

AI Governance Letter · Issue #4

The Trap Question

Why the one question that stops a problem from growing cannot be asked of a data estate

October 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

Dear Board Director,

A bank had been out of service for forty-eight hours. I interviewed everyone who had touched the environment, every account differed, and that was the instrument rather than the obstacle — where the accounts diverged was where the fault was. So I asked one question, built so that only someone concealing the answer would respond a particular way. The cause was known within the hour. The fix took fifteen minutes.

Suppose I had not found it. The bank would not have stayed down. Someone would have put something in to get it running again — a workaround, a nightly correction, a second copy of the data because the first could not be trusted. That is a patch, and the soil for a patch always comes from somewhere else. You dig a second hole to fill the first. The next one is bigger, it is ordered by someone else months later, and nobody is looking at the sequence.

That is a method built on questions and answers. It needs four things: someone who knows, accounts that disagree, behaviour that can be tested, and a fault that can be repaired once it is found. A data estate offers none of them. The established methods of root cause analysis all rest on asking somebody, and that is what fails here.

Nobody Can Say Where the Data Came From

An investigation begins by naming who knows. For the system your board most recently approved, ask yourself who you would call. In most institutions the list runs out quickly.

You are not unusual in that, and the point of the following figures is not to prove you have the problem. Kiteworks, a data governance vendor, surveyed 225 security, IT and risk leaders across ten industries and eight regions for its 2026 Data Security and Compliance Risk Forecast Report. It found that 78% of organisations cannot validate training data before use, and 77% cannot trace where that data originated. I state the sample and the sponsor deliberately: this is a modest vendor survey, not a census, and it should carry exactly the weight that description implies. The direction is not seriously disputed. What it tells you is that nobody has quietly solved this while you were not looking.

If the witness list is empty, the investigation does not begin. The patch gets ordered anyway.

Everything Agrees and the Answer Is Still Wrong

An investigation runs on inconsistency. Data does not oblige.

Take a claim that has circulated for most of this year: that more than half of everything on the web is now written by machines. It appears in vendor briefings, in board packs and in the trade press, and it traces to real work. Graphite sampled 55,400 English-language URLs from Common Crawl, each at least a hundred words long, published between January 2020 and March 2026 and classified as articles or listicles, and ran each one through three commercial detectors. The finding was that 49.9% of sampled articles in the first quarter of 2026 were classified as primarily AI-generated.

Forty-nine point nine per cent is not more than half. The sample is English-language articles and listicles, not the web. An article counted as primarily AI-generated when the majority of its text was flagged — and for one of the three detectors an article merely judged ‘mixed’

was counted as AI outright, so a piece written jointly by a person and a machine lands wholly in the machine column. Graphite's own headline was careful: that AI now writes as many articles as humans do, which is exactly what the number supports. Ahrefs, examining 900,000 newly created pages, reported that 74.2% contained some AI content while only 2.5% were purely AI. The same underlying reality yields 2.5%, 49.9% or 74.2% depending entirely on where the line is drawn.

The origin of the claim is closer to home than it looks. Graphite's own earlier study — one detector, data to May 2025 — was headed More Articles Are Now Created by AI Than Humans, and put the crossover at November 2024. They have since superseded it with the three-detector work above, which runs about three points lower and carries the more careful headline. The board pack is not misreading the research. It is reading, accurately, a version the researchers themselves have withdrawn.

There is no error to find in the citation. The sources are real, cited correctly and attributed correctly. The drift happened in the definition — in what was being counted — and definitional drift leaves no contradiction behind.

The same thing is happening inside your own institution, where it is harder to see. Your retail system and your wealth system both hold customers. Both are correct. Both are current. They do not mean the same thing by the word, and nothing in either will ever say so.

Checking the citation does not catch it, because the drift is in the definition, not the attribution.

In the bank, the weak point announced itself through friction between accounts. A data estate produces no friction. It produces agreement, and agreement is precisely what the method was built to see through.

There Is Nothing to Probe

The question itself is the part I would least willingly give up, and it is the part most completely unavailable.

A trap question works because a person concealing something behaves differently under a designed probe than a person who is not. What is being tested is not the fact. It is the behaviour around the fact. Data has no behaviour. A mis-provenanced record does not hesitate. A poisoned document does not over-explain.

Souly and colleagues established how far that goes in October 2025, working across the UK AI Security Institute, Anthropic, the Alan Turing Institute, and Oxford. They pretrained models from 600 million to 13 billion parameters on datasets from 6 billion to 260 billion tokens and found that roughly 250 poisoned documents compromised every model in the range, about equally. The largest models had been trained on more than twenty times as much clean data and were no safer for it. Two hundred and fifty documents is about 0.00016% of the largest corpus.

Boards should sit with that figure, because it contradicts an intuition almost everyone holds: that a sufficiently large body of data dilutes whatever is wrong inside it. It does not. And those documents do not announce themselves. They do not conflict with anything. They do not respond to probing, because there is nothing there to probe.

The previous letter in this series argued that volume is not coverage. The lesson arrives here from a less comfortable direction: volume is not safety either.

Finding the Fault Does Not Mean You Can Fix It

The method also rested on an asymmetry. Diagnosis was expensive — forty-eight hours — and the fix, once the diagnosis was right, was cheap. Fifteen minutes. That asymmetry is what made the investigation worth the effort.

