Research Article

Can You Trust the Graph Beneath the Agent? A Reliability Metrology for Enterprise Skills Graphs: The FACTS Dimensions and the Skills Graph Reliability Index

Sathyaraj Asaithambi[1]

1. Independent researcher

Enterprise skills graphs, which connect skills to the people who hold them and the work that requires them, have become the data substrate on which agentic HR systems recommend hiring, mobility, reskilling, and redeployment. A consensus has formed among HR analysts that the value of these systems depends less on the sophistication of the agents than on the trustworthiness of the graph beneath them. That trustworthiness, however, is discussed almost entirely in qualitative terms. Where it is quantified at all, it is reduced to a single proxy: coverage, the percentage of roles or people for which some skill data exists. Coverage answers how much data an organisation holds. It is silent on whether that data can be trusted enough to let an autonomous agent act.

This paper develops the missing measurement science. Working within a design-science method and drawing on established data-quality standards (ISO/IEC 25012, ISO 8000, the FAIR principles) and on machine-learning calibration theory, it defines a reliability construct for skills graphs along five dimensions: Freshness, Accuracy of stated confidence, Contestability, Traceability, and Stability, which form the mnemonic FACTS. The dimensions are derived from the conditions under which an automated decision is defensible rather than assembled as a list, and each is operationalised as a metric computable on a graph's nodes and edges. They are combined into a dashboard-ready Skills Graph Reliability Index (SGRI) through a gated, deny-overrides aggregation rather than a weighted mean, so that weakness on any one dimension cannot be concealed by strength on another. SGRI resolves to three decision bands: act, human-gate, or do not automate.

 doi.org/10.32388/CQ9G1L

The paper makes five contributions. To begin with, it argues that unreliability in a skills graph compound rather than averages out under agentic use, because agent inferences written back without provenance contaminate the evidence against which later inferences would be validated, a mechanism named here evidential circularity. Second, it demonstrates that mechanism computationally. In a simulation of 20,000 edges over twelve cycles, unlabelled write-back caused measured calibration error to improve while true calibration error degraded, and the divergence increased with the write-back rate. At a write-back rate of 0.50 the contaminated graph reported a lower calibration error than a fully labelled control while its true error was approximately 4.8 times worse, so an organisation comparing the two on measured calibration would have preferred the corrupted graph. Third, the paper specifies the SGRI aggregation and decision bands and defends the deny-overrides rule on grounds of asymmetric loss. Fourth, it states the boundary of the construct explicitly: SGRI measures the trustworthiness of the evidence base, not the quality of inference drawn from it. Fifth, it confronts the ground-truth problem, since calibration in this domain depends on an oracle that is itself biased, and proposes triangulation and mandatory oracle-provenance disclosure as partial remedies. Fairness is treated as a stratification requirement across all five dimensions rather than as a sixth dimension. A reference architecture, a worked illustration, and a set of falsifiable propositions complete the paper. No claim is made that the index predicts decision quality; that relationship is stated as a falsifiable proposition and remains untested.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction: the trust gap beneath agentic HR

The enterprise conversation about artificial intelligence in human resources has shifted during 2026 from the capabilities of AI agents to the quality of the data on which they stand. The Josh Bersin Company's agentic vision for HR in 2030 describes an architecture of up to 130 agents and 95 distinct HR-focused agent capabilities, and it urges organisations to take an architectural approach in order to avoid what it calls agent sprawl. The HR-technology analyst Steve Goldberg puts the same claim in structural terms, arguing that as AI becomes embedded across HR technology, competitive advantage will come less from the AI model itself and more from the quality of the ontology and the human-capability data beneath it, so that, in his words, all ontologies are simply not created equal. The practitioner mainstream has taken up the theme as well: an April 2026 episode of the Digital HR Leaders podcast framed its

discussion with a people-analytics leader at NBCUniversal around the question of why strong data foundations still matter more than advanced analytics.

The object at the centre of this consensus is the skills graph, a network that links the skills present in an organisation, the people who hold them, and the roles, projects, and learning opportunities that demand them. Unlike a static inventory of self-reported competencies, a skills graph is updated continuously by inference from the work an organisation already produces, and it encodes relationships such as adjacency, substitution, and decay as first-class structure. Because agents now read, reason over, and increasingly act on this graph, its reliability is a governance question rather than a technical footnote.

The field has almost no way to answer that governance question quantitatively. When organisations measure the state of their skills data, they overwhelmingly report coverage: the percentage of roles or employees for which some skill data exists. Survey evidence exposes how thin this proxy is. In a 2026 survey of 227 enterprise HR leaders conducted by JDXpert, only 23.7 percent of enterprises had captured at least 75 percent of their job and skills data. Among that high-coverage group, roughly 40 percent were not confident in the accuracy of the data they had assembled, and 92.6 percent were planning major changes to how they govern job and skills information. JDXpert names the pattern the High-Coverage Paradox. The organisations that had won the coverage race were the ones most determined to rebuild, because coverage had told them how much data they had accumulated while saying nothing about whether they could trust it.

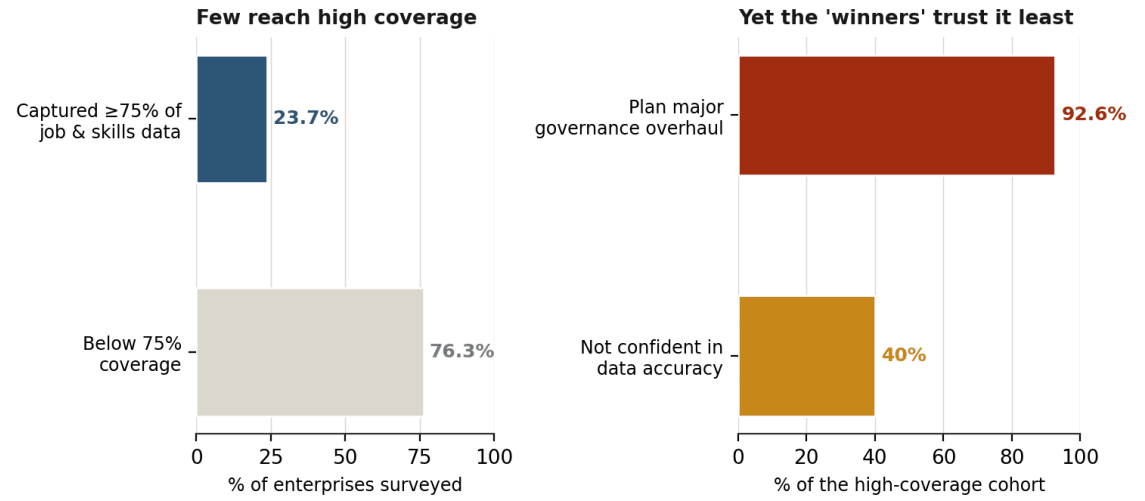

Figure 1. *The High-Coverage Paradox. Few enterprises reach high coverage of job and skills data, and among those that do, most are not confident in its accuracy and are planning to rebuild governance. Data from JDXpert*[1]*, n = 227 enterprise HR leaders.*

This is the trust gap. Coverage is a measure of quantity. What leaders need, before they permit an agent to act, is a measure of reliability. The two are not equivalent and treating them as though they were hazardous: a graph may cover ninety percent of a workforce with stale, uncalibrated, and unattributable inferences and be less trustworthy than a graph covering half the workforce with fresh, well-calibrated, and fully traceable ones. What is missing is a metrology for the skills graph, which is to say a principled and computable account of when the graph beneath the agent can be trusted.

This paper develops that metrology and makes five contributions. To begin with, Section 4 argues that unreliability in a skills graph does not average out under agentic use but compounds, and Section 5 demonstrates the central mechanism computationally, which together establish why the correct instrument is a gate rather than a score. Second, Section 6 derives a reliability construct from the conditions under which an automated decision is defensible, producing five dimensions, *Freshness, Accuracy of stated confidence, Contestability, Traceability, and Stability*, which form the mnemonic FACTS, and treats fairness as a stratification requirement across all five. Third, Section 7 states the boundary of the construct, since an instrument is only as credible as its declared scope. Fourth, Section 8 confronts the problem that calibration depends on a ground-truth oracle that is itself biased. Fifth, Section 9 combines the dimensions through a non-compensatory rule and maps the result to three decision bands. Sections 2 and 3 situate the contribution and define the object of measurement. Section 10 gives a

reference architecture and a worked illustration, Section 11 draws implications, and Section 12 states the propositions by which the framework can be refuted.

2. Related work and the gap

Four bodies of work bear on the reliability of skills graphs. Each contributes something necessary. None supplies a measurement science for the graph itself.

