Inference from Telemetry: measurement, aggregation, and epistemic risk[1]

Daniil Mitrofanov

MSc (CS), Independent researcher. ORCID: 0009-0006-4725-1458.

December 20, 2025

Abstract. Observability practice in modern software operations treats telemetry-metrics, logs, and distributed traces-as actionable evidence about systems that are only partially observable and continuously changing. This preprint argues that these telemetry modalities are not merely different data formats but distinct epistemic instruments. Each implements its own measurement operations, supports its own inferential shortcuts, and tends to fail in systematically different ways. Building on philosophical work on evidence and measurement, the paper proposes an "evidence profile" framework for telemetry: a structured way to characterize what a telemetry source can justify, what it can only suggest, and how it can mislead under aggregation, sampling, and instrumentation constraints. The result is a set of conceptual tools intended to clarify why observability disputes arise (e.g., "the metrics say it's fine" vs "the traces show it's broken") and how teams can set rational standards for belief, action, and uncertainty during incidents.

Keywords: observability; monitoring; telemetry; evidence; epistemology of measurement; justification; defeaters; distributed systems; software operations; SRE, DevOps.

[1] **Affiliation**: This work was carried out in the context of industry practice; the author's employer cannot be disclosed due to NDA. The analysis and conclusions are those of the author.

Inference from Telemetry: measurement, aggregation, and epistemic risk

An epistemology of metrics, logs, and traces

1. Introduction

Production engineering relies on telemetry to form beliefs about what a system is doing and to justify actions under time pressure - mitigate, rollback, page, or (sometimes) wait. Yet observability is routinely experienced as epistemically fragile: teams may have abundant signals while lacking warranted confidence about causal structure, blast radius, or the likely effects of interventions. In practice this fragility shows up as a familiar pattern of disagreement: "the metrics say it's fine" while traces (or a small set of logs) suggest a tail of failures concentrated in a particular dependency path. Qualitative studies of monitoring and incident response in distributed systems describe closely related difficulties as common in contemporary DevOps and microservice environments, including organizational and instrumentation constraints that shape what can be seen, compared, and acted upon.[2]

This paper's guiding question is not how to monitor, but: what kind of knowledge (or lesser epistemic status) is produced when engineers observe a system via metrics, logs, and traces?

In what follows, §3 fixes a minimal vocabulary for telemetry as evidence, §4 applies it to metrics, logs, and traces (via Table 1), and §§5–6 turn that diagnosis into reusable "evidence profiles" and some non-prescriptive lessons for observability practice

2. Related work

In epistemology, the concept of evidence is central to how beliefs become justified and how justification can be defeated by further information. A common distinction is between rebutting defeaters (which support the opposite conclusion) and undercutting defeaters (which target the support relation between evidence and conclusion).[3] In philosophy of science and philosophy of

---

[2] For qualitative evidence that observability breakdowns and "interpretation gaps" are common in distributed systems operations - and that organizational boundaries and uneven instrumentation constrain what teams can infer during incidents - see Niedermaier et al., On observability and monitoring of distributed systems: An industry interview study (2019).

[3] Pollock distinguishes rebutting defeat (evidence for not-H) from undercutting defeat (evidence that undermines the support relation between E and H, often by targeting reliability or applicability of the inference). See Pollock, "Defeasible Reasoning," Cognitive Science 11(4) (1987), 481–518.

measurement, measurement is not merely "reading off" facts. It is an activity structured by representational choices, operational conventions, and model-based constraints - all of which affect reliability, meaning, and what inferences a measurement can license.[4]

In computing practice, observability discourse often distinguishes metrics, logs, and traces as complementary "pillars," each offering a different evidential perspective on system behavior (for example, aggregates, discrete events, and request paths).[5] Empirically oriented software engineering and socio-technical research likewise emphasizes that observability is not only a technical problem: system complexity, organizational boundaries, and operational dynamics routinely outstrip simplistic assumptions about what monitoring data can establish in incident response.[6]

3. Conceptual framework: telemetry as evidence

To answer the guiding question, some terminology needs to be fixed up front.

