



A COGNITIVE DELIVERY PAPER

Reporting AI to the board

A manager's guide to opportunity, adoption, and the evidence
behind both

Jason Cruz Falzon

Creator of the Cognitive Delivery Framework and the Cognitive Delivery Index

AUGUST 2026 · VERSION 1.0

Two questions, one pack, one standard of proof

The board is not asking for good news. It is asking for a position it can defend.

Two things arrived at once. Provision 29 of the UK Corporate Governance Code took effect for financial years beginning on or after 1 January 2026. Boards must now declare whether material controls were effective at the balance sheet date, and name the ones that were not. Organisations with December year-ends are inside that first cycle now and will report on it in early 2027.

Meanwhile AI moved into operations faster than the controls around it matured. McKinsey's 2026 AI Trust Maturity Survey of roughly 500 organisations found active mitigation lagging risk awareness across nearly every category. Organisations know what could go wrong. Fewer can show what they did about it.

So the manager who owns AI delivery now answers two questions from the same board, in the same pack, on the same afternoon. Is this working, and can we stand behind it. Most organisations answer them with different numbers drawn from different places and held to different standards. The investment case gets vendor benchmarks and self-reported productivity. The assurance section gets policy names and a certification logo. Neither survives a director asking how you know.

The order matters, and most guidance on this gets it backwards. A board's agenda item is not AI governance. It is AI: what is running, what it returned, what is coming next, and what was stopped. Governance is not the subject of that item. It is the reason the numbers inside it can be believed.

THE ARGUMENT

The report that wins the investment and the report that survives the scrutiny are the same report. Not because governance is good for you, but because the instrumentation that proves a control fired is the instrumentation that proves a use case moved.

What follows is one test, applied to the numbers that usually appear in board packs, with a standing structure and a ninety-day plan. It is written for the manager assembling the pack.

CONTENTS

01 Two requests, one person

02 The one test

03 One report, not two

04 Six numbers that do not survive

05 What to report instead

06 The standing pack

07 Regulated and listed callouts

08 The groundwork

09 The first ninety days

10 Communicating it upward

A The evidence ladder, one page

B The board pack, drawn up

C A worked example pack

D Sources and how to weigh them

Two requests, one person

Provision 29 is a short piece of drafting with a long tail. The board must monitor the risk management and internal control framework, review its effectiveness at least annually, and declare whether material controls were effective at the balance sheet date. Where they were not, the board explains which ones and what is being done. It covers financial, operational, compliance and reporting controls, and it runs on comply or explain.

There is no special AI rule inside it, and that is the point people miss. If an AI system is material to how the organisation operates, the controls over it are material controls by the ordinary meaning of the words. Nobody has to designate them. They arrive in scope by default, and the board's declaration covers them whether or not anyone mapped them first.

The Code does not define materiality with a universal checklist, which leaves every organisation working out its own line. That interpretive gap is uncomfortable, but it cuts both ways: it also means a manager who maps their AI systems to principal risks early is defining the terms of their own assessment rather than receiving them.

The other half of the ask

At the same time the same manager is asked to justify continued investment. Here the numbers are unhelpful in a different way. McKinsey's November 2025 survey found 88 percent of respondents reporting regular AI use in at least one business function, with roughly a third saying they had begun to scale. Only 39 percent attributed any EBIT impact at enterprise level, and most of those put it below 5 percent. About 6 percent attributed more than 5 percent. That survey is now nine months old and self-reported, so treat the exact figures as indicative rather than settled.

The shape is what matters. Near-universal adoption sits alongside a minority reporting measurable enterprise impact. A board pack has to explain that gap rather than write around it, because every director has read the same headline.

THE POSITION THIS PUTS YOU IN

You are being asked to prove value in a field where most organisations cannot, and to attest to control effectiveness in a domain where most organisations have documentation rather than evidence. Answering either one badly damages the other. A value claim nobody can substantiate makes the assurance section look decorative. An assurance section built on policy names makes the value claim look optimistic.

The way out is not a better narrative. It is a single standard of proof applied to both halves, stated openly, including where you fall short of it.

The one test, and the evidence ladder

Every line in a board pack should answer one question before it earns its place. How do you know?

That question sorts claims into three tiers. The tiers are not a maturity model and there is no prize for reaching the top of all of them. They are a way of labelling what a number can actually carry, so that a decision is not resting on something lighter than it needs.

TIER 1

Assertion

Someone says it. A status colour, a confidence rating, a summary sentence, a percentage with no stated method. Costs nothing to produce and nothing to change. Cannot be checked without asking the same person again.

Legitimate use: judgement, forward view, professional opinion. Label it as such.

TIER 2

Artefact

A document exists. A policy, a register, an approved standard, an impact assessment, a certificate. Can be produced on request. Demonstrates intent and design.

What it does not demonstrate: that the thing described happened, in production, on the day in question.

TIER 3

Attested

A record generated by the system at the moment the thing happened, held somewhere the reporter does not control, capable of showing failures as well as successes, and checkable by someone who does not trust you.

This is the only tier that supports a declaration of effectiveness, because effectiveness is a claim about operation, not about design.