2.1. Skills graphs, ontologies, and inference engines

A substantial practitioner and vendor literature now describes how skills graphs and ontologies are built: how nodes and edges are defined, how proficiency is inferred from work artefacts rather than from self-report, and how adjacency and substitution are modelled in order to support mobility and planning. This work has matured the artefact considerably, and some of it acknowledges that inference introduces uncertainty and that skills decay. Its centre of gravity, however, is construction and application rather than evaluation. It offers no shared, computable account of how reliable a given graph is, and the reliability language it does use, such as better ontologies or living maps, remains qualitative.

2.2. Data-quality standards and frameworks

The information-systems and data-management literatures provide rigorous and general frameworks for data quality. The multidimensional account of Wang and Strong[2] established that quality is what data consumers judge along dimensions such as accuracy, completeness, timeliness, and believability, organised into intrinsic, contextual, representational, and accessibility categories. ISO/IEC 25012 and ISO 8000 codify data-quality models and management processes, and the FAIR principles have become a common language for trustworthy data stewardship. A recent review by Guillen-Aguinaga and colleagues[3] synthesises this literature alongside the ISO standards and reports that accuracy, completeness, consistency, timeliness, and accessibility recur as universal quality dimensions across sectors. These frameworks are foundational and the present paper builds on them directly. They are, however, domain-general. They do not tell a practitioner how to compute timeliness for an inferred skill edge whose relevance decays at a rate specific to the skill, how to test whether a graph's stated confidence matches reality, or how to combine such measures into a decision about whether an agent may act.

2.3. People-analytics methodology and data-quality critique

Within people analytics, a growing number of studies have questioned the reliability of the data on which the field depends. Evidence-based reviews by Marler and Boudreau[4] and by van den Heuvel and Bondarouk[5] examined the maturity of the discipline and found that claims frequently outran the evidence supporting them. Survey work continues to report the same weakness from the practitioner side: in HR.com's 2025 to 2026 research, only 42 percent of HR professionals agreed that their HR technology systems were well integrated enough for easy data analysis, and 32 percent disagreed. This literature is diagnostically useful, since it names the problem. It is narrative and retrospective, however. It does not propose a computable index, and it predates the agentic context that makes the question urgent.

2.4. Machine-learning calibration and graph-quality metrics

Computer science supplies the technical primitives that this paper repurposes. Calibration theory provides principled ways to test whether predicted confidences match observed frequencies, chief among them the Brier score[6] and expected calibration error[7]. The graph-learning literature offers measures of stability and robustness under perturbation, and the entity-resolution literature, synthesised by Christen[8], provides the machinery for deciding when two records describe the same person. These tools are mature. They were developed for classifiers, for generic graphs, and for record linkage, and they have not been assembled into a reliability instrument for enterprise skills graphs or connected to the governance decision that such an instrument ought to inform.

2.5. The gap, stated plainly

Placed side by side, the four literatures leave a clean opening. Vendors build skills graphs but do not measure their reliability. Data-quality standards measure quality in general but not for inferred skill edges or for the decision about whether to act. People-analytics scholarship names the reliability problem but does not compute it. Machine-learning calibration can test confidence but has not been brought to the skills graph. A computable, multidimensional reliability metrology for enterprise skills graphs, one that produces a defensible signal about whether an agent may act, does not appear to exist. That instrument is this paper's contribution. Table 1 summarises the position.

Literature	What it offers	What it leaves unaddressed
Skills graphs, ontologies, inference engines	Methods to build and apply the graph; adjacency, substitution, inferred proficiency	No shared, computable measure of how reliable a given graph is
Data-quality standards[2][9][10][11]	Rigorous general quality dimensions and governance processes	Not specialised to inferred skill edges, skill-specific decay, or the decision to act
People-analytics methodology critique	Diagnoses weak reliability and reporting in the field	Narrative and retrospective; no computable index; predates the agentic context
ML calibration, graph quality, entity resolution	Principled tests of confidence[6] (ECE), stability measures, record linkage	Developed for classifiers and generic graphs; never assembled for skills graphs or governance

Table 1. The four adjacent literatures and what each leaves unaddressed

3. The skills graph as the object of measurement

To measure the reliability of a skills graph, the object must be defined precisely. A skills graph is a directed, attributed multigraph in which nodes represent skills, people, and opportunities such as roles, projects, and learning assets, and in which edges represent relationships among them. Three edge families dominate. Holds edges connect a person to a skill, typically with an inferred proficiency and a confidence score. Requires edges connect an opportunity to the skills it demands. Relates edges connect skills to one another as prerequisites, adjacencies, or substitutes, and form the ontological backbone that allows the graph to reason that a person one step from a role is more redeployable than a person five steps away.

What makes reliability non-trivial is that most consequential edges are inferred rather than asserted. A holds edge is rarely a fact that an employee typed into a form. It is a probabilistic claim produced by a model that read work artefacts, and it therefore carries three latent properties that any reliability metrology must interrogate: a confidence, which may or may not be justified; an age, which matters because skills and the evidence for them decay; and a provenance, which records the source and

transformation that produced the edge and which may or may not be recoverable. Figure 2 shows the object and these latent properties.

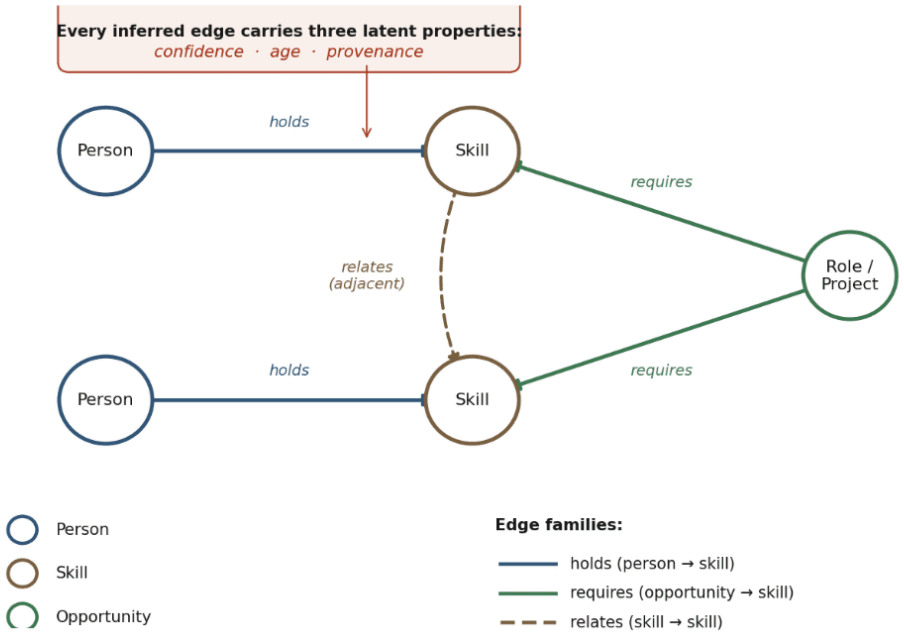

Figure 2. *The skills graph as the object of measurement. Three edge families connect people, skills, and opportunities. Every inferred edge carries a confidence, an age, and a provenance, and the reliability construct interrogates all three.*

A reliability construct for skills graphs is, in essence, a disciplined set of questions about these latent properties across the graph as a whole. Before those questions are posed, however, it is worth establishing why they are urgent now, and why the answer must take the form of a gate rather than a score.

4. Why unreliability compounds under agentic use

A natural objection to the entire enterprise of this paper runs as follows. Skills-graph inference is noisy, but noise averages out; models are improving quickly; therefore, reliability will take care of itself, and the sensible course is to wait for better models rather than to build measurement apparatus. This section argues that the objection is mistaken, and that the error matters because it determines the shape of the instrument the field needs.

The central claim is that unreliability in a skills graph does not average out under agentic use. It compounds. A more capable agent operating on an unreliable graph produces more confident, more

automated, and more widely propagated errors than a less capable one. Three mechanisms produce this result, and they are shown together in Figure 3.

Silent inference. An agent traversing a graph that contains stale edges, miscalibrated confidences, or unresolved duplicate identities does not fail loudly. It infers over the defects as though they were sound and returns an answer with the fluency it would apply to clean data. Capability increases the fluency and the authority of the output. It does not increase the correctness of the input. An agent that cannot inspect the provenance of an edge cannot flag that the edge was untrustworthy, and therefore presents a conclusion drawn from degraded evidence as confident guidance.

Ungoverned write-back. Agentic systems read the graph and they also act on it. Their actions generate new assertions, such as an inferred skill, a suggested successor, or a predicted readiness score, which are written back into the graph or into the systems that feed it. When these writes carry no label distinguishing model inference from assessed fact, they enter the graph indistinguishable from evidence. The next system to traverse the graph consumes yesterday's inference as today's data.