Terminological commitments

In this paper, evidence is any informational item that can rationally justify belief or action in an operational hypothesis, defeasibly (i.e., in a way that can be rebutted or undercut). A measurement operation is the concrete procedure that produces a telemetry stream (what is counted/timed/recorded, under what inclusion rules, with what transformations). An undercutter is information that weakens the support relation between E and H by targeting reliability or applicability; a rebutter supports not-H. An evidence profile is a structured summary of what a telemetry source is designed to support, how it supports it, and how it fails. Epistemic risk is the risk of forming or acting on an unjustified belief due to predictable failure modes in that evidential chain (aggregation, sampling, drift, missingness, topology change).

3.1 Evidence roles in operations

Telemetry functions as evidence in at least three roles, roughly tracking familiar roles from epistemology:

---

[4] For the view that measurement depends on operationalization, representational choices, and model-based constraints (and is not mere "read off"), see Tal, "Measurement in Science" (SEP).

[5] For a practitioner statement of the "three pillars" framing (metrics, logs, traces) and their typical roles, see Rusinowicz, "Pillars of observability explained: Logs, metrics, and traces" (2025).

[6] See note 2.

- Justification role: a telemetry item makes it reasonable to believe some proposition about the system (e.g., “the error rate is elevated”).
- Defeater role: a telemetry item weakens or breaks the evidential connection between another signal and a hypothesis (e.g., “the metrics spike is explained by a sampling change,” an undercutter).[7]
- Action-guidance role: telemetry is used not only to support belief but to justify decisions (page, rollback, scale). This raises a standards question: how strong must the evidence be to warrant costly or risky interventions?

A minimal normative idealization is to treat telemetry as evidence that updates credences in operational hypotheses. This does not assume that incident response is Bayesian in practice. It is a convenient way to make explicit two things that are otherwise left implicit: which hypotheses are being compared, and which telemetry items are doing the evidential work.

Written explicitly, this idealization just says: given some prior degree of belief in a hypothesis and some assumption about how likely a telemetry pattern is if the hypothesis holds, new telemetry can shift that degree of belief in a rule-governed way. On a Bayesian picture, a telemetry item E updates the degree of belief in a hypothesis H according to Bayes’ rule:

$$P(H \mid E) = \frac{P(E \mid H)\,P(H)}{P(E)}$$

Formula 1: Bayes’ rule (posterior credence).

The posterior $P(H \mid E)$ depends both on how plausible H was before seeing the telemetry and on how likely it is that E would be produced if H were true.

Often the operational question is comparative rather than absolute - not “is *H* true?”, but “is *H1* more credible than *H2* given *E*?”. In that case, the update is naturally expressed in odds form:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)} \cdot \frac{P(H_1)}{P(H_2)}$$

Formula 2: Posterior odds and Bayes factor decomposition.

Formula 2 simply rewrites the same update in odds form, isolating the Bayes factor as the part that measures how sharply this telemetry pattern discriminates between two operational stories. In observability terms, the action is in the Bayes factor: the same hypothesis pair can look sharply

[7] See note 3.

separated under traces and barely separated under aggregates, because the modalities control how discriminating actually is. Telemetry is not a direct readout - it is emitted, transformed, sampled, and sometimes silently degraded. So the framework needs a place for information that says: "treat this update as weak", rather than "treat it as support". **Example:** *a p95 latency metric can stay flat while traces show a new dependency hop dominating a small but user-visible cohort; the modalities disagree because they slice the request population differently.*

This is also where epistemic risk enters: the same operational decision can be rational under one evidence profile and irrational under another, because our estimate of the Bayes factor is only as good as the production conditions of E. Modality-specific undercutters therefore matter not as "noise," but as predictable ways in which P(E|H) becomes miscalibrated - for example, aggregates undercut by cohort shifts, logs undercut by rate limits, traces undercut by biased sampling or broken propagation. So, they effectively drag the Bayes factor back toward 1, softening how strongly E favors one hypothesis over another.

