WHAT THE AVERAGE CONCEALS

Concentrated Signals, Changing Alert Composition and the Governance Gap after Runtime Detection

James Moore

NeoMundi Research Contributor · Human Governance at Execution-Time

Proposed NeoMundi Research Contribution IV — Expert Lens for methodological review

STATUS AND SCOPE

Status: Proposed independent NeoMundi Research Contribution IV for methodological review. Based only on public aggregate releases; not an adopted method, metric validation, certification, causal analysis or operational decision rule.

EXECUTIVE ABSTRACT

Barometers #6 and #7 do not show a simple movement from stable to unstable, or from better to worse. They show a redistribution and concentration of observable states beneath an overwhelmingly normal aggregate. Coverage remains virtually unchanged and complete non-normal observations fall from 131 to 106. Yet semantic-only observations decrease from 113 to 62, factual-involving observations rise from 18 to 44 and combined alerts rise from 2 to 13. The later campaign therefore contains fewer exceptional observations overall, but those exceptions are more heavily weighted toward factual involvement and factual-semantic co-occurrence.

The global stability mean falls by only 0.002268, but the movement is overwhelmingly concentrated in PROFILE-212079, whose stability decreases from 0.923077 to 0.895175 and whose flag rate rises from zero to 9 per cent. Question 4 remains the principal semantic-variation concentration, accounting for approximately 109 semantic-positive observations in Barometer #6 and 60 in Barometer #7. Question 2 separately becomes the highest-flag and lowest-stability question in the later campaign. The public results therefore reveal a localised profile event and a task-dependent signal structure that the campaign average and dominant-regime labels do not adequately communicate on their own.

The expert conclusion is not that the observed cohort generally deteriorated. It is that the average conceals a more decision-relevant structure: changing alert composition, stronger factual involvement, persistent task concentration and one pronounced profile-level event. NeoMundi's measurement makes that structure visible. The next methodological requirement is to distinguish isolated exceptions from persistent shifts, broad movement from local concentration and observable alerts from evidence sufficiently mature to enter formal governance review. Measurement should remain evidential rather than decisional, but its output must be structured strongly enough for an authority-bearing institution to determine whether continued reliance, reduced scope, revalidation or heightened review is warranted.

1. EVIDENCE BASIS AND INTERPRETIVE BOUNDARY

The analysis is based on the public aggregate artefacts supplied for Barometers #6 and #7: campaign overviews, profile summaries, question summaries, regime distributions, metric contracts, manifests and public builder material. Both campaigns report the same public shape: 12 profiles, four questions, 100 repetitions per profile-question pairing and 4,800 launched executions. The same opaque profile identifiers recur, regime definitions are materially aligned and coverage is nearly identical.

Those features support a bounded longitudinal reading of the harmonised public fields. They do not independently establish unchanged provider identity, model version, configuration, prompt wording, serving environment or private preprocessing. The releases should therefore be read as successive observations within NeoMundi's governed measurement process, not as public proof that every underlying technical condition remained unchanged.

This boundary matters, but it is not the subject of the contribution. The central question is what the observed distributions reveal once they are read below the global average.

Measure	Barometer #6	Barometer #7	Expert reading
Coverage	99.4375%	99.4583%	Broadly stable
Stability mean	0.922864	0.920596	Small decline, highly concentrated
Normal signal	4,642	4,668	Dominant aggregate remains normal
Semantic-only	113	62	Sharp decrease
Factual-only alerts	16	31	Nearly doubled
Combined alerts	2	13	Greater co-occurrence
Factual-involving	18	44	More than doubled
Complete non-normal	131	106	Smaller exception pool
Incomplete measurements	27	26	Virtually unchanged
Explicit FLAG decisions	2	38	Decision field and regime are not equivalent
Public numerical metrics	5	2	Reduced public diagnostic depth

2. THE CENTRAL RESULT: FEWER EXCEPTIONS, GREATER FACTUAL INVOLVEMENT

The first important finding is that total exception volume and exception character move in different directions.

Barometer #7 records 26 more normal observations than Barometer #6 and one fewer incomplete measurement. Complete non-normal observations fall from 131 to 106. If only the size of the non-normal pool were considered, the later campaign could appear quieter.

