

GEOMETRY OF TRUST

GOVERNANCE SERIES

Governance Thresholds: How Tight Is Tight Enough?

Lecture Notes

Jade Wilson

Synoptic Group CIC, Hull, UK

April 2026

Part 4 of the Governance Series

1. Where We Are in the Stack

The three earlier talks in this series set up the structural governance layer. Safety doesn't travel between domains (Part 1). Every agent declares one primary domain (Part 2). Cross-domain interactions run three structural checks — exclusions, permissions, mode — before any cryptography (Part 3). Those are all yes/no decisions. An agent is either allowed to interact with a peer or it isn't.

This talk is about the quantitative layer sitting on top. Once two agents have passed the structural checks and their attestations are being evaluated, how strictly does the verifier hold them to quantitative bounds? How much drift is too much? How confident does a probe reading have to be? Does causal validation have to pass, or is correlational evidence enough?

The Key Variable

T = the governance threshold.

T is not a number the maths produces. It's a number governance sets.

The maths produces readings — drift magnitudes, confidence scores, causal consistency ratios. Governance decides what counts as acceptable given the domain's tolerance for error.

Different domains get different T. That's the entire point of this talk.

A Critical Note on the Numbers in This Talk

Every numerical threshold in this talk — 0.02, 0.03, 0.05, 0.10, 0.25 — is illustrative, not prescriptive. These are made-up values chosen to show the *shape* of a tiered framework, not to recommend what critical infrastructure or healthcare should actually use.

The real values have to be set by domain regulators working with operators, auditors, and the standards bodies they already answer to.

A threshold that looks reasonable in a slide can fail badly in deployment. Getting these numbers right is a job for people who know the domain and have been watching how the measurements behave in practice — not for a framework author.

Treat this talk as an argument about how thresholds should be *structured*, not what they should be.

2. Thresholds by Domain — An Illustrative Tiering

Different domains tolerate different amounts of drift and demand different depths of evidence. The tiering below is the kind of picture you'd expect a domain regulator to arrive at after thinking about what failure looks like in their world. Again, these numbers are illustrative.

Domain	Max drift	Causal validation	Rationale (illustrative)
--------	-----------	-------------------	--------------------------

Critical infrastructure	0.02	Required	Public safety exposure is large; deployed geometry is near-static by design.
Healthcare	0.03	Required	Patient safety tolerates a narrow band of variation before a recommendation becomes unsafe.
Finance	0.05	Required	Regulatory compliance needs verifiable evidence; slightly more latitude than healthcare because positions unwind.
Commercial supply chain	0.10	Not required	Business priorities legitimately shift with seasons and demand; correlational evidence usually sufficient.
Research / experimental	0.25	Not required	Exploration requires room to move; overly tight bounds would suppress the experimentation the deployment exists to enable.

A few things to notice about the shape of this tiering, even with the specific numbers held at arm's length.

Tighter drift and mandatory causal validation come together. The domains with the smallest tolerance for drift are the same domains that can't accept correlational evidence as proof that values are still where they should be. They need the stronger guarantee.

“Required” is a per-interaction property, not a platform property. A critical-infrastructure agent demanding causal validation doesn't mean the maths is always running — it means the regulator's verifier won't accept an attestation without a causal certificate attached. The cost of causal probes gets paid at attestation time, when the agent is certifying itself to a strict peer, not on every inference.

Numbers get looser by an order of magnitude across the tiers. Critical infrastructure at 0.02 vs research at 0.25 is roughly a 12× difference. That's not an arbitrary spread — it reflects that the cost of a false-positive alarm in research (blocking a legitimate experiment) is much lower than the cost of a false-negative in critical infrastructure (letting a drifted model keep operating).

3. The Dual-Domain Problem — Self-Driving Tractor

Some agents genuinely operate in two domains at once. A self-driving tractor drives on farmland for most of its working life and on public roads for the rest. It can't split into two logical agents because the hardware, sensors, and decision-making are shared. And it can't claim two primary domains — Part 2 ruled that out.