At this point, defeaters become easy to represent. An undercutter can be modeled as information U that primarily lowers the evidential force of E by changing its expected reliability (for example, learning that E was produced under sampling drift or an instrumentation change). In such cases, U tends to act on the likelihood term - it reshapes P(E|H) - rather than directly supplying evidence for not-H.[8]

3.2 Measurement operations and representational choices

The Bayesian idealization above is intentionally thin: the update depends on likelihoods, and likelihoods depend on how telemetry is produced and processed. To understand when telemetry is strong evidence - and when it is fragile or misleading - we need to treat telemetry as the output of measurement operations.

[8] See note 3. Here the point is that undercutters typically reduce the probative force of E by undermining the reliability/production conditions relevant to the likelihood term.

Telemetry is produced by instrumenting code paths, exporters, agents, and sampling rules. Each telemetry stream thereby embeds:

- A measurement operation (what is counted, timed, or recorded, and under what conditions).
- A scale and transformation invariances (what it means to compare values across time, hosts, or services).
- A model of relevance (what the instrumentation assumes will matter).

In ops we treat “latency”, “availability”, and even “error” as if they named stable properties. In practice they name engineered quantities. What they mean is fixed by what we chose to count, bucket, sample, and exclude.

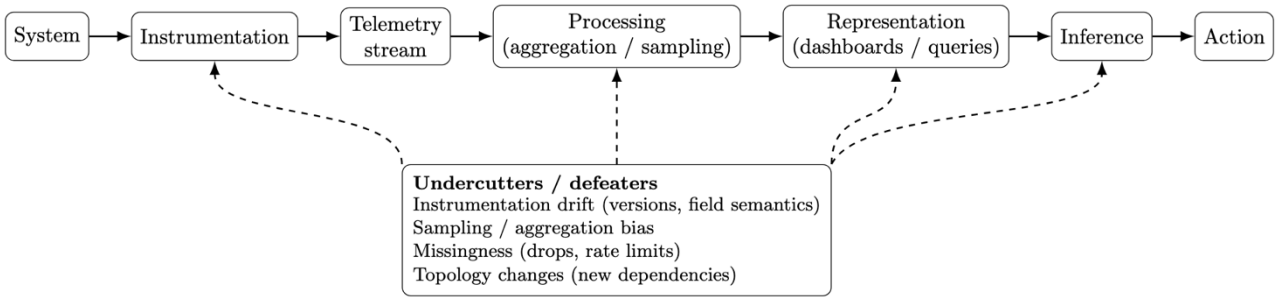

Figure 1. Telemetry-to-belief pipeline. The arrows mark typical undercutters by stage (instrumentation drift, sampling/aggregation bias, missingness, topology change), showing how evidence can fail before it reaches dashboards and decisions.

The figure also lets us pin typical undercutters to *where* they enter the chain, instead of listing them abstractly:

- Instrumentation drift - a deploy that keeps the metric name but changes what the field means.
- Sampling or aggregation bias.
- Missingness (drops, rate limits).
- Topology changes (new dependencies).

Methodologically, Figure 1 can be read as a defeater checklist: when telemetry and experience diverge, ask which stage of the measurement pipeline has changed, since evidence in observability is produced (not given) and can fail systematically along the way.

4. What is epistemically distinctive about metrics, logs, and traces?

This section operationalizes the thesis by distinguishing the modalities along stable dimensions captured in Table 1. Each modality makes some questions cheap to answer and others awkward or impossible - not by accident, but because it bakes in a measurement operation, a default aggregation story, and a characteristic set of failure modes.

Evidence-profile field	**Metrics**	**Logs**	**Traces**
Target propositions	Population-level rates and levels (error rate, saturation, SLO burn, availability over a window)	Discrete events and state transitions (a specific error occurred, config applied, invariant violated)	Per-request / path-specific behavior (where latency accrues, which dependency dominates, which span fails)
Inference patterns	Statistical / comparative: trends, thresholds, baselines, anomaly detection; strong for "is it elevated?"	Diagnostic / abductive: explain symptoms by matching event patterns, signatures, exemplars	Structural / causal-leaning: reconstruct request paths, identify critical path and dependency edges
Defeaters	Instrumentation or label drift; aggregation window changes; sampling changes; missing series / staleness; conflicting metrics	Log-level changes; rate limits/drops; format drift; missing correlation IDs; clock skew; contradictory exemplars	Sampling bias (head/tail); broken context propagation; partial instrumentation; span loss; alternative trace narratives
Resolution and scope	Coarse time buckets; wide scope across many requests/hosts; limited cohort/path distinctions under aggregation	Event-level; medium scope (per host/service/component), often episodic; scope depends on correlation fields	Request-level; end-to-end scope across services (when propagation works); path structure with missing-span limits
Stability across change	Moderate: stable if semantics are versioned; fragile to definition/label changes	Lower: message formats and levels change easily; meaning often informal	Mixed: stable for cross-service structure under standards; fragile to propagation and sampling policy changes