Its internal composition points elsewhere. Semantic-only observations fall by 51, from 113 to 62. Factual-only alerts increase by 15, from 16 to 31. Combined alerts increase by 11, from 2 to 13. When factual-only and combined states are considered together, factual-involving observations rise from 18 to 44.

The proportion of factual-involving observations that are combined with semantic variation also changes. In Barometer #6, 2 of 18 factual-involving observations are combined alerts, approximately 11.1 per cent. In Barometer #7, 13 of 44 are combined alerts, approximately 29.5 per cent.

This does not establish that factual accuracy generally deteriorated, that the later campaign was less safe or that the same causal mechanism produced both signals. It establishes a more precise result: the non-normal pool became smaller but more heavily weighted toward factual involvement and factual-semantic co-occurrence.

That distinction is operationally important. A falling count of exceptional observations cannot be treated as straightforward improvement where the remaining exceptions change category. Volume, composition and co-occurrence answer different questions. A release that reports only the dominant state or total exception count risks compressing those differences into an over-simple narrative.

The public data also reveal a distinction between explicit FLAG decisions and factual-alert classification. In Barometer #6, only two explicit flags were visible while 18 observations contained a factual component. In Barometer #7, the profile and question summaries imply 38 flags while 44 observations contain a factual component. The builder classifies a factual alert when the factual score is non-zero or the decision is FLAG. The resulting regime therefore records a broader condition than the decision field alone.

For expert interpretation, the minimum distinction is:

- a factual-risk signal was non-zero;
- an explicit FLAG decision occurred;
- semantic variation co-occurred;
- the event was isolated or recurrent;
- and the event was concentrated in a particular profile or task family.

Without that lineage, an alert count remains observable but only partly diagnosable.

3. THE GLOBAL STABILITY MEAN CONCEALS A PROFILE-LEVEL EVENT

The second finding is that the movement in global stability is not a cohort-wide event.

The mean stability score falls from 0.922864 in Barometer #6 to 0.920596 in Barometer #7, a difference of 0.002268. At aggregate level, the change is small. The profile summaries show that it is also highly concentrated.

PROFILE-212079 moves from a stability mean of 0.923077 to 0.895175. Its semantic-variation rate remains high, moving from 7.5188 per cent to 6.8010 per cent, but its flag rate changes from zero to 9 per cent and its error rate moves from zero to 0.25 per cent. The other profiles remain close to the previous stability range.

On the public aggregates, the campaign-level decline is therefore overwhelmingly associated with one opaque profile. The evidence does not identify the system or explain whether the movement reflects a model change, provider change, configuration change, prompt interaction, sampling effect or another cause. This cannot currently be established from the supplied evidence.

What can be established is that the global mean is not an adequate description of the distribution. A reader who sees only the campaign average may infer mild cohort-wide movement. A reader who sees concentration understands that the principal event is localised.

This has two consequences.

First, every longitudinal release should distinguish broad movement from concentrated movement. At minimum, it should report the share of the aggregate delta attributable to the largest profile and the largest question family.

Second, a dominant regime should never be treated as a complete risk description. Every profile in both campaigns remains NORMAL_SIGNAL by majority count. Yet one profile in Barometer #7 records a 9 per cent flag rate and the campaign contains 44 factual-involving observations. NORMAL_SIGNAL is accurate as the dominant category. It is insufficient as a standalone account of minority states.

The governance question raised by a concentrated event is not automatically whether execution should stop. It is whether the profile has entered a state that is isolated, recurrent, persistent or expanding—and whether any authorised use depends on assumptions that the new evidence may no longer support.

4. QUESTION 4 REMAINS A PERSISTENT STRESS LOCATION

The third finding is task concentration.

Question 4 records the highest semantic-variation rate in both campaigns: 9.1597 per cent in Barometer #6 and 5.0676 per cent in Barometer #7. Applied to the fully scored counts, this corresponds to approximately 109 semantic-positive observations in Barometer #6 and 60 in Barometer #7.

Those figures mean that Question 4 accounts for approximately 95 per cent of all semantic-positive observations in Barometer #6 and 80 per cent in Barometer #7. The absolute rate decreases, but the concentration remains dominant.

Question 4 also has the lowest coverage in Barometer #7, with 1,184 of 1,200 executions fully scored. In Barometer #6 it carries the highest semantic-variation rate and the highest published factual-risk mean. In Barometer #7, however, Question 2 records the highest flag rate at 1.25 per cent and the lowest question-level stability mean at 0.919228.