Most board reporting on AI sits at tiers one and two and is presented in language that implies tier three. That is not usually dishonesty. It is what happens when the reporting format was designed for a status update and the obligation quietly became an attestation.

RULE

A control you can describe is an artefact. A control you can evidence firing, and failing, is attested. Provision 29 asks for the second and will accept the first only until somebody checks.

Applying the test in practice

The test is fast. Read any line in the pack and ask what a director would have to do to verify it. If the answer is "ask you", it is tier one. If the answer is "request a document", it is tier two. If the answer is "query a system that you do not administer", it is tier three.

HOLDS UNDER QUESTIONING

- Exception counts pulled from the pipeline, with the query stated
- A baseline dated before the change, measured the same way after
- Named incidents with detection and response times
- Coverage figures with the denominator shown
- An explicit list of what is not yet instrumented

FALLS APART UNDER QUESTIONING

- A green status with no defined threshold
- Percentage improvements with no stated baseline date
- "All teams follow the policy"
- A certification logo used as a control
- A section with no exceptions listed at all

The panels are not a scoring scheme. They are the two ways the same fact tends to be written, and the difference is almost always whether the writer had access to a system of record when they wrote it.

One line, at three tiers

Take a sentence that appears in almost every AI board pack and watch what changes as the evidence behind it rises.

TIER	HOW THE LINE READS	WHAT A DIRECTOR CAN DO WITH IT
Assertion	AI-generated code is reviewed before it reaches production.	Believe you, or not. There is no third option.
Artefact	Our engineering standard requires human review of AI-generated changes before merge. The standard is approved, dated and available.	Confirm the rule exists. Still cannot tell whether anyone followed it last Thursday.
Attested	Of 1,240 changes merged this quarter, 1,190 carried a recorded review. Fifty were merged under the documented break-glass path, each with a named approver. Drawn from the merge log.	Test the claim, see the exceptions, and ask about the fifty. Which is the conversation worth having.

Illustrative figures.

The third version is not longer because it is padded. It is longer because it contains the exceptions, and the exceptions are what make it evidence rather than a statement of intent. A board that never sees a denominator has no way of knowing whether it is being shown a performance or a measurement.

Why opportunity and assurance are the same report

The instinct is to treat governance as a cost imposed on the value case. The available evidence points the other way, though it needs handling carefully.

Protiviti and BoardProspects surveyed 772 board members and C-suite executives in the fourth quarter of 2025, publishing in March 2026. Only 26 percent of boards discussed AI at every board meeting. Split by outcome, the pattern is stark: among organisations reporting strong returns from AI, 63 percent put AI on every board agenda. Among those reporting low returns, 13 percent did. The same split shows in confidence, with 95 percent of high-return organisations confident in their AI integration against 33 percent of low-return ones.

McKinsey's 2026 AI Trust Maturity Survey found organisations with explicit ownership for responsible AI averaging 2.6 on its maturity scale against 1.8 for those without a clearly accountable function. It also found organisations investing heavily in responsible AI practices were markedly more likely to report EBIT impact above 5 percent.

READ THIS BEFORE YOU PUT THOSE NUMBERS IN A PACK

All of the above is correlation. In both surveys the outcome and the practice are reported by the same respondent, which invites a halo: a person who feels good about their AI programme rates every item high. The causal direction is unresolved, and boards plausibly discuss AI more often because it is going well rather than the reverse. Both firms sell governance advisory services, so the finding sits with their commercial interest. Say all of this in the pack yourself. A director who spots an unstated weakness in your strongest slide will discount the rest.

The mechanism, which is less mysterious than the correlation

Set the surveys aside and the connection is structural. To show a board that a control fired, you need a record of the moment it fired. To show a board that a use case moved a number, you need a record of the same work before and after. In practice these come from the same places: the pipeline, the ticketing system, the deployment log, the incident record.

An organisation that cannot evidence its controls usually cannot evidence its benefits either, and for the same underlying reason. Nobody instrumented the work. The governance gap and the value gap are two readings of one missing capability.

This is why splitting the pack into a value section written by delivery and a risk section written by compliance produces two documents that do not reconcile. They are drawing on different evidence to describe the same activity, and a good director will notice.

A USEFUL REFRAME FOR THE CONVERSATION UPWARD

Instrumentation is not a compliance cost charged against the AI programme. It is the measurement capability the AI programme needed anyway to know whether it was working. The regulatory obligation simply removed the option of not building it.

Six numbers that do not survive the question

These appear in most AI board packs. Each is defensible as context and indefensible as evidence, and the difference matters more now that a declaration hangs off the pack.

1. Licence counts presented as adoption

Seats purchased is procurement data. Seats activated is provisioning data. Neither is use, and use is not dependence. The question a board is actually asking is what would break, and how badly, if this stopped on Monday. Report that instead.

2. Self-reported productivity

The 2025 DORA study of nearly 5,000 technology professionals found over 80 percent believed AI had increased their productivity. The same study found AI adoption associated with increased software delivery instability. Both findings are real and they are not in conflict, because one measures belief and the other measures outcome. Perception is worth tracking. It is not a result, and presenting it as one invites the obvious question about who was asked.