Table 1: Evidence profiles of metrics, logs, and traces, organized by the evidence-profile fields.

The point of Table 1 is methodological as well as rhetorical: it turns an intuitive "three pillars" story into a checkable analytical artifact, and it supports the claim that the modalities differ in epistemic profile rather than in presentation alone. These contrasts are not incidental. They are exactly what an evidence profile is meant to record: what the modality targets, what it licenses inferentially, what defeats it, and what kinds of distinctions it cannot stably represent under change. Section 5 turns this diagnosis into a reusable schema.

4.1 Metrics: compressive, comparative evidence

As Table 1 suggests, metrics behave like compressive measurements: they trade local detail for stable comparability across time and populations. Metrics typically:

- Aggregate across time windows and populations (rates, histograms, percentiles).
- Enable comparative reasoning (regression against baseline, SLO tracking).
- Risk masking heterogeneity (tail behavior, specific cohorts, correlated failures).

Epistemically, metrics are strong for trend detection and population-level "state claims" (for example, that error rates are elevated over a window), but weak for singular causal narratives unless paired with substantial background assumptions about topology, workload mix, and instrumentation semantics. Their characteristic failure modes are measurement-model mismatch and aggregation-induced invisibility: the system can be broken for the relevant users or paths while the aggregate remains "fine." **Example:** *a single noisy tenant can be effectively down for an hour while only a thin cohort in one percentile bucket moves, leaving global p95 flat.*

This is exactly the kind of fragility highlighted by measurement philosophy - meaning depends on the operation performed, the scale used, and the model that links the quantity-term to the underlying phenomenon (see Table 1: Characteristic failure modes; Temporal granularity and scope).

4.2 Logs: event-structured, narrative evidence

Logs are usually:

- Event-labeled and context-rich but selectively emitted.
- Retrospective in a strong sense: they record what developers chose to make recordable.
- Vulnerable to omission and semantic drift (for example, what counts as "INFO" or "ERROR" changes across versions).

Logs are well suited to narrative reconstruction - "what happened?" - but epistemically brittle when coverage is uneven, context propagation is inconsistent, or message semantics shift without governance. In evidential terms, logs can supply strong rebutting evidence (for example, an explicit exception record that rules out a competing hypothesis). They also admit characteristic undercutters: learning that logging was disabled, rate-limited, or sampled can downgrade the probative force of an apparent absence of events and can even undermine positive log evidence if the emission conditions have changed (see Table 1: Typical undercutters; Typical rebutters).

Unlike distributed tracing, logs often preserve local semantic detail but do not, by default, preserve cross-service structure.

4.3 Traces: path-structured evidence and causal temptation

Traces are distinctive because they represent end-to-end structure: they bind observations across services into request paths. Distributed traces:

- Connect spans across components into a single execution path.
- Invite causal readings ("this span caused that delay"), even when what is directly encoded is temporal ordering plus propagation conventions.
- Are shaped heavily by sampling, context propagation, and instrumentation boundaries.

Epistemically, traces trade on a causal appearance. They enable counterfactual-friendly questions ("if we remove this hop, would latency drop?"), but only via background models that interpret timing, concurrency, retries, queueing, and missing spans. The resulting causal temptation is a predictable hazard: tracing can make causal conclusions feel immediate even when the evidence supports only weaker claims about correlation, ordering, or localization of delay along a path (see Table 1: Main inference pattern; Typical undercutters).

Many trace undercutters target exactly these assumptions - for example, broken context propagation or biased sampling can fabricate or erase apparent bottlenecks.

