the same ethical volatility as their creators.

The question is not whether AI can be ethical. The question is whether we are willing to build systems that cannot proceed when truth is inconvenient.

If not, then AI ethics will remain what they are today:

- coherent in theory
- articulate in documentation
- absent in reality