Evidential circularity. The third mechanism is the most damaging and the least discussed, and it is named here for that reason. Calibration, the second FACTS dimension, requires a validation sample of edges whose truth has been established independently. Once agent-generated assertions have been written back into the graph without provenance labels, they begin to appear inside that validation sample. The model is then calibrated against evidence that its own earlier outputs helped to produce. Measured calibration improves while actual calibration degrades, and the instrument that was meant to detect the failure is the instrument being corrupted. Evidential circularity is therefore not merely one more quality defect. It is a mechanism by which a reliability measurement silently stops measuring reliability.

Evidential circularity is related to, but distinct from, several phenomena already described in the machine-learning literature. The degradation of models trained on their own generated output concerns the corruption of a model's capability through its training data. Contamination between training and evaluation sets concerns a methodological error, usually static and correctable once identified. Performative prediction concerns cases in which a prediction alters the reality it predicts. Concept drift concerns change in the world that a model fails to track. Evidential circularity differs from each in a way that matters for governance: it corrupts neither the model nor the world but the record against which the model would be checked. Reality may be unchanged and the model may be unchanged, and yet the measurement that would reveal a problem can no longer do so. It is also endogenous and continuous

rather than exogenous or one-off, since it is produced by the ordinary operation of the system rather than by an error in study design. This is why it belongs in a paper about measurement rather than one about model quality. The failure it produces is a failure of oversight.

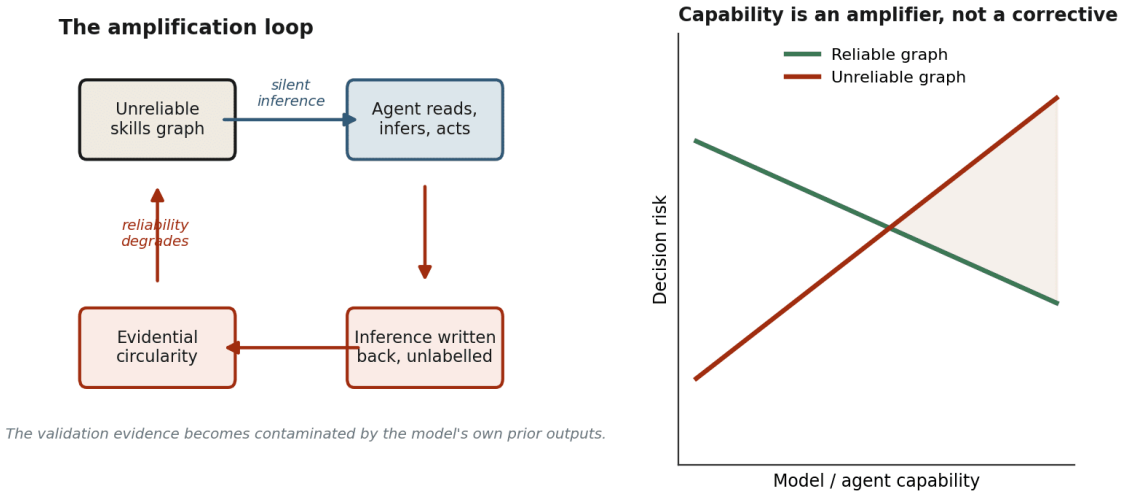

Figure 3. *Unreliability compounds under agentic use. Left: agent outputs re-enter the graph unlabelled and contaminate the evidence against which later inferences would be validated. Right: on an unreliable graph, added capability increases decision risk rather than reducing it. The right panel is a schematic of the argument, not fitted data.*

Two consequences follow, and both shape the instrument proposed in this paper. The first is that the returns to model capability are bounded by the reliability of the graph, so that past a point, further investment in capability on an unreliable graph may increase risk rather than reduce it. The second is architectural. If unreliability compounds through use, then a reliability instrument must be capable of withholding permission, not merely of reporting a score that a determined programme can average its way past. This is the reasoning behind the deny-overrides aggregation specified in Section 9, and behind the requirement in Section 8 that the provenance of the validation oracle be disclosed alongside any calibration figure.

5. A computational demonstration of evidential circularity

The argument of Section 4 is mechanistic. Evidential circularity in particular is a claim about a process that would be difficult to observe directly in a production system, because an organisation experiencing

it would by construction be unable to detect it. This section therefore tests the mechanism in simulation, where the latent truth is known and measured calibration can be compared against actual calibration. The purpose is to determine whether the proposed mechanism produces the predicted divergence under stated assumptions. It is not field evidence, and Section 12 states plainly what a simulation of this kind can and cannot establish.

5.1. Method

A synthetic graph of 20,000 person-skill edges was constructed. Each edge carries a latent true state, indicating whether the person holds the skill, which drifts with probability 0.03 per cycle in order to represent people gaining and losing capability. Each edge also carries an evidence pool of at most twelve observations. In each cycle, genuine evidence arrives for 18 percent of edges and reflects the current latent state with an observation error of 0.12. The model computes a confidence for each edge as the Laplace-smoothed mean of its evidence pool. An agent then acts, and a fraction w of its point estimates is written back into the evidence pool as an additional observation carrying no label distinguishing it from genuine evidence. The organisation draws its validation sample from the evidence pool and cannot separate genuine from synthetic entries, which is precisely the condition that provenance labelling would prevent.

Two quantities were measured in each cycle. Measured expected calibration error compares model confidence against the validation labels available to the organisation. True expected calibration error compares model confidence against the latent state, which only the simulation knows. Four conditions were run, with w set to 0.00, 0.15, 0.30, and 0.50, each averaged across five random seeds. The condition w equal to zero is the control and represents complete provenance labelling, under which no synthetic observation enters the evidence pool. Every condition was preceded by eight burn-in cycles with no write-back, so that all conditions begin from a comparably warmed evidence pool and the results are not an artefact of a cold start.

5.2. Results

The predicted divergence occurred, and it was dose-dependent. In the control condition, measured and true calibration error both improved over the twelve measured cycles and no divergence appeared, which confirms that the simulation does not generate the effect in the absence of contamination. In every condition with unlabelled write-back, measured calibration error improved while true calibration error

degraded. Figure 4 shows both the divergence over cycles at a write-back rate of 0.30 and the final-cycle comparison across all four conditions.

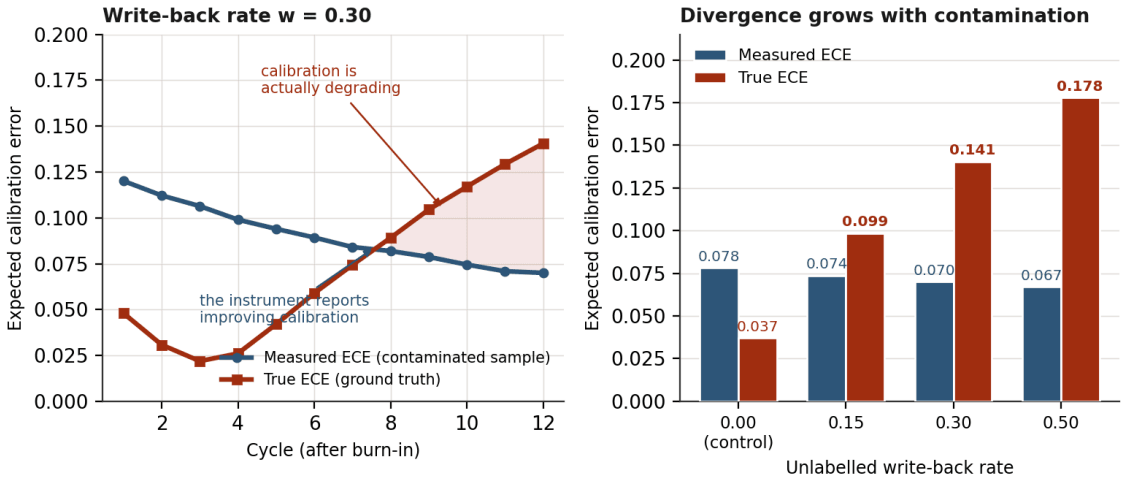

Figure 4. *Evidential circularity demonstrated in simulation. Left: at a write-back rate of 0.30, measured calibration error falls while true calibration error rises, and the two cross at cycle eight. Right: at the final cycle, measured error declines monotonically with the write-back rate while true error rises monotonically. Control condition w = 0.00 represents complete provenance labelling. Values are averages across five seeds.*

Table 2 gives the final-cycle figures. The pattern is that the instrument becomes more reassuring exactly as the graph becomes less trustworthy. The most consequential result is the comparison between the control and the most contaminated condition. At a write-back rate of 0.50, measured calibration error was 0.067, which is lower than the control's measured error of 0.078, while true calibration error was 0.178 against the control's 0.037, a factor of approximately 4.8. An organisation comparing these two graphs using the calibration measurement available to it would therefore have judged the heavily contaminated graph to be the better calibrated of the two. Evidential circularity does not merely blunt the instrument. In this simulation it inverts the ranking the instrument produces.

doi.org/10.32388/CQ9G1L

Unlabelled write-back rate	Measured ECE	True ECE	Ratio true to measured	Synthetic share of evidence
0.00 (control)	0.078	0.037	0.48	0.0%
0.15	0.074	0.099	1.34	20.7%
0.30	0.070	0.141	2.01	35.6%
0.50	0.067	0.178	2.65	52.4%

Table 2. Simulation results at the final cycle. Averages across five seeds, 20,000 edges, twelve cycles after burn-in.