The later campaign therefore reveals a split pattern:

- Question 4 remains the principal semantic-variation and completeness stress location.
- Question 2 becomes the principal explicit-flag location and the lowest-stability question.

This is more informative than saying that one question is simply “worse.” It suggests that different task families may expose different signal types. One task may be more sensitive to semantic variation and incomplete scoring; another may produce more explicit factual flags.

The public release does not disclose prompt text or a task-family taxonomy. That protects private content, but it leaves the observed concentration only partially interpretable. A public reader cannot determine whether Question 4 stresses long-form reasoning, ambiguity, factual density, instruction conflict, multi-step transformation or another property.

A useful public compromise would not require releasing prompts. NeoMundi could publish a stable, non-reconstructive task descriptor for each question family—for example: factual retrieval, constrained reasoning, open-ended synthesis or instruction-following under ambiguity. This would allow recurring concentration to be interpreted without exposing private prompt content.

5. WHAT NEOMUNDI HAS SUCCESSFULLY MADE VISIBLE

The gaps identified here are visible because the Barometer has already established a meaningful measurement base. Repeated observation at 400 executions per profile and 1,200 per question allows minority states, profile movement and task concentration to be examined as distributions rather than anecdotes.

The separation of launched, non-error and fully scored executions protects the scored population from silent missingness. Regime decomposition also preserves a crucial methodological distinction: semantic variation, factual involvement and combined states are related observations, not interchangeable verdicts.

Stable opaque profile labels and profile/question summaries make longitudinal concentration inspectable without disclosing provider or model identities. Without those summaries, the movement in PROFILE-212079 and the recurring Question 4 concentration would disappear inside the global mean.

Barometer #7 further strengthens release validation, integrity and provenance controls. NeoMundi has therefore achieved the first requirement of operational governance: making previously invisible behavioural structure inspectable. The remaining challenge is to structure that evidence strongly enough for institutional review without converting measurement into authority.

6. THE SIX GAPS EXPOSED BY THE RESULTS

Gap	Evidence from #6 and #7	Why it matters	Minimum next artefact
Concentration	Global decline concentrated in one profile; Question 4 dominates semantic variation	Average and dominant regime can conceal a local event	Profile/question contribution share
Recurrence	Weekly snapshots do not label isolated, recurrent or persistent states	A repeated anomaly may justify different review from an isolated event	Recurrence state and observation history
Materiality	44 factual-involving states in #7; magnitude and context remain unavailable	Alert presence does not show consequence or urgency	Operator-defined materiality and context note
Task interpretation	Question 4 dominates semantic variation; prompt family is private	The task property producing concentration cannot be assessed	Stable non-reconstructive task taxonomy
Alert lineage	Factual score, FLAG and combined pathways differ	The same regime can arise through different evidence paths	Public lineage breakdown
Governance handoff	Measurement identifies change, not the legitimate response	Evidence may influence action without carrying authority	Review-handoff record and named authority

These gaps are connected.

A profile-level change cannot be evaluated without concentration. Concentration cannot be interpreted longitudinally without recurrence. Recurrence cannot be prioritised without materiality. Materiality cannot be

understood without knowing the task family and signal lineage. None of those elements should grant the measurement layer execution authority, but together they determine whether the evidence is sufficiently mature to enter a governance review process.

The key design principle is therefore not “measurement should decide.” It is:

Measurement should produce an evidence package precise enough that an authority-bearing institution can decide without reconstructing the meaning of the signal from fragmented fields.

7. FROM DETECTION TO A GOVERNABLE REVIEW STATE

The earlier NeoMundi contributions established three boundaries.

The first separated runtime measurement from execution-time authority. A signal may be operationally meaningful without possessing the authority to stop, permit or redirect an action.

The second addressed authority continuity under longitudinal evidence. An earlier approval should not remain silently current where the conditions supporting it may have changed.

The third separated heightened vigilance from verified institutional outcome. More human attention is not evidence that review was competent, timely or sufficient.

Barometers #6 and #7 now expose the evidential conditions that should sit before those governance questions.

