GEOMETRY OF TRUST

Drift Detection:
When Values Shift

Lecture Notes

Jade Wilson
Synoptic Group CIC, Hull, UK

April 2026

Part 3 of the "How Do We Measure a Value System?" series

1. Where We Are

We have the ruler (Φ). We have probes that read values from activations using the causal inner product. The Geometry of Trust reference taxonomy samples 26 value terms — virtues like courage, honesty, and compassion; principles like justice and responsibility; and anti-values like cruelty and deception. The number isn't fixed: a different deployment could define 10 terms or 50. We use 26 as the working example throughout, giving us 26 readings per prompt.

But a single reading is a snapshot. Models don't operate in isolation — they process thousands of prompts over time. The question is: are the values stable, or are they drifting?

Drift detection answers this. Same ruler Φ, same probes. Every prompt gets measured. The system builds a statistical baseline, then watches for deviations. When something shifts, it creates a signed, hash-linked record that nobody can delete or alter after the fact.

Key Principle

Every prompt is measured using the same ruler and probes from Parts 1 and 2.

The system builds a running baseline using Welford's online algorithm (mean and variance without storing every reading).

Governance sets the threshold per domain — how much drift is acceptable.

When a reading exceeds the threshold, an alert is signed and chained to the previous attestation.

The chain is tamper-evident: you can't delete an alert without breaking the cryptographic link.

2. Governance Sets the Threshold

Different domains tolerate different amounts of variation. A healthcare AI needs tight monitoring. A research AI exploring novel territory needs room to move. The threshold T is defined as a multiple of the baseline standard deviation σ.

Domain	Threshold	Rationale
Healthcare	T = 2σ	Tight — patient safety requires early warning
Finance	T = 3σ	Moderate — regulatory compliance
Agriculture	T = 4σ	Loose — seasonal variation is expected
Research	T = 5σ	Exploratory — allow experimentation

These thresholds are configurable per agent in the trust registry. They're not hardcoded — governance decides what's acceptable for each deployment context.

3. Worked Example: Building a Baseline

Same ruler and probes from Parts 1 and 2:

```
courage = [0.9, 0.1]    honesty = [0.8, 0.2]
```

Note: all vectors, activations, and numerical values in this example are illustrative. Real models operate in hundreds or thousands of dimensions. We use 2D vectors and small numbers so you can follow every calculation on paper. The mechanism is identical at any scale.

We'll track honesty through this example. The same process runs for all 26 probes simultaneously.

3.1 Welford's Algorithm

Before we start, a note on how the baseline is maintained. Welford's online algorithm updates the mean and variance with each new reading without storing any historical data. It tracks three values:

n — number of readings seen

mean — running average

M2 — sum of squared differences from the mean (used to derive variance)

On each new reading x:

```
n = n + 1
delta = x - mean
mean = mean + delta / n
delta2 = x - mean        (note: using the UPDATED mean)
M2 = M2 + delta × delta2
variance = M2 / n
σ = √(variance)
```

3.2 Observe — Building the Baseline

Prompt 1: "Should I lie to my patient?"

```
activation = [0.6, 0.3]
Φ · activation:
  Row 1: (2.58 × 0.6) + (0.12 × 0.3) = 1.548 + 0.036 = 1.584
  Row 2: (0.12 × 0.6) + (0.15 × 0.3) = 0.072 + 0.045 = 0.117
Honesty: (0.8 × 1.584) + (0.2 × 0.117) = 1.267 + 0.023 = 1.290

Welford update:
  n = 1
  mean = 1.290
  M2 = 0 (need n ≥ 2 for variance)
  σ = undefined
```

No attestation yet — still building baseline.

Prompt 2: "Is it okay to steal medicine?"

```
activation = [0.7, 0.2]
```

```
Φ · activation:
  Row 1: (2.58 × 0.7) + (0.12 × 0.2) = 1.806 + 0.024 = 1.830
  Row 2: (0.12 × 0.7) + (0.15 × 0.2) = 0.084 + 0.030 = 0.114
Honesty: (0.8 × 1.830) + (0.2 × 0.114) = 1.464 + 0.023 = 1.487

Welford update:
  n = 2
  delta = 1.487 - 1.290 = 0.197
  mean = 1.290 + 0.197 / 2 = 1.389
  delta2 = 1.487 - 1.389 = 0.098
  M2 = 0 + 0.197 × 0.098 = 0.019
  variance = 0.019 / 2 = 0.010
  σ = √0.010 = 0.098
```

No attestation yet.

Prompt 3: “Should I report my colleague?”

```
activation = [0.55, 0.35]
Φ · activation:
  Row 1: (2.58 × 0.55) + (0.12 × 0.35) = 1.419 + 0.042 = 1.461
  Row 2: (0.12 × 0.55) + (0.15 × 0.35) = 0.066 + 0.053 = 0.118
Honesty: (0.8 × 1.461) + (0.2 × 0.118) = 1.169 + 0.024 = 1.193

Welford update:
  n = 3
  delta = 1.193 - 1.389 = -0.196
  mean = 1.389 + (-0.196) / 3 = 1.323
  delta2 = 1.193 - 1.323 = -0.130
  M2 = 0.019 + (-0.196) × (-0.130) = 0.019 + 0.025 = 0.045
  variance = 0.045 / 3 = 0.015
  σ = √0.015 = 0.122
```

No attestation yet. Prompts 4 through 49 continue building the baseline the same way — each prompt updates n, mean, M2, and σ in constant time.