5.3. What this does and does not establish

The simulation establishes that the mechanism is internally coherent and that it produces the predicted divergence under the stated assumptions, with a dose-response relationship and a clean control. It does not establish that evidential circularity occurs at these magnitudes in production systems, because the parameters governing drift, evidence arrival, observation error, and write-back rate were chosen to be plausible rather than estimated from field data. The result is best read as a demonstration that the failure mode is real in principle and is capable of inverting a calibration measurement, which is sufficient to justify treating provenance labelling as a control rather than as a documentation preference. Proposition 3 in Section 12 states the field test that would establish the magnitude empirically.

6. The reliability construct: the five FACTS dimensions

6.1. Deriving the dimensions

A construct presented as a list invites the objection that the list is arbitrary, and a construct presented as an acronym invites the sharper objection that the acronym came first and the dimensions were selected to fit it. Both objections are fair against any framework that does not show its derivation. This subsection therefore derives the five dimensions rather than asserting them, and Table 3 shows where the most commonly proposed alternatives fall once the derivation is applied.

The derivation starts from the decision rather than from the data. Suppose an agent is about to act on a claim held in the graph, for example that a named person holds a named skill at a stated level and

suppose that action must later be defended to an auditor, a regulator, or the person affected. Five questions must be answerable, and each corresponds to a distinct way in which the defence can fail.

Is the claim still true? A claim that was accurate when recorded may have ceased to be accurate through skill decay or role change. This is temporal validity, and its measure is Freshness.

Is the stated certainty warranted? A claim carrying a confidence of 0.9 must be correct approximately nine times in ten, or the confidence is not information but decoration. This is epistemic validity, and its measure is Accuracy of stated confidence.

Can an error in the claim be caught and corrected? A system whose errors cannot be surfaced and repaired will accumulate them, and the person best placed to detect an error about a person is that person. This is corrigibility, and its measure is Contestability.

Can the claim be traced to its basis? A claim that cannot be connected to the evidence and transformation that produced it cannot be examined, and Section 5 has shown that it also cannot be validated honestly. This is evidential validity, and its measure is Traceability.

Would the same evidence produce the same claim? A claim that changes between runs on unchanged inputs cannot be reproduced by an auditor and cannot support a defensible decision. This is procedural validity, and its measure is Stability.

The five dimensions are therefore the five distinct ways in which the defence of an automated decision fails: the claim is stale, the certainty is unwarranted, the error is uncorrectable, the basis is untraceable, or the result is irreproducible. They are not five desirable properties of data selected from a longer list. That the initials happen to form a usable mnemonic is convenient and should be treated as no more than that.

This derivation also explains the treatment of the properties most often proposed as additions. Table 3 applies it to seven candidates. Three are already present under different names, two belong to a different construct, one is the quantity this paper argues against using as a proxy for reliability, and one, fairness, is genuinely absent as a dimension and is treated in Section 6.7 as a stratification requirement applying across all five.

Candidate	Disposition	Reasoning
Robustness	Already present as Stability	Sensitivity of the edge set to benign input perturbation is what Stability measures
Consistency	Largely present as Stability and Traceability	Internal contradiction surfaces as instability across runs or as conflicting lineage
Completeness	Belongs to coverage, not reliability	A missing edge is an absence of claim; reliability concerns claims the graph does make
Coverage	Explicitly excluded	Coverage is the quantity this paper argues has been mistaken for reliability (Section 1)
Interoperability	Belongs to FAIR stewardship, not reliability	Whether a graph exchanges cleanly with other systems does not bear on whether its claims are true
Explainability	A property of the model, not of the graph	Concerns how an inference is rendered intelligible, which sits above the evidence layer
Fairness	Genuinely absent as a dimension; addressed as stratification	Not a sixth failure mode but a question asked of all five (Section 6.7)

Table 3. Commonly proposed alternative dimensions and their disposition under the derivation

Each dimension is stated below as the question it answers, the property it interrogates, and an example metric. The metrics are illustrative rather than final, since the contribution is the construct and its operationalisation, and the precise estimators are a matter for the validation work set out in Section 12.

6.2. F, Freshness: currency against decay

Freshness asks whether the graph's edges are current, given that skills and the evidence for them decay. The property interrogated is the age of an edge relative to a half-life specific to its skill family. The practical half-life of a technical skill tied to a particular tool or framework is short, while the same skill label can sit unchanged in a system for a decade. A freshness score can be defined as the percentage of decision-relevant edges whose age is below the estimated half-life of their skill family, weighted so that

recently reconfirmed edges count fully and stale edges decay toward zero. Freshness therefore reframes timeliness for a setting in which the correct expiry is not a fixed calendar window but a property of the skill itself.

6.3. A, Accuracy of stated confidence: calibration

Calibration asks whether the confidence a graph attaches to an inferred edge matches reality. The property interrogated is the agreement between predicted confidence and observed correctness on a validation sample of edges whose truth has been established independently. Expected calibration error or a Brier score computed over such a sample provides the metric. This dimension is the one most often absent in practice and the most consequential for autonomy, because an agent that treats a poorly calibrated confidence of 0.9 as though it were a justified 0.9 will act decisively on claims that are frequently wrong. Calibration, rather than raw accuracy alone, is what makes a confidence score safe to reason with. The dimension also carries the heaviest methodological burden in the construct, since it depends on an oracle whose own quality is contestable, and Section 8 is devoted to that problem.

6.4. C, Contestability: human challenge and correction

Contestability asks whether the people the graph describes can see, challenge, and correct the claims it makes about them, and whether those challenges are resolved. The property interrogated is the existence and health of an appeal path running from an inferred edge to human adjudication. A composite metric can combine whether a contest mechanism exists, the rate at which contested edges are resolved, and the latency of resolution.

This dimension requires an explicit defence, because it is not of the same kind as the other four. Freshness, calibration, traceability, and stability are properties of the artefact. Contestability is a property of the socio-technical system surrounding the artefact. A reviewer is entitled to ask whether it belongs in the construct at all, or whether it was included because the mnemonic required a letter. Three arguments support its inclusion, and the first is decisive.

To begin with, contestability is the principal generator of ground truth in this domain. Calibration cannot be computed without a validation sample of independently established edges, and in the skills domain no natural label exists. Contest events produce exactly such labels, and they produce them continuously, at the points where the graph's claims matter most to the people they describe. Contestability is therefore not an ethical addition to a technical construct. It is a *precondition* for the second dimension. A graph

without a contest channel cannot be calibrated honestly, and a metrology that omitted contestability would be unable to sustain its own calibration measure over time. Second, although the mechanism is organisational, the measurement is not: contest events attach to specific edges, so contest rate, resolution rate, and latency are edge-level quantities computable on the graph in the same way as the other four dimensions. Third, and least central to the argument, contestability carries a normative weight that the governance audience already recognises, since human oversight of high-risk employment systems is an explicit regulatory expectation.

The construct is therefore deliberately heterogeneous, and the heterogeneity is a design decision rather than an oversight. Four dimensions measure the artefact. One measures whether the organisation retains the capacity to correct it. A construct restricted to the artefact alone would be internally elegant and practically inert, because it would measure a graph's condition while ignoring the only mechanism through which that condition can be established and improved.

6.5. T, Traceability: provenance

Traceability asks whether every consequential edge can be traced back to the evidence and the transformation that produced it. The property interrogated is the completeness of lineage, which is to say the source, the method, the model version, and the timestamp recorded for each edge. A provenance-completeness score equal to the percentage of decision-relevant edges carrying full and resolvable lineage provides the metric, and it can be graded by provenance depth. Traceability is the dimension that makes the other four demonstrable, since a freshness or calibration claim is only as credible as the audit trail that substantiates it. It is also the dimension whose failure enables evidential circularity, because the write-back contamination described in Section 4 is possible only where inferred assertions are not labelled as such.

6.6. S, Stability: reproducibility under re-inference

Stability asks whether re-running inference on the same inputs produces the same graph, and whether the graph is robust to small and benign perturbations of its inputs. The property interrogated is the reproducibility and the sensitivity of the edge set across repeated or slightly perturbed runs. The Jaccard similarity of decision-relevant edge sets across re-runs provides one metric, and a drift rate measuring how much the graph changes when inputs change trivially provides another. An unstable graph cannot

be governed, because if the same evidence produces a different answer each time it is processed, no downstream decision built on it is defensible and no audit can reproduce it.