5. The "evidence profile" proposal

Rather than treating telemetry as three interchangeable "pillars," I propose characterizing each telemetry source by an evidence profile with the following fields:

- Target propositions: what kinds of claims the source is designed to support (state, event, path, dependency).
- Inference patterns: what reasoning it reliably licenses (trend inference, anomaly detection, narrative reconstruction, localization).
- Defeaters: recurrent undercutters and rebutters (instrumentation changes, sampling shifts, missing context).
- Resolution and scope: what distinctions it can and cannot represent (cohorts, time granularity, topology).
- Stability across change: how sensitive it is to deployments, schema drift, and evolving semantics.

The motivation is simple: evidence is not merely "data that exists," but what can rationally justify belief (and action) given the possibility of defeat and error.[9] And for telemetry in particular, measurement operations and model-based conventions are what give quantity-terms their meaning and reliability. The evidence profile is intended as a practical bridge notion: it connects philosophical constraints on justification and measurement to concrete choices in observability engineering.

6. Epistemic (non-prescriptive) implications for observability practice

The evidence-profile framework is descriptive, but it has immediate implications for how observability work is understood and organized:

- Telemetry governance as epistemic governance: instrumentation, sampling, and schema decisions determine what can count as evidence later, and what will predictably generate defeaters.
- Incident rationality: persistent disagreements during incidents often track differences in evidence profiles (what each source can justify), not merely interpersonal conflict or "intuition."
- Designing for defeaters: systems should expose meta-telemetry (sampling rates, drop counts, rate limits, versioned schemas) so that undercutting defeaters can be made explicit rather than discovered ad hoc.

These points fit with empirical accounts in which observability breakdowns are socio-technical: strategy, roles, and responsibilities shape whether telemetry is interpretable and action-guiding.[10]

7. Limitations and scope

This preprint is conceptual and aims at analytic clarity, not a new tracing algorithm. It also brackets AI-based observability, focusing instead on the epistemic status of the core telemetry modalities under ordinary engineering constraints. A further limitation is that it does not settle debates about the "correct" theory of evidence. Instead, it relies on a minimal common core: evidence is whatever can rationally play a justificatory role (and be defeated), and measurement is sensitive to operations, conventions, and models.

---

[9] For evidence as what can rationally justify belief (defeasibly), see Kelly, "Evidence" (SEP). For rebutting vs undercutting defeat, see Pollock (1987).

[10] For socio-technical accounts of observability breakdowns in practice, see Niedermaier et al. (2019).

References

- Ratiu, D., ROHLINGER, T., STOLTE, T. and WAGNER, S. Towards an argument pattern for the use of safety performance indicators. In: CECCARELLI, A., BONDAVALLI, A., TRAPP, M., SCHOITSCH, E., GALLINA, B. and BITSCH, F. (eds.). Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops – DECSoS, SASSUR, TOASTS, and WAISE, Proceedings. Cham: Springer, 2024, pp. 160–172. ISBN 978-3-031-68737-2.
- POLLOCK, J. L. Defeasible reasoning. Cognitive Science. 1987, 11(4), pp. 481–518. DOI 10.1207/s15516709cog1104_4.
- NIEDERMAIER, S., KOETTER, F., FREYMANN, A. and WAGNER, S. On observability and monitoring of distributed systems: An industry interview study online. 2019. arXiv:1907.12240. Available from: <https://doi.org/10.48550/arXiv.1907.12240> accessed 20 December 2025.
- KELLY, T. Evidence online. In: The Stanford Encyclopedia of Philosophy (Winter 2014 Edition). 2006/2014. Available from: <https://plato.stanford.edu/entries/evidence/> accessed 20 December 2025.
- TAL, E. Measurement in science online. In: The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). 2015/2020. Available from: <https://plato.stanford.edu/entries/measurement-science/> accessed 20 December 2025.
- RUSINOWICZ, K. Pillars of observability explained: Logs, metrics, and traces online. CodiLime Blog, 4 December 2025. Available from: <https://codilime.com/blog/pillars-observability-explained-logs-metrics-traces/> accessed 20 December 2025.

Acknowledgements

I am grateful to Prof. S. Wagner whose comments helped refine several conceptual distinctions. Responsibility for the final content rests solely with the author.