

AI Governance Letter · Issue #3

Agreement Is Not Evidence

Why a system's confidence measures repetition rather than corroboration — and the two questions that replace “is our data good?”

September 2026 · Reto Gruenenfelder, IT AI Advisory

Impressum

IT AI Advisory Sdn. Bhd. (202601023072)
B-3A-8, Icon City, Jalan SS8/39, 47300 Petaling Jaya, Selangor, Malaysia
Responsible for content: Reto Gruenenfelder
reto@itaiadvisory.com · itaiadvisory.com

Published for informational purposes only. Not legal, regulatory, or compliance advice. Full disclaimer:
itaiadvisory.com/disclaimer

Dear Board Director,

A few weeks ago I was checking a figure for the previous letter. It concerned AI-related breaches, and it appeared everywhere — the same number, the same framing, across blog posts, vendor briefings and summaries of summaries. Most of them cited IBM as the source, and most pointed to IBM's 2026 report as the reference.

Then I went to the 2026 report itself. The number was not there. IBM publishes this report every year, and the figure it carries changes each time. The number circulating everywhere was from the 2025 edition. What had happened was straightforward: writers had referenced IBM's latest publication without checking whether it still contained the number they were citing. It did not. They had attributed a 2025 finding to a 2026 document that contradicted it.

It took a few minutes to untangle, and only because I went looking for the origin instead of counting the agreements. A dozen apparently separate sources had agreed with one another, which looked like confirmation. It was not. They had all inherited the same error from the same place, and each repetition had made the mistake look more settled than it was. That distinction — between sources that agree and sources that independently corroborate — is the subject of this letter, and it sits underneath every AI system your organisation buys or builds.

Five Sources, or One Source Five Times

Tell a person the same story five times and they stop listening. They say: yes, I know that one, tell me something new. To a human being, repetition carries almost no additional weight, because we recognise that we are hearing the same thing again.

A training algorithm has no such reflex. It doesn't recognise five closely correlated documents as one story retold. They contribute as five, and the pattern they share is reinforced accordingly. What a person discounts as redundancy, the machine records as confirmation.

The consequence is uncomfortable. A system can be most confident precisely where its source material is most repetitive — and repetition is what published text produces best. One well-written claim propagates into hundreds of near-identical restatements, and every restatement makes the underlying pattern appear better supported than it is.

This is not a defect that better engineering removes. It is a property of learning from a body of text in which some things are said very often and other things are said once.

A model counts its sources. It does not trace them.

The Half That Cannot Be Checked

There is a second failure, and it is the more serious of the two, because it leaves nothing behind to inspect.

Consider the discipline captured in the term MECE — mutually exclusive, collectively exhaustive. Applied to the material a system learns from, the first half asks whether the sources overlap: are these genuinely separate pieces of evidence, or the same evidence

wearing different clothes? That is the problem described above, and in principle it can be examined, because the sources exist and can be traced.

The second half asks whether the set is complete — whether everything that ought to be represented is represented. No machine can establish this, because absence leaves no trace. No signal exists for a subject nobody wrote down. No confidence score falls because a market, a language, a product line or a decade is missing. Seen from inside the material, a gap and a certainty are indistinguishable.

Only a person who knows the field can say what should have been there. That is not a technical judgement. It is an expert one, and it requires knowing the domain well enough to notice what is not being said.

For an institution operating across ASEAN and Europe this is not an abstraction. It is worth asking how much of the material behind a given system reflects the markets you actually operate in — the customers, the products, the languages, the local conventions — and who in your organisation has looked. A system built largely on material from elsewhere will not announce that it is out of its depth. It will answer in the same even tone it uses for everything else.

Nor can the gap be closed by adding more. Volume is not coverage.

Suppose a mortgage model is judged to rest on too little data, and the remedy is to feed it a great deal more — including mortgage rates drawn from Latin American markets into a European book. The volume rises. The apparent coverage rises. But those rates carry a different inflation regime, a different currency, different enforcement of collateral and different refinancing behaviour. Nothing in them describes the market being lent into. They do not sit inertly beside the relevant material; they shift what the system treats as normal.