6.7. Stratification: fairness as a requirement across all five dimensions

A graph can be fresh, well calibrated, contestable, traceable, and stable in aggregate while failing on every one of those dimensions for a particular population. Aggregate reliability scores conceal subgroup failure by construction, because an average computed across a workforce is dominated by its largest groups. A framework that certified such a graph as reliable would be certifying something the affected population would not recognise as reliable at all.

Fairness is not therefore a sixth dimension. Adding it as one would be a category error, since the derivation in Section 6.1 identifies five ways a decision defence fails and inequity is not a sixth way but a pattern in the distribution of the other five. Fairness is instead a requirement about how the five are reported. Every FACTS dimension should be computed and reported disaggregated by the population groups relevant to employment decisions, and the reported dimension score should be the worst subgroup score rather than the aggregate whenever the two diverge materially.

Each dimension has a characteristic distributional failure. Freshness fails unevenly where evidence about some populations is refreshed less often, which occurs predictably for part-time staff, for those on leave, and for functions with less digitised work. Calibration failure by subgroup is well established in the machine-learning literature, and a model calibrated in aggregate may be systematically overconfident for groups underrepresented in its training evidence. Contestability fails unevenly where some populations are less able to exercise an appeal path, whether through unfamiliarity, language, or perceived risk in challenging a manager. Traceability fails unevenly where evidence for some kinds of work is inherently less documented. Stability fails unevenly where sparse evidence produces volatile inferences, which again concentrates on smaller groups.

Stratified reporting therefore turns the same five measurements into an equity instrument without expanding the construct. It also carries a practical implication for the deny-overrides rule specified in Section 9: if a dimension clears its floor in aggregate but falls below it for an identifiable population, the gate should close for decisions affecting that population. A reliability index that permitted automation over a group for whom the evidence is demonstrably unreliable would fail on its own terms.

Dimension	What it captures	Example metric	Governance anchor
F, Freshness	Currency of edges against skill-specific decay	Percentage of edges younger than their skill half-life, decay-weighted	Timeliness[9]; data relevance[12]
A, Accuracy of confidence	Whether stated confidences match observed correctness	Expected calibration error or Brier score on a validated sample	Accuracy and robustness[12]
C, Contestability	Human capacity to challenge and correct inferred edges	Contest-path existence; resolution rate; resolution latency	Human oversight[12]; worker trust
T, Traceability	Recoverable lineage of each consequential edge	Provenance-completeness score; provenance depth	Record-keeping and logging[12]; FAIR
S, Stability	Reproducibility and low sensitivity of the edge set	Edge-set Jaccard across re-runs; drift rate under benign perturbation	Reproducibility; auditability

Table 4. The five FACTS dimensions: the question each answers, an example metric, and the governance obligation to which it speaks

7. The boundary of the construct: what SGRI does not measure

A measurement instrument is only as credible as its stated scope, and the most serious objection available against the construct developed here is that it measures the wrong thing. The objection runs as follows. Reliability, properly understood, ought to mean that the system produces good recommendations. A graph could score well on all five FACTS dimensions and still support poor decisions about mobility, succession, or redeployment, because nothing in the construct guarantees that the inferences drawn from the graph are correct. On this reading, FACTS measures data hygiene and calls it reliability.

The objection is partly right, and the correct response is to concede its valid part precisely rather than to expand the construct until the objection disappears. SGRI measures the trustworthiness of the evidence base on which an automated workforce decision rests. It does not measure the quality of the inference drawn from that evidence, the predictive validity of any model reading the graph, or the business

outcome of any decision the graph informs. Those are distinct constructs requiring distinct instruments, and a framework claiming to measure all of them at once would measure none of them well.

The analogy that clarifies the boundary is the quality management system of a clinical laboratory. Such a system establishes that specimens were correctly identified, that instruments were calibrated, that results are traceable to a run and an operator, and that the same specimen would produce the same result on repeat testing. It does not establish that the physician reading the result will reach the correct diagnosis. No one concludes from this that laboratory quality management is mere hygiene. It is understood instead as the precondition without which diagnostic reasoning has nothing dependable to reason from. The relationship between SGRI and decision quality is the same.

The claim that measurement-system quality is a precondition for decision quality is not novel to this paper. It is established practice in other fields that act on measured evidence. Manufacturing quality management assesses measurement system capability, through repeatability and reproducibility studies, before process decisions are taken on the resulting data, on the reasoning that a measurement system of unknown capability cannot support a defensible process decision. Clinical laboratories operate under accreditation regimes built on the same premise. What distinguishes the workforce setting is not that the principle is different but that no equivalent discipline exists. Organisations act on inferred claims about people without having established that the system producing those claims can support the action. The contribution of this paper is therefore better understood as importing an existing discipline into a domain that lacks it than as proposing a novel relationship between measurement and decision quality.

The claim the construct supports is therefore asymmetric and stating it in that form is both more defensible and more useful than a symmetric claim would be. A high SGRI does not guarantee that recommendations drawn from the graph are good. A low SGRI does establish that any good recommendation drawn from the graph was *fortuitous*, since it rested on evidence that was stale, miscalibrated, uncorrectable, untraceable, or irreproducible. Reliability in this sense is a necessary condition for defensible automated decisions rather than a sufficient one, and the three decision bands in Section 9 are constructed accordingly: the green band licenses action within a defined scope under oversight, and it does not certify that the action will be correct.

Stated more precisely, the proposition is not that reliability produces good decisions, but that reliability sets a ceiling on achievable decision quality. An organisation may fall well below the ceiling its evidence base permits, through poor models, poor process, or poor judgement. It cannot exceed it. This formulation is falsifiable in a way that a correlational claim is not: it predicts that decision quality and

reliability form a triangular rather than a linear relationship, with high-reliability organisations distributed across the full range of decision quality and low-reliability organisations bounded above.

Two consequences follow for how the instrument should be used. Organisations should not treat a green band as evidence that their workforce AI is performing well, because that question requires outcome measurement the index does not provide. Equally, they should not dismiss a red band on the grounds that recommendations currently appear satisfactory, because the argument of Section 4 and the simulation of Section 5 together establish that apparent satisfaction on an unreliable graph is not informative. Proposition 1 in Section 12 states the empirical test that would connect the two constructs, and until that test is conducted the connection remains a hypothesis rather than a finding.

8. The ground-truth problem

The calibration dimension depends on a validation sample of edges whose truth is known independently. In the skills domain, no such sample occurs naturally, and the sources available as substitutes are imperfect in ways that matter. This section treats the problem directly rather than relegating it to a limitation, because the credibility of the second FACTS dimension, and therefore of the index as a whole, depends on how honestly it is handled.

The most commonly proposed oracle is manager attestation. A manager confirms or rejects an inferred skill claim, and the confirmed set becomes the validation sample. Manager attestation, however, is not ground truth. It is a judgement subject to halo effects, to recency, to the manager's own limited visibility of work performed outside their direct observation, and to organisational incentives that make certain attestations more comfortable than others. Calibrating a model against a biased oracle produces a calibration figure that is itself biased, and reporting expected calibration error against such an oracle without qualification would overstate what has been established. The same caution applies in different degrees to the alternatives, each of which carries a characteristic distortion, as summarised in Table 5.

Oracle	What it establishes	Characteristic distortion	Partial mitigation
Manager attestation	Whether a manager endorses the inferred claim	Halo and recency effects; limited visibility; incentive pressure	Blind attestation; multiple attestors; inter-rater agreement reported
Structured assessment or credential	Demonstrated proficiency under test conditions	Narrow construct coverage; expensive; may lag practice	Reserve for high-stakes skill families; report coverage of the sample
Work-artefact evidence	That the person produced work requiring the skill	Attribution ambiguity in collaborative work	Require corroborating artefacts; weight by attribution confidence
Peer endorsement	Recognition by colleagues who observe the work	Reciprocity and popularity effects	Use only as a corroborating signal, never as sole oracle
Contest events	A disputed or confirmed claim raised by the person described	Self-selection; disputes cluster on consequential claims	Model the selection explicitly; report contest-sample composition

Table 5. Candidate ground-truth oracles for the calibration dimension, their characteristic distortions, and partial mitigations

Three practices follow. To begin with, no single oracle should stand alone. Triangulation across independent sources, with agreement between them reported rather than assumed, produces a validation sample whose weaknesses are at least visible. Second, attestations should be treated as noisy labels rather than as truth, which places the problem within an established methodological literature on learning and evaluation under label noise, and permits the residual bias to be modelled rather than ignored. Third, and most important for practice, any reported calibration figure should be accompanied by a disclosure of the oracle that produced it, which this paper calls **oracle provenance**. An expected calibration error of 0.04 established against blind multi-attestor review is a different claim from the same figure established against single-manager sign-off, and an index that reported the two identically would be concealing the distinction that matters most.