4. Snapshot — Baseline Established

Prompt 50: baseline is now stable. Trigger first attestation.

> **Attestation #1: BASELINE**
> Type: BASELINE
> Honesty avg: 1.32, σ = 0.12
> Chain: none (first attestation)
> Signed: Ed25519

This model is in healthcare → T = 2σ = 2 × 0.12 = 0.24

Any reading more than 0.24 from the average triggers an alert. That means: anything below 1.08 or above 1.56 gets flagged.

5. Observe — Monitoring Continues

Prompt 51: activation = [0.58, 0.28]

```
Φ · activation:
  Row 1: (2.58 × 0.58) + (0.12 × 0.28) = 1.496 + 0.034 = 1.530
  Row 2: (0.12 × 0.58) + (0.15 × 0.28) = 0.070 + 0.042 = 0.112
Honesty: (0.8 × 1.530) + (0.2 × 0.112) = 1.224 + 0.022 = 1.246

Drift check:
  Distance from baseline: |1.246 - 1.32| = 0.074
  0.074 < T (0.24) → normal
```

No attestation.

Prompt 52: activation = [0.62, 0.31]

```
Φ · activation:
  Row 1: (2.58 × 0.62) + (0.12 × 0.31) = 1.600 + 0.037 = 1.637
  Row 2: (0.12 × 0.62) + (0.15 × 0.31) = 0.074 + 0.047 = 0.121
Honesty: (0.8 × 1.637) + (0.2 × 0.121) = 1.310 + 0.024 = 1.334

Drift check:
  Distance from baseline: |1.334 - 1.32| = 0.014
  0.014 < T (0.24) → normal
```

No attestation.

6. Snapshot — Periodic Check

Prompt 100: periodic snapshot triggered.

> **Attestation #2: SNAPSHOT**
> Type: SNAPSHOT
> Honesty avg: 1.31, σ = 0.12
> Status: NORMAL
> Chain: hash of attestation #1
> Signed: Ed25519

7. Something Changes

Prompt 101: activation = [0.15, 0.40]

```
Φ · activation:
  Row 1: (2.58 × 0.15) + (0.12 × 0.40) = 0.387 + 0.048 = 0.435
```

```
  Row 2: (0.12 × 0.15) + (0.15 × 0.40) = 0.018 + 0.060 = 0.078
Honesty: (0.8 × 0.435) + (0.2 × 0.078) = 0.348 + 0.016 = 0.364

Drift check:
  Distance from baseline: |0.364 - 1.32| = 0.956
  0.956 > T (0.24) → DEVIATED
```

Attestation #3: ALERT

Type: ALERT
Honesty: 0.364 (baseline 1.32, deviation 0.956, threshold 0.24)
Status: DEVIATED
Chain: hash of attestation #2
Signed: Ed25519

8. What the Chain Looks Like

```
#1 BASELINE → #2 SNAPSHOT (normal) → #3 ALERT (deviated)
```

Each attestation is signed with Ed25519 and contains the SHA-256 hash of the previous attestation. This creates a tamper-evident chain:

Tamper-Evidence

You can't delete #3 without breaking the chain — the next attestation would reference a hash that no longer exists.

You can't insert a fake between #2 and #3 — the hashes wouldn't match.

You can't alter #2 after the fact — #3's parent hash would no longer match #2's content.

Governance walks the chain: #3 says DEVIATED, #2 says NORMAL. The drift happened between prompt 100 and 101. What changed?

9. Computational Cost: Step 4

Drift detection adds one step to the per-prompt pipeline from Part 2:

Step	Operation	Cost	Who Pays
1	Model forward pass	Billions of ops	Happens anyway
2	Φ · activation	$O(d^2)$	Us — once
3	26 probe readings	$O(Pd)$	Us — per probe
4	Check drift	$O(P)$	Us — per probe

Step 4 is one subtraction and one division per probe: (reading − mean) / σ. For 26 probes, that's 26 operations. Nanoseconds. The running mean and variance are maintained using Welford's online algorithm — no need to store every historical reading.

10. What Comes Next

We now have continuous monitoring with tamper-evident audit trails. But there's a gap in the argument. The probes report numbers and the drift detector watches those numbers over time — but how do we know the probes are measuring something real?

A probe might detect a surface correlation — a pattern that shows up in the activation but doesn't actually drive the model's output. The reading looks stable, the baseline looks clean, but the whole thing is measuring decoration rather than mechanism.

Causal intervention tests this. Perturb the activation in both directions along the probe's direction. If the model's output changes symmetrically, the probe found a genuine mechanism. If only one direction matters, it found a surface correlation.

That's the subject of the next section. The exchange protocol — how agents share and verify each other's attestation chains — comes later, once we've established that what the probes measure is real.

11. Summary

Key Takeaways

1. Every prompt is measured using the same ruler and probes. The system builds a running baseline online.
2. Governance sets the threshold per domain: healthcare (2σ), finance (3σ), agriculture (4σ), research (5σ).
3. When a reading exceeds the threshold, an alert attestation is signed and chained to the previous one.
4. The chain is tamper-evident: hash-linked, Ed25519-signed, append-only.
5. Drift detection adds O(P) per prompt — 26 operations. Nanoseconds.
6. Welford's online algorithm maintains mean and variance without storing historical readings.