That is the part worth holding onto. Irrelevant material is rarely neutral. A system will find a pattern in whatever is placed in front of it, whether or not the pattern means anything, and data that does not belong does not so much dilute the answer as bend it. It bends it quietly, because the output looks the same either way.

Which is the same discipline again, seen from the other side. Mutual exclusivity asks what has been duplicated. Collective exhaustiveness asks what is absent. Neither asks how much there is, and neither is improved by adding.

Why “Is Our Data Good?” Cannot Be Answered

Boards do ask about data. The question is almost always some version of: is our data good? It is a reasonable instinct and an unanswerable question, because “good” carries no test. Management can answer yes in complete honesty, and the board is no wiser than before it asked.

MECE replaces one unanswerable question with two answerable ones.

Are these sources independent of one another?

Not how many were used, but how many are separate. A billion documents descended from a hundred thousand originals are a hundred thousand documents and a great deal of amplification.

Who determined that the set was complete, and on what basis?

Not whether the data is large, but who decided what belonged in it — and whether that person knew the domain well enough to notice what was missing.

Neither can be satisfied by a policy document. Someone who did the work can answer both, and someone who did not can answer neither.

Questions Worth Raising at Board Level

As in previous letters, these are framed so that a statement of policy will not satisfy them.

1. When management reports what the analysis concluded, do we know whether that rests on independent sources or on one source repeated?

Volume of agreement and weight of evidence are different measures. Only one of them is a governance standard.

2. Who decided what belonged in the data behind this system, and what qualified them to decide?

Completeness is a human judgement. If nobody is named, nobody made it.

3. When a supplier says the system is “trained on our data”, does that mean trained, fine-tuned, or retrieved at the moment of the query?

The three differ materially in where accountability sits. The phrase used in the sales conversation is frequently not the one that describes the system.

4. How much of the material behind this system reflects the markets we actually operate in?

If nobody knows, that is itself the answer, and it is worth recording.

5. What in this system would tell us that something is missing?

If the answer is nothing, then absence is invisible, and every gap will present itself as an answer.

6. What would we accept as evidence that an output is corroborated rather than merely agreed?

A board that cannot answer this has no test to apply and will accept confidence as a substitute.

A Final Observation

Boards already hold this principle. It is the foundation of the second line of defence: an assurance function that merely repeats the first line's conclusion has told you nothing, and every director understands why. Independence of evidence is not a technical concept. It is the ordinary discipline of governance.

The systems now being placed inside your institution do not apply that discipline, because they were not built to. They weigh what is frequent, and they cannot see what is absent. Neither is a reason to reject them. Both are reasons to keep judgement about completeness in human hands—and to know whose hands those are.

Your board already knows that five people repeating one rumour is not five sources. The question is whether the systems you are buying know it, and if they do not, who in your organisation does.

About the Author

Reto Gruenenfelder advises executive and non-executive boards on AI governance through IT AI Advisory Sdn. Bhd. in Kuala Lumpur. He served on the Technology and Cybersecurity Committee of Bursa Malaysia from 2019 to 2022, and spent nearly twenty years at IBM in financial services across Switzerland, Europe and Asia Pacific. His AI Governance Guide, fifteen board-level frameworks for AI oversight, is available at itaiadvisory.com/guide.

Primary Source Documents

The following link directly to the original material referenced above.

→ **[IBM: One in Four Malicious Breaches are AI-Enabled — 29 July 2026](#)**

→ **[IBM: Cost of a Data Breach Report 2026 — full report](#)**

Published for informational purposes only. Not legal, regulatory, or compliance advice.

Full disclaimer: itaiadvisory.com/disclaimer

© 2026 IT AI Advisory Sdn. Bhd. (202601023072). This letter may be shared freely in its original form.