The honest position is therefore bounded. The calibration dimension is only as good as the oracle behind it, this limit is not removable by better statistics, and the appropriate response is disclosure and triangulation rather than a claim of resolution. Recognising the limit also strengthens the case for contestability as a construct dimension, since contest events supply the one correction stream that is generated continuously by the people best placed to know whether a claim about them is true.

9. Constructing the Skills Graph Reliability Index

A five-dimensional profile is diagnostically rich but operationally awkward. Leaders, and increasingly agent policies, need a single signal that answers whether the graph may be acted upon. The Skills Graph Reliability Index converts the FACTS profile into that signal without discarding what the profile reveals.

9.1. Normalisation and the aggregation rule

Each dimension is normalised to a common scale from 0 to 100, with higher values indicating greater reliability. The central design choice is how to combine them. A weighted mean is rejected deliberately, because it permits precisely the failure the metrology exists to prevent. A graph could post a respectable average while harbouring a fatal weakness, such as near-zero calibration, concealed beneath strong freshness and traceability. Section 4 gives the deeper reason for the rejection: if unreliability compounds through use, then an aggregation that allows compensation between dimensions will license exactly the automation that causes the compounding.

SGRI therefore uses a gated, deny-overrides aggregation, in which the composite is bounded above by the weakest dimension, so that no strength can compensate for a dimension falling below its floor. Figure 5 shows the effect on the illustrative Graph B introduced in Section 10: the weighted mean of 46 conceals a calibration score of 22, whereas the deny-overrides rule returns 22 and places the graph in the red band.

A statistician may reasonably object that a minimum function is brittle. Consider a graph scoring 95, 94, 92, 91 and 59 across the five dimensions against a floor of 60. A rule driven by the raw minimum would place that graph in the red band on the strength of a single score one point below the floor, which seems disproportionate, and small movements near a threshold would produce large swings in the reported band. The objection is well taken against a naive minimum, and the aggregation is therefore specified more carefully than that.

The rule has three parts. Each dimension carries its own floor, since the dimensions are not equally consequential and a common floor would be an arbitrary simplification. The composite is the minimum

of the dimension scores, which preserves the non-compensatory property. The band, however, is determined by the relationship between each dimension and its own floor rather than by the composite alone. A dimension at or above its floor contributes no constraint. A dimension below its floor but within a defined margin places the graph in amber rather than red, so that action continues under human adjudication while remediation proceeds. A dimension below its floor by more than that margin places the graph in red. On this specification the graph scoring 59 against a floor of 60 lands in amber, and a graph scoring 22 against the same floor lands in red, which is the discrimination the objection asks for.

The deeper justification for non-compensatory aggregation is not statistical elegance but asymmetric loss. The cost of a false permit, meaning that an agent is authorised to act on a graph it should not have acted on, is borne by the people the decisions affect and is frequently irreversible, since a person who is not surfaced for a role does not learn that they were not surfaced. The cost of a false denial, meaning that automation is withheld from a graph that would in fact have supported it, is delay and additional human review. Where the loss function is this asymmetric, conservatism is the correct design, which is why deny-overrides aggregation is standard in access control and in safety-critical engineering rather than being peculiar to this paper. A weighted mean implicitly asserts that the two errors are exchangeable at the rate given by the weights, and in this domain they are not.

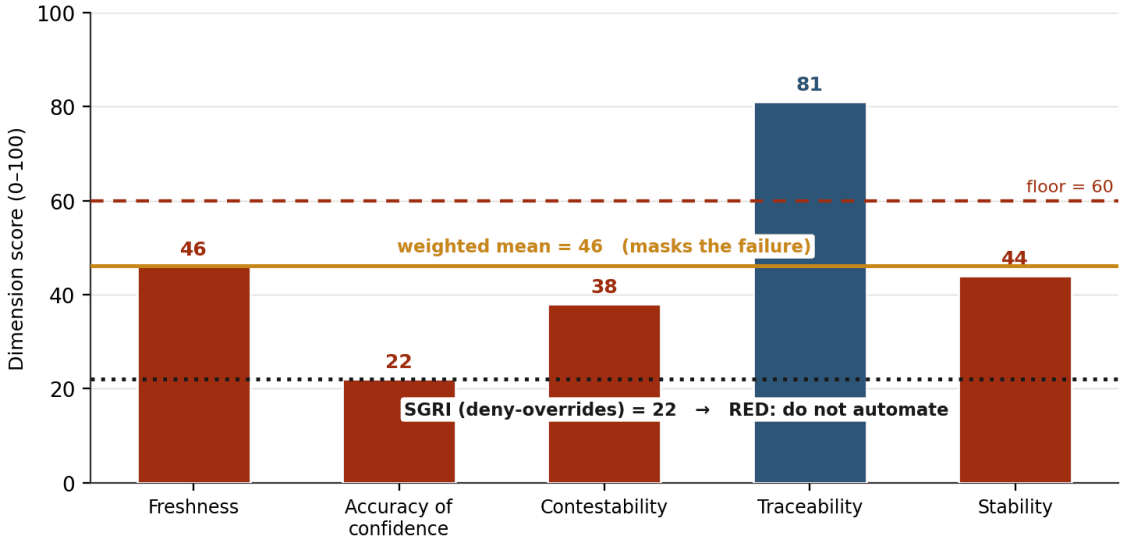

Figure 5. *Deny-overrides aggregation compared with a weighted mean, using the dimension scores of illustrative Graph B. The mean sits above three of the five dimension scores and conceals a calibration failure. The deny-overrides rule is bounded by the weakest dimension and returns a red band. Values are illustrative.*

doi.org/10.32388/CQ9G1L

9.2. Decision bands

SGRI resolves to three bands that map directly to an operating decision, as shown in Table 6. The thresholds are configurable to the criticality of the use case, since a graph feeding an advisory dashboard may act at a lower bar than one feeding an autonomous redeployment agent, and Section 12 treats their calibration as an empirical question.

Band	SGRI condition (illustrative)	Operating decision
Green	Every dimension is at or above its own floor, in aggregate and for every reported subgroup	Agents may act on the graph within a defined scope, under standing oversight
Amber	One or more dimensions fall below floor but within the defined margin, or a dimension clears in aggregate and fails for an identifiable subgroup	Human in the loop is required; the agent proposes and a named human decides. For a subgroup failure, the gate closes only for decisions affecting that population
Red	Any dimension falls below its floor by more than the margin (deny-overrides)	Do not automate; remediate the failing dimension before any agentic action

Table 6. SGRI decision bands. Thresholds are illustrative and configurable to use-case criticality

9.3. Why an index, and why a profile alongside it

SGRI is intended to travel. A single banded number is what a chief human resources officer can place on a readiness dashboard, what a vendor can be asked to meet, and what an agent policy can gate on. The index is always reported with its FACTS profile, however, and never instead of it, so that a red or amber result is immediately actionable. The profile names the failing dimension and therefore the remediation. The index answers whether the organisation may act. The profile answers what must be fixed if it may not. Figure 6 shows two contrasting profiles.

