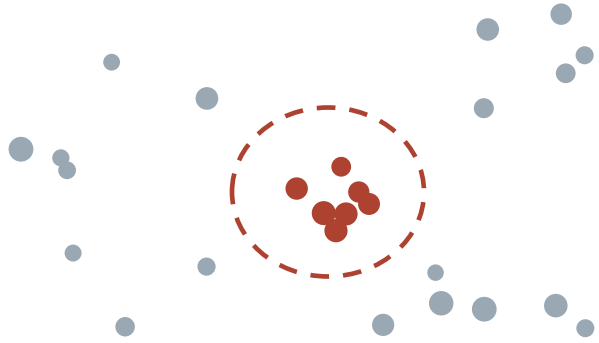
WORLD PRIVACY FORUM

The Object of Evaluation in Agentic AI

Authorship, grouping, and the provenance of authority

Pam Dixon · Founder & Executive Director

ITU-T SG17 · Session 3: Evaluating & Assessing Trustworthiness in Agentic AI · Geneva · 9 July 2026



We are evaluating the wrong object.

The framework asks how to evaluate the agent. The harm is not in the agent.

Component-based evaluates data, tools, and platforms. Process-based evaluates the perception–memory–planning–execution loop. Both audit the parts; neither reaches the grouping the system enacts across them.

a population — but where is the object of evaluation?

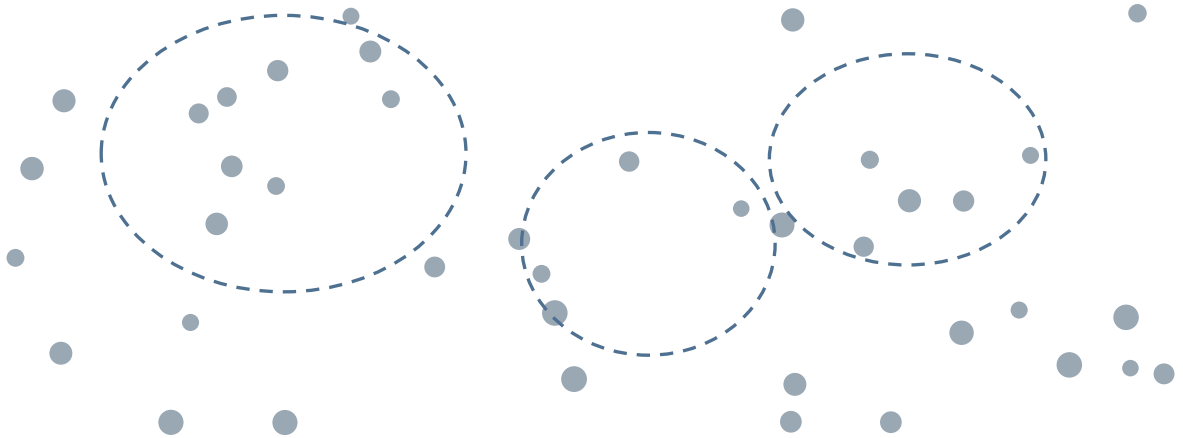
THE OBJECT

Trustworthiness lives in the grouping the system acts on — not in a component, and not in the loop.

You cannot discover the grouping.

No method recovers a canonical grouping from data.

Kleinberg’s impossibility theorem: no clustering function satisfies scale-invariance, richness, and consistency at once. Pick any two; the third fails.



same data, different groupings — none of them given by the data

THE FLOOR

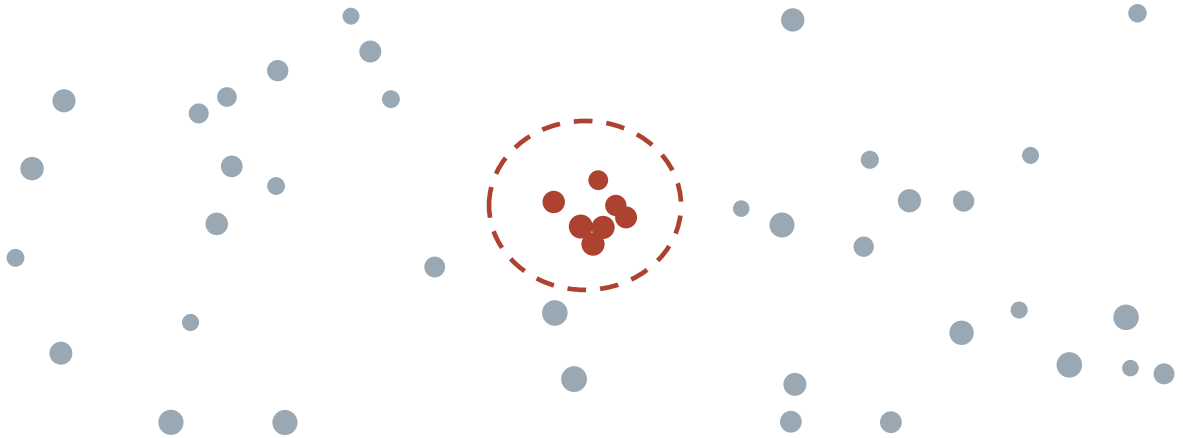
The grouping is not in the data. It has to be supplied from outside.

