

Synthetic Data & the Sovereignty Problem

When models learn from the output of other models, the question of who controls the well becomes a national one.

By **Abhijeet Vyas**, Integral Consulting ●

The open internet used to be a record of what humans wrote. Increasingly, it is a record of what machines wrote — text, images, and code generated by models, published at a scale no human population could match. This matters for a reason that has nothing to do with taste: the internet is also the training ground for the next generation of models.

When a model trains on the output of earlier models, errors stop being random and start being **inherited**. The rare, the unusual, and the true-but-inconvenient get smoothed away; the plausible average gets amplified. Researchers call the degenerate end-state model collapse. Long before collapse, though, comes something subtler and more commercially relevant: a slow, invisible narrowing of what systems know, with no single failure to point to.

Provenance stops being hygiene and becomes strategy

For an organisation, the practical question is brutally simple: can you account for what your

systems learned from? A model fine-tuned on data of unknown lineage — or built on a base model whose training corpus no one can describe — is an asset you cannot fully defend, in front of a regulator, a court, or your own board.

The organisations that can answer “where did this come from?” will move faster than the ones that cannot — because they can stand behind what they ship.

This is why provenance has moved from an archivist's concern to the foundation of AI governance. Data lineage, model lineage, and the discipline of knowing what entered the pipeline are no longer optional documentation. They are the difference between a system you operate and a system you merely host.

Scale the problem up, and it becomes sovereignty

Now raise the same question from the organisation to the state. A country whose administration, finance, health systems, and public discourse increasingly run on AI is a country whose critical infrastructure depends on three things: the data those models trained on, the models themselves, and the compute they run on. Whoever controls those three controls something that looks less like a product category and more like a utility.

If the data well is contaminated with synthetic output — and if the models are trained, tuned, and hosted elsewhere — then sovereignty erodes quietly, without a single decision being taken. No treaty is signed. No border is crossed. The dependency simply accumulates, one procurement at a time.

THE CORE IDEA

Sovereignty in the AI era is not about building everything at home. It is about being able to account for — and if necessary replace — what you depend on.

What this means in practice

For nations, it argues for treating high-quality, verifiable, human-generated data as a strategic reserve; for investing in curation and provenance infrastructure, not just compute; and for procurement standards that ask of every AI system the question no vendor volunteers: what was this trained on, and who can attest to it?

For organisations, the agenda is nearer at hand. Inventory the models you use and the data they touched. Prefer suppliers who can document lineage over those who merely promise performance. Treat your own first-party data — clean, consented, human-generated — as the appreciating asset it has quietly become. And write provenance requirements into contracts now, while you still have the leverage of the buyer.

The uncomfortable truth is that the window for this discipline is the present one. Provenance cannot be retrofitted onto a system after the fact any more than lineage can be restored to a river downstream of the spill. The organisations — and the countries — that treat the well as infrastructure will own their systems. The rest will rent theirs, on terms they did not set.

If your systems depend on data and models you cannot account for, that is the conversation to have.

Integral advises organisations on data provenance and AI governance aligned to ISO 42001 and the NIST AI Risk Management Framework. contact@integralconsulting.in