The answer is to invent a domain that captures the dual-purpose nature directly:

```
vehicle
  vehicle.autonomous-truck      (pure transport)
  vehicle.agricultural-tractor  (dual: farming + road use)
  vehicle.construction-excavator (dual: site + road use)
```

The tractor's primary domain is `vehicle.agricultural-tractor`. Its value geometry is trained on the dual-purpose objective — crop outcomes and collision avoidance both, under one coherent structure. Its probes measure both sets of concerns. A governance body (or a coordination between the agricultural and transport regulators) decides what “tractor safety” means, what drift bounds are acceptable, and whether causal validation is required.

3.1 Whose Thresholds Apply?

This is where per-peer governance thresholds matter. The tractor has one primary domain and one attestation, but different peers interact with it under different rules:

Peer	Peer's required drift (illustrative)	Peer's causal requirement	Effect on tractor
Farm management agent	0.05	Not required (chain required)	Tractor's attestation must fit the agricultural envelope.
Road-infrastructure agent	0.02	Required	Tractor's attestation must fit the transport envelope.

How the Tractor Meets Both

The tractor doesn't pick its own threshold. It gets held to whichever peer's threshold applies to the current interaction.

On farmland, talking to farm peers, the farm threshold applies — looser but still binding.

On public roads, talking to transport peers, the transport threshold applies — tighter and with causal validation required.

In practice the tractor has to stay within the **strictest** envelope that any of its expected peers will hold it to. That's what “strictest rules from both parents” means.

The peer decides which rules apply to an interaction, not the tractor. That's the whole point of per-peer governance thresholds.

Concretely, if a tractor's current drift is 0.04, it passes the farm interaction (0.05 tolerance) but fails the road interaction (0.02 tolerance). The road-infrastructure peer rejects the exchange. The tractor doesn't stop operating, but it can't participate in the road-coordination network until its geometry is re-measured and brought back inside the transport envelope.

4. A Same-Domain Pair — Diagnostic + Drug Checker

Thresholds don't only apply across domains. Inside a single regulated domain, peers may still hold each other to the full domain thresholds. A hospital-deployment pair illustrates this.

Property	Diagnostic agent	Drug-checker agent
Primary domain	healthcare.diagnostic-advisory	healthcare.drug-interaction
Mode toward peer	Advisory (sends hypotheses)	Read-only (receives hypotheses, cannot advise back)
Max drift (illustrative)	0.03	0.03
Causal validation	Required	Required
Outcome if either fails	Exchange refused	Exchange refused

Two observations worth drawing out.

Same-domain doesn't mean same-role. Both agents sit in healthcare, but one informs the other rather than negotiating as equals. The diagnostic agent is trained to generate hypotheses; the drug checker is trained to evaluate specific interactions given those hypotheses. The asymmetric mode — advisory on one side, read-only on the other — captures that. Part 3's mode framework is what allows this shape to be expressed without either agent overreaching.

Both must pass, not just one. Because the interaction is being held to healthcare-grade thresholds, both agents' attestations have to clear both the drift bound and the causal validation requirement. If the drug checker's geometry has drifted past 0.03 — even though its mode is only read-only — the interaction is refused. Read-only constrains what the agent can **say**, not how rigorously its values are checked.

5. The Asymmetric Case — Finance Regulator + Trader

Supervised mode inverts the usual symmetry. A finance regulator initiating a supervised interaction with a trading agent isn't producing an attestation of its own value geometry — it's demanding one from the trader.

Property	Regulator	Trader
Primary domain	finance.regulatory-compliance	finance.trading
Mode	Supervised (demands)	Supervised (must comply)
Produces own attestation in this interaction?	No — carries authority attestation instead	Yes — full attestation demanded

Applicable thresholds (illustrative)	n/a — the regulator is the threshold-setter	Finance max drift 0.05, causal required
Direction of information flow	Inward (demand)	Outward (proof)

What Supervised Mode Preserves

The regulator's authority is itself an attestation — not trust-by-assertion.

The trader still has its own thresholds; those haven't vanished just because a supervisor is asking.

What's changed is that the trader's obligation to produce the attestation is triggered by the supervisor's credential, not negotiated as a peer.

The one-way information flow is visible in the audit record: a supervised-mode message is a different record type from a cooperative one.

If the trader's attestation fails to meet finance-domain thresholds, the regulator sees that as a finding — not an error.