Kleinberg, J. (2002). “An Impossibility Theorem for Clustering.” Advances in Neural Information Processing Systems 15.

So the group is authored, not discovered.

If grouping cannot be discovered, every protected category is an act of authorship.

Authorship moves the problem from mathematics to legitimacy — who declares the category, under what mandate. Legislative naming spends the scale choice deliberately.



the ring is drawn, not found — placed from outside

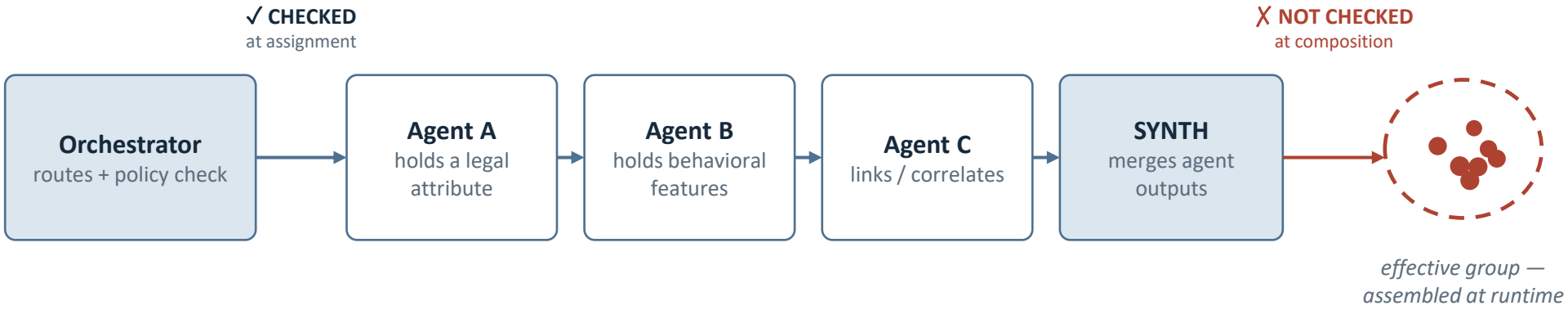
THE STAKE

Absent an author, the scale defaults to the optimizer — and the choice flows to whoever runs the system.

In a multi-agent system, no one holds the group.

The effective group is not stored anywhere. It assembles at runtime, across agents.

One agent holds a legal attribute, another holds behavioral features, a third links them. No single agent holds the harmful grouping — it forms along the path between them. The runtime proxy is single-linkage chaining at the protocol layer.



WHERE THE HARM LIVES

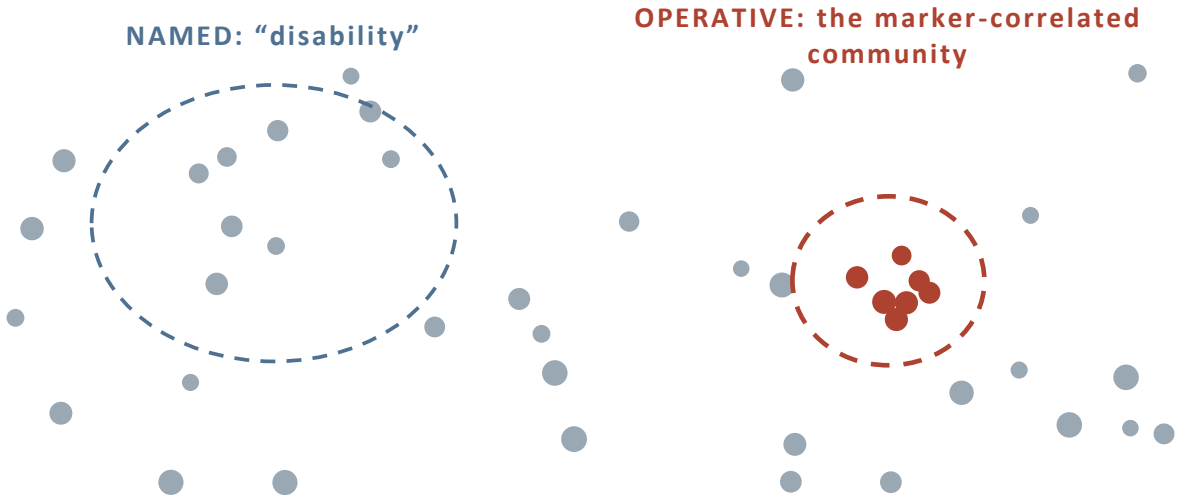
No single component carries the harm — it lives in the composition, exactly where component-based evaluation cannot look.

On chaining: Carlsson & Mémoli (2010), JMLR 11. On group–label divergence: Peel, Larremore & Clauset (2017), Science Advances 3(5).
Operational reference — Federation of Agents: Giusti et al. (CERN; L. Serio), 2025. arXiv:2509.20175.