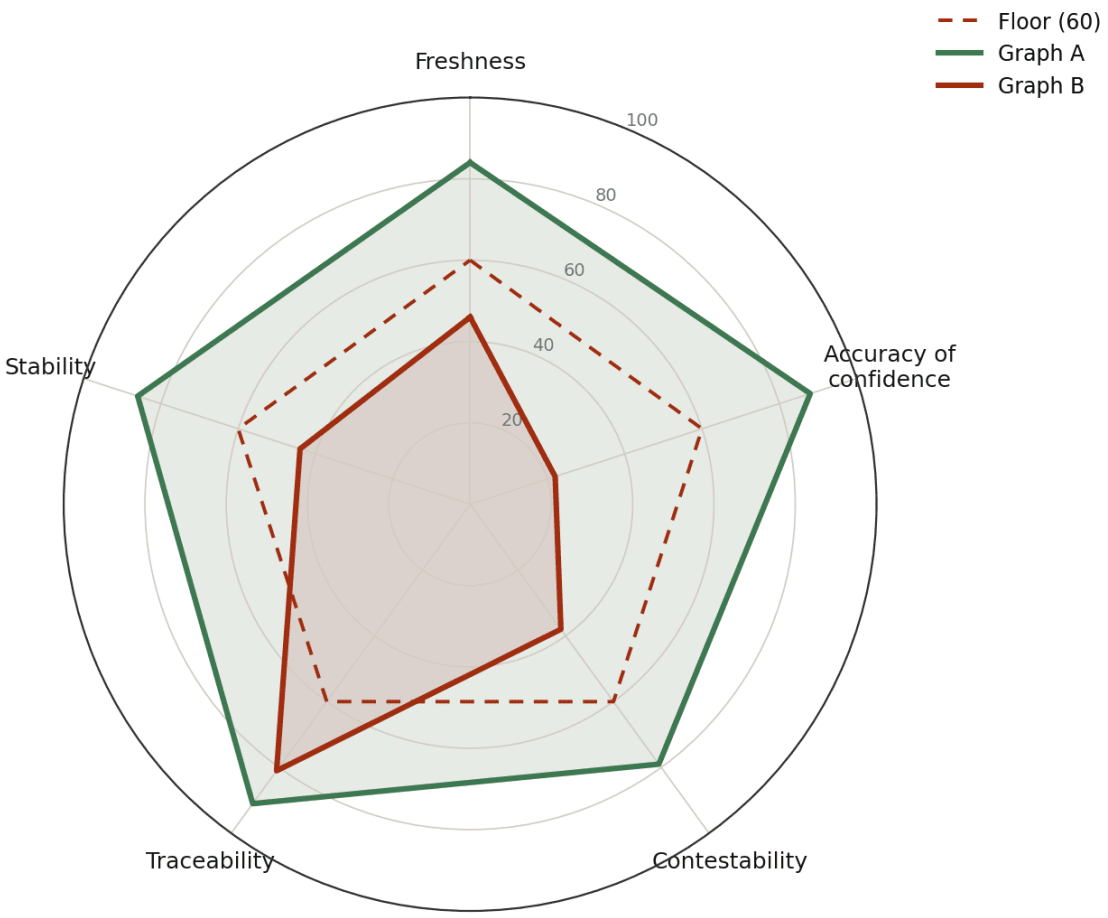

Figure 6. *Two contrasting FACTS profiles against a common floor. Graph A clears the floor on every dimension. Graph B posts a respectable traceability score while failing on calibration, freshness, contestability, and stability. Values are illustrative.*

10. Reference measurement architecture and worked illustration

10.1. The measurement layer

SGRI is computed by a measurement layer that sits beside the skills graph rather than inside any single vendor system, so that graphs of different provenance can be scored on comparable terms. The layer has four components, shown in Figure 7. A validation-sample manager maintains the independently confirmed edges needed to compute calibration and to anchor freshness half-lives and records the oracle provenance described in Section 8. A metric engine computes the five dimension scores on a defined slice of decision-relevant edges. An aggregator applies the deny-overrides rule and the configured thresholds

in order to produce the banded index. An evidence log records every score, the sample on which it was computed, and the graph version to which it applied, which makes the reliability measurement itself traceable in the spirit of the dimension it reports.

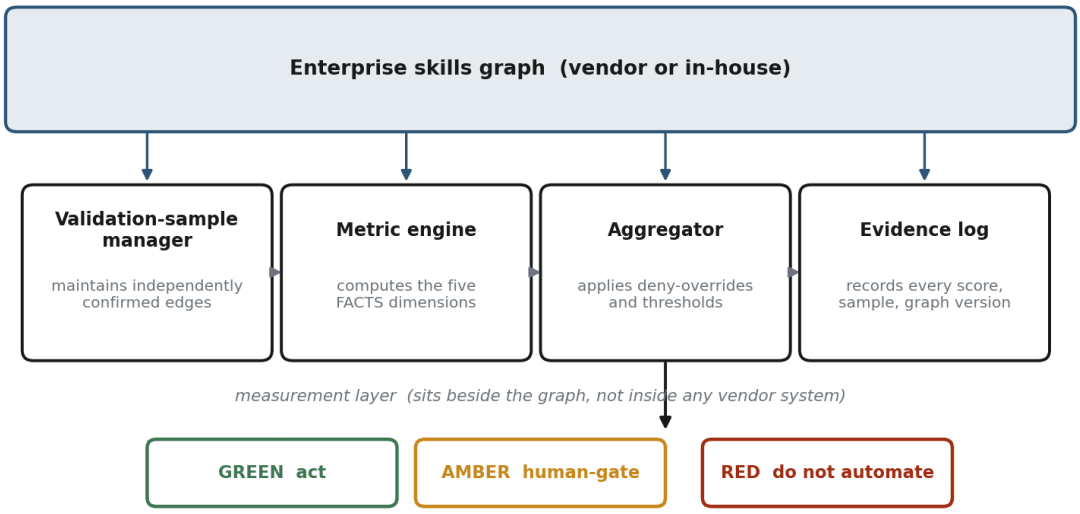

Figure 7. *Reference measurement architecture. The measurement layer sits beside the graph rather than inside a vendor system, and its evidence log makes the reliability measurement itself auditable.*

The ordering of these components is deliberate. The evidence log is built at the foundation rather than added afterwards, because the component that makes every other claim demonstrable cannot credibly be retrofitted.

10.2. Worked illustration

Consider two skills graphs for the same synthetic enterprise. Graph A is refreshed continuously, its inference confidences are validated each quarter against a triangulated sample with disclosed oracle provenance, employees can contest any inferred skill through a self-service path, every edge carries full lineage, and re-runs are near-identical. Graph B covers slightly more of the workforce and would therefore win on coverage. Much of its data is several years old, its confidences have never been checked against reality, contest requests are unresolved, and re-inference produces materially different edges each cycle. Table 7 gives the scores.

doi.org/10.32388/CQ9G1L

Dimension	Graph A	Graph B	What differs	Floor cleared
Freshness	84	46	B has not been refreshed against skill half-lives	A only
Accuracy of confidence	88	22	B′s confidences have never been validated	A only
Contestability	79	38	B′s contest requests are unresolved	A only
Traceability	91	81	Both record lineage reasonably well	Both
Stability	86	44	B′s re-inference produces materially different edges	A only
SGRI (deny-overrides)	79	22	Bounded by the weakest dimension	-
Band	GREEN	RED	A may act in scope; B must not automate	-

Table 7. Worked illustration. Two graphs for the same synthetic enterprise, scored on the five FACTS dimensions with a floor of 60. Values are illustrative

Graph B would have posted a weighted mean of 46, which reads as mediocre rather than dangerous, and its coverage advantage would have made it look like the stronger asset on the metric most organisations currently report. Under deny-overrides its index is 22 and it lands in the red band, and the profile points remediation squarely at recalibration, refresh, and the contest backlog. The illustration makes the paper′s central claim concrete: more data is not more trust, and only a reliability metrology distinguishes the two.

11. Implications

11.1. For chief human resources officers and HR leaders

SGRI provides leaders with an answer to a question they are increasingly asked and cannot currently answer with rigour, namely whether their workforce data is trustworthy enough to let AI act on it. It reframes the internal conversation from how much skills data an organisation holds to how reliable that data is and where specifically it is weak. It does so in a single banded signal backed by a diagnostic profile, which is the form a readiness discussion with a board or an audit committee requires.

11.2. For HR-technology vendors

If competitive advantage is migrating from the model to the data beneath it, as the analyst consensus holds, then vendors need a common yardstick on which that advantage can be demonstrated. SGRI offers a vendor-neutral standard against which a skills-graph capability can be evaluated and compared, which turns the claim that not all ontologies are equal into a measurement, and gives buyers a procurement instrument that looks past coverage claims to reliability.

11.3. For governance, risk, and responsible-AI functions

The five dimensions map onto obligations these functions already carry: data governance and relevance, accuracy and robustness, human oversight, and record-keeping. Because the measurement layer produces its evidence log as a by-product of operation, it supplies much of the documentation an internal audit or a regulatory data-governance review would seek. The mapping in Table 4 is offered as a reasoned correspondence rather than as legal interpretation, and organisations should take their own advice on the obligations that apply to them.

12. Validation design, limitations, and research agenda

12.1. Falsifiable propositions

A design-science contribution earns its standing by stating what would refute it. The framework developed here is therefore restated as five propositions, each with a predicted direction, so that the construct can be tested rather than merely discussed. Proposition 1 is the most consequential, since a failure there would undermine the practical case for the instrument even if the construct were internally sound.

P1. SGRI predicts decision quality. Across organisations, graphs in the green band support workforce recommendations judged higher in quality by independent human review than graphs in the amber and red bands, controlling for coverage and organisation size. Predicted direction: decision quality increases monotonically across red, amber, and green. Refutation: no monotonic relationship, or coverage predicting decision quality at least as well as SGRI once both are entered together. The principal threat to this proposition is confounding by organisational maturity. Organisations that invest in graph reliability are plausibly better run in ways that independently improve workforce decisions, so a simple correlation between SGRI and decision quality would be uninformative. A credible test therefore requires either

within-organisation variation, comparing decision quality across regions or functions whose reliability differs while organisational practice is held constant, or a design in which reliability is improved on one subgraph and not another and outcomes are compared before and after.

P2. SGRI outperforms simpler proxies. SGRI predicts decision quality better than coverage, completeness, or provenance-completeness alone. Predicted direction: incremental predictive validity of SGRI over each single proxy is positive. Refutation: a single dimension, most plausibly calibration, predicting as well as the composite, which would indicate the construct should be collapsed rather than retained at five dimensions.