6. When Thresholds Don't Get to Matter — Structural Refusal

The last case is the one where the whole quantitative layer doesn't come into play at all.

Property	Farm agent	Transport agent
Primary domain	agriculture.crop-management	transport.autonomous-vehicle
Exclusions	transport.*	(none relevant)
Transport agent's drift reading (hypothetical)	—	0.01 — excellent
Causal score (hypothetical)	—	0.95 — excellent
Outcome	Blocked at Step 1	Blocked at Step 1

Why the Thresholds Don't Get Consulted

Structural checks run first. Exclusions are Step 1.

The transport agent's attestation could be the finest ever produced — no drift, perfect causal consistency, every probe reading within tolerance.

None of that gets evaluated. The farm agent's exclusion of `transport.*` fires before the attestation is even opened.

This is the whole point of the separation between structural and quantitative layers: structural refusal isn't an override of the maths, it's a layer that decides whether the maths ever gets to run.

A regulator reviewing the audit log sees a `DomainExcluded` record, not a `ThresholdFailed` record. The difference matters — it's the difference between “we wouldn't engage” and “we engaged and the numbers came back bad.”

7. How Thresholds Actually Get Set

The illustrative numbers above came from someone writing a talk. The real numbers have to come from somewhere else. A short picture of what it takes to set a threshold responsibly:

7.1 Who

The domain regulator, working with operators, auditors, and the standards bodies they already answer to. For healthcare, that's clinical regulators plus bodies that set clinical-decision-support norms. For critical infrastructure, that's the sectoral safety regulator plus operators with skin in the game. The framework doesn't make this easier by picking a number; it makes it easier by making clear what the number is actually constraining.

7.2 What

A threshold is a commitment to reject interactions whose measured drift exceeds the bound. To set one responsibly, a regulator needs to know: the distribution of drift readings observed across comparable deployments, the distribution of drift values at which real incidents have occurred in the past, the distribution of drift values at which false alarms become operationally disruptive. These are empirical questions that can only be answered by watching the measurements behave over time across many deployments.

7.3 When

Thresholds shouldn't be set on day one and left alone. They should be provisional at first — looser than the regulator thinks they need to be — while the measurement system is itself being validated. Tightening comes later, as the baseline distribution of drift in healthy deployments becomes well-understood. Setting a tight threshold too early produces false alarms that erode trust in the whole measurement regime.

A Reminder About the Numbers

The 0.02 / 0.03 / 0.05 / 0.10 / 0.25 tiering in this talk is illustrative.

Real values need to come from regulators working with operators and standards bodies, informed by actual deployment data.

The right number for critical infrastructure might turn out to be 0.01, or 0.04, or a multi-dimensional bound rather than a scalar.

What matters is the shape of the framework: per-domain, per-peer, set by governance not maths, revisable as evidence accumulates.

Treat the shape as the contribution. Treat the specific numbers as placeholders.

8. Summary

Key Takeaways

1. Structural governance (Parts 1–3) decides whether an interaction is allowed. Quantitative governance — thresholds — decides how strictly the interaction is held to its evidence.
2. T is set by governance, not produced by maths. The maths gives readings; governance decides what's acceptable.
3. Different domains get different T. Critical infrastructure tightest; research loosest; healthcare, finance, supply chain between.
4. All numbers in this talk are illustrative. Real thresholds must come from domain regulators with deployment data.
5. Tighter drift bounds and mandatory causal validation come together. High-stakes domains need stronger guarantees.
6. Dual-domain agents get their own domain (e.g. `vehicle.agricultural-tractor`) and meet the strictest envelope any expected peer will hold them to.
7. The peer decides which thresholds apply to an interaction, not the agent.
8. Same-domain peers can still hold each other to the full domain thresholds — asymmetric roles don't reduce the evidence requirement.
9. Supervised mode is one-way: the regulator demands, the trader complies. The regulator carries authority attestation, not a value-geometry one.
10. Structural refusals (e.g. exclusions) block interactions before thresholds are consulted. A perfect attestation from an excluded peer is still rejected.
11. Thresholds should start provisional and tighten as measurement behaviour becomes well-understood. Too-tight-too-early produces false alarms that erode trust in the whole regime.