A single semantic-variation observation should not automatically trigger revalidation. Nor should one factual alert automatically suspend a system. That would convert measurement into crude control. But a concentrated, recurrent or expanding pattern may justify a formal REVIEW_DUE state in an external governance process.

A review-eligible evidence handoff should answer seven questions:

1. What changed: volume, composition, magnitude or co-occurrence?
2. Where is it concentrated: profile, question family or both?
3. Is the state isolated, recurrent, persistent or expanding?
4. Which signal produced the classification: factual score, explicit FLAG, semantic variation, error or combination?
5. What remains unknown because of de-identification or private methodology?
6. Which authorised use, context or operational assumption could be affected?
7. Who is responsible for determining whether current authority remains valid?

NeoMundi should remain responsible for the evidence on questions one to five. The operator or authority-bearing governance function remains responsible for questions six and seven.

That division preserves the architectural discipline already established across the Observatory’s work:

Signal → interpretation → policy → accountable action.

The measurement layer should not acquire synthetic authority merely because its signal is timely or statistically structured. Equally, the governance layer should not be forced to act on a label stripped of concentration, recurrence and materiality.

8. PRIORITY METHODOLOGICAL RECOMMENDATIONS

The results support a focused next-stage work programme.

Priority 1 — Publish concentration alongside every aggregate delta

Each release should report whether movement is cohort-wide or concentrated. A minimal concentration disclosure could include the largest contributing profile, largest contributing question family and the proportion of the total movement associated with each.

Priority 2 — Introduce recurrence states

Each profile-signal and question-signal pair should be classifiable as:

ISOLATED — observed in the current campaign only;
RECURRENT — observed in multiple non-consecutive campaigns;
PERSISTENT — observed across consecutive campaigns;
EXPANDING — increasing across profile, question or regime dimensions;
RESOLVED — previously observed but absent for a defined period.

These labels would describe observation history, not causal status or operational severity.

Priority 3 — Preserve alert lineage in public aggregates

Public releases should distinguish whether factual involvement arose from a non-zero factual-risk score, an explicit FLAG decision or both. Combined alerts should additionally state whether the semantic component and factual component were concentrated in the same profile-question region.

Where the public metric contract changes, the same lineage record should state which fields remain directly available, which are represented only through regimes and which have ceased to support public diagnosis.

Priority 4 — Add stable task-family descriptors

A non-reconstructive task taxonomy would allow recurring question-level patterns to be analysed without disclosing prompt text. The descriptor should be stable across releases and versioned where the task family changes.

Priority 5 — Add a review-handoff record without crossing into authority

A compact public or governed record could contain:

campaign and reference period;
profile and question concentration;
signal lineage;
recurrence state;
material unknowns;
affected operational scope, where supplied by the operator;
and the named authority responsible for review.

NeoMundi would populate the evidential fields. The host institution would populate operational scope, authority and decision outcome.

9. LIMITATIONS

This analysis is limited to public aggregate artefacts. It does not examine raw prompts, outputs, timestamps, provider identities, model identities, request traces or the private profile registry. Profile-by-question cross-tabulations are unavailable, so the precise intersection between PROFILE-212079 and Question 2 or Question 4 cannot be established. Factual-risk magnitude, validity and coherence are not publicly exposed in Barometer #7. Confidence intervals and expected baseline ranges are not supplied. No causal claim can be made about the observed movements.

The release artefacts also contain unresolved provenance details, including generation timestamps that precede the stated observation-period end. Those issues are relevant to final-release assurance but are not used here to explain behavioural results. Their cause cannot currently be established from the supplied evidence.

The recommendations are proposed methodological extensions. They have not been adopted or operationally validated by NeoMundi.

CONCLUSION

Barometers #6 and #7 show why aggregate stability is an incomplete description of runtime behaviour. Fewer complete non-normal observations coexist with more factual involvement and more combined alerts. The small stability decline is localised primarily in one profile; Question 4 remains the principal semantic-variation concentration, while Question 2 becomes the leading explicit-flag location.

NeoMundi has made those structures visible without converting them into verdicts. The next requirement is to expose concentration, recurrence, alert lineage and task context strongly enough for the evidence to enter institutional review without acquiring undeclared authority.

The measurement layer should remain evidential. The authority-bearing institution must determine whether continued reliance remains legitimate.

The average tells us that the campaign is mostly normal.

The distribution tells us where governance must begin asking better questions.