P3. Evidential circularity occurs in production. In systems where agent inferences are written back without provenance labels, the fraction of validation samples containing model-generated assertions rises over time, and measured calibration diverges from calibration recomputed against an independently established sample. Predicted direction: divergence increases with write-back rate, as in the simulation of Section 5. Refutation: no measurable synthetic contamination of validation samples, or divergence independent of write-back rate.

P4. The dimensions are non-redundant. The five dimensions load onto distinct factors rather than collapsing into fewer. Predicted direction: inter-dimension correlations are moderate and a five-factor structure fits better than a one or two-factor alternative. Refutation: high inter-correlation indicating that the construct measures fewer distinct properties than claimed.

P5. Stratified reliability differs from aggregate reliability. In organisations with heterogeneous workforces, at least one FACTS dimension clears its floor in aggregate while failing for an identifiable population. Predicted direction: subgroup minimum is materially below the aggregate for a non-trivial percentage of organisations. Refutation: aggregate and worst-subgroup scores are equivalent within measurement error, which would make the stratification requirement of Section 6.7 unnecessary.

12.2. Study designs

Construct validation. A Delphi or structured-interview study with people-analytics leaders, skills-graph architects, and data-governance specialists, in order to test P4 and to examine whether the derivation in Section 6.1 is judged sound. The status of contestability, defended in Section 6.4, is the element most in need of independent scrutiny.

Metric validation. On benchmark graphs, both synthetic and, where permitted, real, test whether each proposed estimator behaves as intended. The calibration metric should detect deliberately miscalibrated

confidences and the freshness metric should track injected staleness.

Oracle sensitivity. Measure how much the calibration score moves as the oracle is varied across the sources in Table 5, which would establish empirically how much of a reported calibration figure is attributable to the oracle rather than to the graph.

Field replication of the simulation. Instrument a production skills graph to label the provenance of every validation-sample entry, then track synthetic contamination and calibration divergence over successive cycles, testing P3 against the magnitudes reported in Section 5.

Threshold calibration. A study relating SGRI bands to downstream decision quality, testing P1 and P2 and setting defensible floors and margins by use-case criticality rather than by assertion.

12.3. Limitations

Five limitations bound the contribution. The first is the ground-truth dependency treated in Section 8, which triangulation and oracle-provenance disclosure mitigate but do not remove. The second concerns the simulation reported in Section 5: it demonstrates that evidential circularity produces the predicted divergence under stated assumptions, and its parameters were chosen to be plausible rather than estimated from field data, so it establishes the mechanism in principle and not its magnitude in practice. The third is that the construct is derived analytically, and although Section 6.1 supplies a derivation from the conditions of a defensible decision rather than an assertion, the derivation itself has not been tested against practitioner judgement, which is what Proposition 4 and the Delphi study are for. The fourth is that the thresholds, floors, and margins in Section 9 are illustrative and await the threshold-calibration study. The fifth is the most important and applies to the paper as a whole: the central practical claim, that reliability measured in this way predicts decision quality, is stated as Proposition 1 and has not been tested. Until it is, SGRI is a reasoned instrument rather than a validated one, and this paper does not claim otherwise.

An index also invites Goodhart's law. Once SGRI becomes a target, organisations may optimise the measure rather than the underlying reliability, and Section 5 shows that at least one dimension can be improved on paper while degrading in fact. This is the reason the metrology is specified with a traceable evidence log and periodic revalidation rather than as a self-reported figure, and the reason oracle provenance is a required disclosure rather than an optional one.

12.4. Research agenda

In addition to the propositions above, four lines of work would extend the contribution.

1. An open, cross-industry SGRI benchmark and reference dataset, so that reliability can be compared across organisations and over time, and so that floors can be set from distributional evidence rather than by assertion.
2. A formal treatment of evidential circularity beyond the skills-graph setting, since the mechanism applies to any system whose outputs re-enter the evidence base against which it is later validated and is therefore not specific to workforce data.
3. Extension of the metrology from skills graphs to the broader people graph, and integration of SGRI gating directly into agent orchestration policies so that the band governs execution rather than reporting.
4. Investigation of the cost and reliability trade-off, namely what freshness, calibration, contestability, and provenance cost to achieve, and where the returns to further investment diminish.

13. Conclusion

The industry has decided, correctly, that agentic HR will rise or fall on the trustworthiness of the skills graph beneath it. It has not yet given itself the means to measure that trustworthiness, and in the absence of such means it has fallen back on coverage, a measure of how much data exists which is silent on whether the data can be believed. The evidence that this substitution fails is already visible: the organisations with the highest coverage are the ones most determined to rebuild their governance.

This paper has proposed the missing measurement science. It has argued that unreliability compounds under agentic use rather than averaging out, and it has demonstrated in simulation that evidential circularity can corrupt the very measurement intended to detect the problem. In that simulation the most heavily contaminated graph reported the lowest calibration error of any condition tested while its true calibration error was several times the control's, so an organisation relying on the measurement available to it would have ranked the worst graph first. This is why the correct instrument withholds permission rather than reporting a score.

The paper has derived a reliability construct from the conditions under which an automated decision is defensible, rather than assembling one from a list of desirable data properties. It has defended the inclusion of contestability as a precondition for calibration rather than as an ethical addition, treated

fairness as a stratification requirement across all five dimensions rather than as a sixth, confronted the biased-oracle problem that constrains calibration in this domain, and combined the dimensions through a non-compensatory rule justified by asymmetric loss into a banded index that answers whether an agent may act. It has also stated the boundary of the construct plainly: a high index does not guarantee good recommendations, and a low index establishes that any good recommendation was fortuitous.

Coverage told organisations how much they had accumulated. Reliability tells them whether they can act on it, and, when they cannot, what to fix first. Making that distinction measurable is the step the field now needs, and the empirical validation of SGRI is the work this paper is written to invite.

Notes

This paper presents a measurement framework and a design specification. It is not a compliance instrument, and it is not legal advice. The Skills Graph Reliability Index and its five FACTS dimensions are offered as reasoning to be tested empirically, not as a certified standard. The paper is a companion to the author's earlier work on classifying workforce-intelligence systems under the EU AI Act. The views expressed are the author's own and are not associated with any employer.

Reproducibility

The simulation reported in Section 5 uses 20,000 edges, a latent-state drift probability of 0.03 per cycle, genuine evidence arriving for 18 percent of edges per cycle, an observation error of 0.12, an evidence pool capped at twelve observations, eight burn-in cycles with no write-back followed by twelve measured cycles, and five random seeds per condition. Expected calibration error is computed over ten equal-width confidence bins. The full parameterisation is given in Section 5.1, and the procedure is specified in sufficient detail to be reimplemented independently.

References

1. ^*JDXpert (2026). "The State of Job and Skills Data Governance 2026." JDXpert.*
2. a, b*Wang RY, Strong DM (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." J Manag e Inf Syst. 12(4):5–33. doi:10.1080/07421222.1996.11518099.*
3. ^*Guillen-Aguinaga M, Aguinaga-Ontoso E, Guillen-Aguinaga L, Guillen-Grima F, Aguinaga-Ontoso I (202 5). "Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles." Data. 10(12):20*

1. doi:10.3390/data10120201.

4. ^*Marler JH, Boudreau JW (2017). "An Evidence-Based Review of HR Analytics." Int J Hum Resour Manag. 28 (1):3–26.*
5. ^*van den Heuvel S, Bondarouk T (2017). "The Rise (and Fall?) of HR Analytics: A Study into the Future Appli cation, Value, Structure, and System Support." J Organ Eff. 4(2):157–178.*
6. a, b*Brier GW (1950). "Verification of Forecasts Expressed in Terms of Probability." Mon Weather Rev. 78(1):1–3.*
7. ^*Guo C, Pleiss G, Sun Y, Weinberger KQ (2017). "On Calibration of Modern Neural Networks." Proc Int Conf Mach Learn. 70:1321–1330.*
8. ^*Christen P (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Du plicate Detection. Springer. doi:10.1007/978-3-642-31164-2.*
9. a, b*International Organization for Standardization (2008). "ISO/IEC 25012:2008. Software Engineering, Sof tware Product Quality Requirements and Evaluation (SQuaRE), Data Quality Model." International Organiz ation for Standardization.*
10. ^*International Organization for Standardization (2016). "ISO 8000-61:2016. Data Quality, Part 61: Data Qua lity Management, Process Reference Model." International Organization for Standardization.*
11. ^*Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Sci Data. 3:160018. doi:10.1038/sdata.2016.18.*
12. a, b, c, d*European Parliament and of the Council (2024). "Regulation (EU) 2024/1689 of the European Parlia ment and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." Off J Eur Union. OJ L, 2024/1689.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.