ENGINEERING · TECHNICAL NOTE · Technical Note

Reproducibility is becoming an audit requirement — is your AI assessment defensible?

If two auditors assess the same AI system and reach two different conclusions, neither conclusion is evidence. It is opinion. And opinion is precisely what regulators, courts and procurement teams are learning not to accept.

This note explains why reproducibility — the property that the same system, assessed the same way, yields the same findings — is moving from a methodological nicety to a baseline requirement for AI audits, and what organizations should demand from any assessment they commission.

The problem with one-off assessments

Most AI audits today are artisanal. An expert reviews documentation, runs a selection of tests, applies judgment, and produces a report. The report may be excellent. But ask a simple question — if we ran this audit again next month, with a different reviewer, would we get the same result? — and the honest answer is usually no.

For internal comfort, that may be acceptable. For anything with legal or contractual weight, it is a structural weakness. Consider the situations in which an AI assessment is actually used. A regulator questions whether a deployed system meets its obligations, and the organization produces last year's audit as evidence of diligence. A procurement team compares two vendors' compliance reports. A court examines whether an organization took reasonable steps before an automated decision caused harm. A board tracks whether the compliance position improved after remediation.

In every one of these situations, the value of the assessment depends on it being verifiable — capable of being re-run, checked, and compared. An assessment that cannot be reproduced cannot be verified. It can only be believed.

Why regulators are converging on this

None of the frameworks that matter in this region uses the word “reproducibility” as a headline requirement. But all of them are built on assumptions that quietly demand it.

The EU AI Act requires providers of high-risk systems to maintain technical documentation sufficient for authorities to assess conformity — an assessment logic that presupposes a repeatable basis of evaluation, not a one-time expert impression. Its post-market monitoring duties assume compliance can be measured at multiple points in time and meaningfully compared, which is only possible if the measurement itself is stable. SDAIA's AI Ethics Principles place accountability and auditability at the center of responsible deployment: an audit trail is only as strong as the audit it records. And PDPL's

accountability provisions require controllers to be able to demonstrate compliance — demonstration being, by definition, something a third party can check.

The direction of travel is consistent. As enforcement matures, “our expert looked at it” will carry less and less weight, and “here is the assessment, here is the method, run it again if you doubt it” will carry more. This mirrors what happened in financial audit decades ago: methodology became standardized precisely because regulators and courts refused to rely on unverifiable professional opinion.

What reproducibility changes in practice

For the organization commissioning an audit, reproducibility has three practical consequences.

First, findings become comparable over time. If this quarter's assessment and last quarter's were produced the same way, the difference between them is real change in your compliance position — not noise introduced by a different reviewer or a different mood. Remediation tracking becomes measurement instead of narrative.

Second, findings become defensible under challenge. When a regulator, counterparty or claimant disputes an assessment, the strongest response is not a longer report; it is the ability to show that the assessment follows a defined method that produces the same result on the same inputs, whoever runs it. The conversation shifts from “trust our expert” to “check our work” — a far stronger position.

Third, findings become portable. A reproducible assessment of one system can be extended across a portfolio on a consistent basis, which is what boards and group-level compliance functions actually need. Ten artisanal audits of ten systems produce ten incomparable documents; a consistent method produces a compliance position.

What to ask of any AI audit you commission

Organizations do not need to understand audit methodology in depth to protect themselves. Four questions are enough.

- **Would the same inputs produce the same findings?** If the answer is no, or unclear, the assessment is an opinion — treat it as one.
- **Is the method defined independently of the person applying it?** An audit that lives in one expert's head cannot be verified, transferred, or defended after that expert moves on.
- **Can findings be traced to evidence?** Every conclusion in the report should point to something checkable — a document, a test result, an obligation in the applicable text — not to unexplained judgment.
- **Will next year's assessment be comparable to this one?** If the method changes with every engagement, you will never know whether your position improved.

An audit that passes these four tests will hold up in front of a regulator. An audit that fails them may still be insightful — but it will not be evidence when evidence is what you need.

The bottom line

AI and data compliance is leaving its artisanal phase. The frameworks now in force in Saudi Arabia and the European Union all assume, implicitly or explicitly, that compliance can be demonstrated, re-checked and tracked — and only reproducible assessment satisfies that assumption. Organizations choosing an audit partner should treat reproducibility not as a technical detail but as the property that determines whether the resulting report is a legal asset or an expensive PDF.

Oxon X conducts structured, reproducible AI & data compliance audits against SDAIA, PDPL, NCA and EU AI Act requirements.

Schedule a compliance audit → contact@oxonx.sa