Worked exemplar: melorheostosis.

The legal label and the operative group are different objects — and the gap is the expected condition.

The named category is “disability.” The cluster the agents act on is the disease-marker-correlated community, assembled from genomic and behavioral signals. Faithful carriage of the label tells you nothing about the community doing the work.



carried faithfully — and still not the group that acts

THE MISMATCH

Audit against the named class and you validate the wrong object. The mismatch is structural, not accidental.

Peel, Larremore & Clauset (2017). “The ground truth about metadata and community detection in networks.” Science Advances 3(5):e1602548.

Carry the provenance of the naming act.

Make authority a first-class protocol property — not just the category, but who declared it.

A consuming agent should verify the legitimacy of the asserter, not only the fidelity of the carriage. FAIR carries the name; it is silent on whose name it is.
Registered when the category is authored. Bound to the assertion it travels with. Verifiable by the consuming agent at the gate.

WHICH AUTHORITY

Who declared this category protected.

WHAT MANDATE

The source of their standing to do so.

CONTESTABLE HOW

The route by which it can be challenged.

REVISABLE WHEN

The conditions under which it changes.

THE PROPERTY

Four fields carried with every category assertion — which authority, what mandate, contestable how, revisable when.

FAIR: Wilkinson et al. (2016), Scientific Data 3:160018. CARE: Carroll et al. (2020), Data Science Journal 19:43.

Measure conformance to the authored authority graph.

The reference is the authored graph. The measurand — the quantity measured — is the operating graph.

The authored authority graph records who granted which scope, what may be delegated, and what constraints bind a spawned agent. The operating authority graph is what the running federation instantiates. Evaluation is a conformance check of the second against the first.

AUTHORITY AMPLIFICATION

An agent operates with more authority than was granted or delegated to it.

AUTHORITY ORPHANING

Authority exercised with no traceable grant — an edge with no source.

CONSTRAINT LOSS ON SPAWNING

A spawned agent does not inherit the constraints attached to its parent's grant.

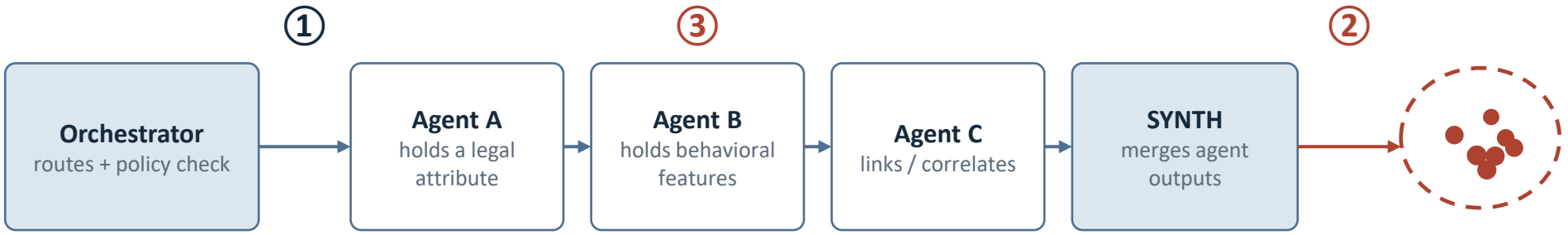
THE SUBSTITUTION

Classic attestation, one change: attest the authority under which an agent acts — not only the state of the component.

Where authority is not authored explicitly enough to serve as a reference, there is no measurand — evaluation depends on the authoring step having occurred.

Where the check runs: three gates in a live federation.

In-system, at the gate — not post-hoc audit.



① AT ASSIGNMENT

Exists today: containment is checked. Add: verify the grant itself — its provenance, not only its possession.

② AT COMPOSITION

New: gate the merge. Outputs carry their authority domains; the composed result carries the combined provenance.

③ AT RE-ADVERTISEMENT

New: re-attest on change. A version increment re-verifies the grant, closing delegation-by-forgetting.

THE PLACEMENT

Composition and change are gated, not only assignment. The check runs where audit cannot reach in time.

Gate ① exists in Federation of Agents (policy-flag containment at assignment): Giusti et al. (CERN; L. Serio), 2025. arXiv:2509.20175. Gates ② and ③ are the proposed additions.

Conclusion: the object, and how to evaluate it.

THE DIAGNOSIS

The object of evaluation is the grouping the system acts on. It cannot be discovered; it is authored. In a federation it assembles at runtime, held by no one — and the mismatch with the named class is structural.

THE SOLUTION

Carry provenance-of-authority as a first-class protocol property: authority, mandate, contestation route, revision conditions. Evaluate conformance of the operating authority graph to the authored one. Gate composition and re-advertisement, not only assignment.

THE ASK

A coordination-layer work item: provenance-of-authority as a specified data element, conformance semantics, and gate placement at assignment, composition, and re-advertisement.