With a trained model the asymmetry inverts. Suppose you do establish that particular material should never have entered the training data. Personal information inside a model is not a record you can locate and delete; it is distributed across a very large number of parameters, and there is as yet no settled technical definition of what would count as having successfully removed it. In the same Kiteworks survey, more than half of organisations reported having no mechanism at all to remove a data subject's data from a trained model.

So even a correct diagnosis may leave you with no proportionate remedy — only retraining, which for most institutions is not a remedy but a rebuild. This matters directly for any commitment your organisation has made about deletion, in a privacy notice or in a contract, that does not stop at the boundary of your databases.

The Digging Does Not Stop

Put the four absences together — no one to ask, nothing that disagrees, nothing to probe, and no repair once the fault is found. The conclusion is not that these systems are harder to investigate. It is that nothing ever settles what went wrong, so the patching never has a reason to stop.

A system that cannot be diagnosed is not a system that fails once. It is a system that gets patched. A reconciliation here, a filter there, a second copy kept because the first could not be trusted, a rule added to catch the case that surfaced last quarter. Each one is ordered by a different person. Each one is reasonable on the evidence available to the person ordering it. And each one is more soil taken from somewhere else. A patch is also not neutral: the reconciliation stops the mismatch showing, the second copy stops the bad figure surfacing. In a system with no witness and nothing that contradicts itself, those were the only signs there were.

Every patch works. Each one removes a symptom, and the symptom was the evidence.

The Objections Worth Answering

Two responses deserve to be met rather than avoided.

The first is that data lineage and observability tooling exists, and is improving. It is worth having. But lineage records what was captured at the moment of ingestion, which is precisely what most organisations did not capture, and it cannot be reconstructed afterwards for data already absorbed. Observability catches what is statistically anomalous, which is a real capability and not the one needed here — the poisoning work shows that the damaging

case need not be anomalous at all. 250 ordinary-looking documents are not an outlier in 260 billion tokens. They are indistinguishable from the background, which is the entire problem.

The second response is to change the supplier. It is the most available answer and the least useful one. It does not survive contact with a supervisor, whose next questions are who selected that supplier, who oversaw them, and what this board was told at approval — which returns the problem to the room it was trying to leave. And it changes nothing on its own, because the estate stays where it is. A new provider inherits the same sources, the same definitions and the same absent records, and arrives at the same place more expensively.

Questions Worth Raising at Board Level

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

1. If we had to establish today where the data behind this system came from, who would we ask?

The first step of any investigation is naming the witnesses. If no one can be named, the investigation cannot begin, and that is worth knowing before an incident rather than during one.

2. How many times has this system already been patched, and who is keeping count?

Each individual fix is recorded somewhere. The sequence usually is not. If nobody can say how many workarounds this system is now carrying, nobody can say how large the hole has become.

3. When a figure reaches this board, do we know what it counted — not merely who published it?

Accurate attribution is not the same as accurate meaning. A correctly cited number can still support a claim its own source does not make.

4. If material that should never have entered this system were found in it, what would we actually do?

The answer determines whether diagnosis is worth anything. If the only remedy is retraining, then the board is carrying a risk it cannot cheaply discharge.

5. What commitments have we made about deleting data that we could not honour if that data has reached a model?

Privacy notices and client contracts were mostly written for databases. Check whether ours promise something the architecture cannot deliver.

6. What would tell us that this system had been compromised through its data rather than through its code?

Security investment assumes an intruder who behaves. Data does not behave, and if the honest answer here is nothing, that gap is not covered by anything else we are doing.

A Final Observation

I do not want to overstate this. The investigative method I have described still works, and it still works well, wherever people know something and can be asked. Most of what goes wrong in an institution is still that kind of thing.

But the systems now placed at the centre of your operations have a failure mode that sits entirely outside it. They can be wrong for reasons that leave no inconsistency, no witness and no behaviour to test — and they will report that wrongness with the same composure as everything else. The forty-eight hours I spent in that bank were unpleasant, but they ended, because someone in the building knew and could be caught.

That is the part which no longer holds. And a fault that cannot be found does not wait quietly. It gets patched, by one person after another, each of them acting reasonably, none of them holding the whole sequence.

You cannot fill a hole by digging another one.

I have no method to offer in place of the one that fails here, and I would be wary of anyone who claimed to. What I can tell you is where the cost falls. Not on whoever ordered the first patch, but on whoever is in the chair when the ground finally gives way.

About the Author

Reto Gruenenfelder advises executive and non-executive boards on AI risk and governance. Since 2017 he has worked as an independent advisor, unaffiliated with any technology vendor, and served as Technology Advisor to the Board of Bursa Malaysia. Before that he spent nearly twenty years at IBM in financial services across Switzerland, Europe and Asia Pacific.

His view is usually sought before a board approves an AI programme, or when one is in trouble. In both cases the report belongs to the board only.

Primary Source Documents

The following link directly to the original material referenced above.

→ [Souly, Rando, Carlini, Gal, Kirk et al.: Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples — arXiv, 8 October 2025](#)

→ [Graphite: AI Now Writes as Many Online Articles as Humans Do — Common Crawl sample, 55,400 URLs](#)

→ [Graphite, superseded: More Articles Are Now Created by AI Than Humans — one detector, data to May 2025](#)

→ [Ahrefs: 74% of New Webpages Include AI Content — study of 900,000 pages](#)

→ [Kiteworks: 2026 Data Security and Compliance Risk Forecast Report — survey of 225 leaders](#)

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.