

VIKRAM JHA - DEFENSIBLE AI AUDIT

THE REACHABILITY AUDIT - TWELVE QUESTIONS PER ENDPOINT

Version 1.0 - 23 August 2026

<https://vikramjha.work/artifacts/agent-reachability-audit>

Twelve questions that ask, endpoint by endpoint, whether your system grants on the basis of what reached it rather than who asked.

WHY IT EXISTS

Reachability is not identity. CVE-2026-9198 is the extreme published version - a login endpoint that, in the vulnerability record's own wording, mints superuser tokens to any network caller. But the shape is rarely that obvious. It usually looks like a service that was correct when only one thing could reach it, and is now reachable by an agent that did not exist when the boundary was drawn. Answer these from the code and the config, not from memory and not from the design document: the gap between the design and the deployment is the thing you are looking for.

WHAT IT CONTAINS

1. Twelve questions in four groups - identity, authority, evidence, revocation
2. A scoring convention where 'nobody could tell me today' counts as a failure
3. The escalation path most audits miss, because each endpoint is correct on its own
4. How to read a failing column as an architecture rather than an endpoint
5. What a failing set does and does not let you conclude

THE TWELVE QUESTIONS

Score each Y (yes, demonstrably), N (no), or ? (nobody could tell me today). A '?' counts as an N - it is not a smaller problem, it is the same problem plus an unknown.

IDENTITY

#: 1

Question: Does this endpoint require a credential at all, or does arrival suffice?

What a failure means: The base case. Include endpoints protected only by network position.

IDENTITY

#: 2

Question: If a credential is required, does it identify a unique principal - or a role, a team, or a shared service account?

What a failure means: A credential four things share does not identify a caller. It identifies a group, and every record afterward will say so.

IDENTITY

#: 3

Question: Could the calling agent's credential have been obtained by anything else that can reach the network segment?

What a failure means: Static secrets in environment variables, files or images all fail this. So does anything an agent can read and re-present.

IDENTITY

#: 4

Question: Does the endpoint distinguish an agent acting for itself from an agent acting for a user?

What a failure means: If both arrive as the same principal, the delegation is invisible at enforcement and no downstream record can reconstruct it.

AUTHORITY

#: 5

Question: Is the grant scoped to specific actions, or is it 'authenticated therefore permitted'?

What a failure means: Callers are known and unbounded.

AUTHORITY

#: 6

Question: Does the grant have a stated expiry enforced at this endpoint, not just a token TTL set upstream?

What a failure means: Expiry that is not enforced where the action happens is a description, not a control.

AUTHORITY

#: 7

Question: Can the caller widen its own grant by asking for more, re-authenticating, or invoking a second endpoint that issues a broader credential?

What a failure means: The escalation path most audits miss, because each endpoint is correct on its own.

AUTHORITY

#: 8

Question: Is the grant attenuating - can a caller pass a narrower authority downstream, and does the downstream honor the narrowing?

What a failure means: If authority passes whole or not at all, every hop runs at the widest scope on the chain.

EVIDENCE

#: 9

Question: Does a call here produce a record naming the principal, the action, and the authority relied on?

What a failure means: You may be well controlled and unable to demonstrate it.

EVIDENCE

#: 10

Question: Would that record survive the credential being rotated or the account being deleted?

What a failure means: A record that resolves an identity by lookup, rather than recording it, degrades to nothing.

EVIDENCE

#: 11

Question: Could you answer 'who authorized this?' about a single call from six months ago, using only what is retained today?

What a failure means: Retention sized for debugging rather than for examination.

REVOCATION

#: 12

Question: If you revoked this caller's authority right now, what is still in flight that would land anyway?

What a failure means: Go and measure it rather than estimating it. The time-to-safe-state protocol has the method.

READING THE COLUMNS

Q1-Q4

What it means: The estate has no concept of an agent principal. Everything downstream of that is cosmetic, and this is the only thing worth fixing first.

Q5-Q8

What it means: You have identity but not authority. Callers are known and unbounded.

Q9-Q11

What it means: You may be well controlled and unable to demonstrate it. Different problem, different fix, and it usually surfaces at the worst possible moment.

Q12

What it means: Nobody has measured revocation. Estimating it is how a containment plan becomes a remediation plan that never happens.

HOW TO USE IT

1. List every endpoint an agent in your estate can call. Include internal ones - especially internal ones, since that is where the assumption lives.
2. Answer the twelve questions from the code and the config, not from memory or the design document.
3. Score Y, N or ?, and treat every ? as an N.
4. Read the columns. One endpoint failing question 7 is a finding about that endpoint; question 7 failing everywhere is an architecture.
5. Treat a failing column as a question to answer, not an incident to declare. This tells you a class of grant is being made on the basis of reachability, which is a precondition rather than an exploit.

Built from public standards and general practice. No client information appears anywhere in this document. Regulatory instruments move - check anything cited here against its primary source on the day you rely on it.
Free to use and adapt; attribution welcome, not required.