That report is from September 2025 and predates the current generation of agentic tooling, so treat its magnitudes as dated even where the direction holds.

3. Attributed savings with no baseline

A saving requires a before. The before has to be dated earlier than the intervention and measured the same way as the after. Hours-saved figures derived by multiplying a headcount by an assumed percentage are an assumption wearing a number's clothing. If there is no baseline, the honest line is that there is no baseline and here is what is being established from this quarter.

4. One-sided delivery metrics

DORA's finding that AI adoption now correlates with higher throughput while remaining associated with higher instability is the cleanest illustration available. Throughput alone argues for expansion. Instability alone argues for restraint. Only the pair supports a decision, and a board given one half will make a worse choice than a board given neither.

5. A certification standing in for a control

ISO/IEC 42001 evidences that an AI management system exists, was audited against a recognised standard, and is subject to surveillance. That is genuinely worth having and worth reporting. It does not evidence that any particular control changed any particular behaviour at any particular moment, and a board pack should not let it carry that weight.

6. A borrowed statistic nobody checked

This one deserves a worked example, because it is the failure mode that damages a pack fastest.

WORKED EXAMPLE: THE 95 PERCENT

The claim, now near-ubiquitous in slide decks: 95 percent of enterprise generative AI pilots fail to deliver measurable impact. Source: The GenAI Divide, Project NANDA, MIT Media Lab, July 2025.

What a sceptical director finds in ten minutes. The cover describes the work as preliminary findings. The sample is self-selected, and press accounts of its size do not agree with each other or with the document. The 5 percent sits in a funnel chart measuring task-specific tools reaching production, which is a much narrower thing than all AI pilots. The project that published it builds agentic AI infrastructure, and the report concludes that agentic AI infrastructure is the way across the divide.

And the finding almost nobody quotes: within the same sample, externally partnered deployments reached production roughly twice as often as internal builds. The report lists "the best enterprises build their own" among the myths it is correcting. The lesson usually drawn from the statistic is close to the opposite of what the document says.

One caveat on this analysis, which proves the point. The critique above rests substantially on secondary commentary rather than a full audit of the primary. If a number cannot be checked easily, it should not be load-bearing, and that applies to the critique as much as to the claim.

The damage is not that the statistic is wrong. It is that a director who checks one number and finds it soft will apply that discount to every other number in the pack, including the ones you got right.

Three checks before a borrowed number goes in

1. **Find the primary, not the article about it.** If the primary sits behind a form, or the sample size differs between the document and its coverage, that discrepancy is itself the finding.
2. **Read what the underlying exhibit measures.** Headline claims and the charts beneath them frequently describe different populations. The gap between the two is where a number quietly stops meaning what it appears to mean.
3. **Ask who benefits from the conclusion.** Not to dismiss the source. Commercial interest does not make a finding false, but it does determine how much weight the finding can carry on its own.

A number that fails any of the three can still appear as context. It cannot appear as support for a decision, and in a well-built pack those two roles look visibly different on the page.

What to report instead

Six working rules. They are unglamorous and they survive contact with an audit committee.

RULE	WHY IT HOLDS
Report pairs, never singles	Speed with stability. Coverage with exceptions. Volume with rework. A single number almost always supports the opposite decision to the pair it belongs to.
Date the baseline before the intervention	A baseline established after the change is a description, not a comparison. Where none exists, say so and start one, on the record, this quarter.
Count from systems of record	Surveys measure belief and belief moves with mood, recent incidents and who sent the survey. Use surveys for sentiment and label them as sentiment.
Name exceptions and near misses	A section reporting no exceptions is not a clean section. It is an unmonitored one, and an experienced director reads it that way.
Separate measured from inferred	Both belong in a pack. Mixing them without a marker is what turns a reasonable estimate into an accusation of spin later.
Give every number an owner	Not a team. A person who can be asked, and who knew their name was attached before the pack was printed.

THE ONE-SENTENCE TEST

Every number in the pack should reduce to: this figure, measured this way, from this system, as at this date, owned by this person. If it will not reduce, it is an assertion, and it should be labelled as one rather than dressed as something else.

On reporting bad news

A pack that improves every quarter is not reassuring to anyone who has sat on a board for long. Deterioration reported early, with a diagnosis and a date, buys more credibility than improvement reported without a method. The pack is establishing whether you are a reliable narrator, and that judgement is formed over several cycles rather than in one.

This has a practical consequence for the first pack under a new standard. It will look worse than the one before it, because it is measuring things that were previously asserted. Say that explicitly in the covering note, before somebody concludes performance fell off a cliff.

WORDING THAT HELPS

"The figures below are lower than last quarter's because they are now drawn from the deployment log rather than team self-assessment. The underlying position has not deteriorated. The measurement has become honest, and the gap between the two numbers is itself the finding."

The standing pack

Eight standing items, in the order a board actually wants them. What is running and what it produced comes first. The assurance sits underneath as the warrant, not on top as the subject.

ITEM	WHAT IT ANSWERS	EVIDENCE IT NEEDS	TIER
1. What is live	What AI is running, where, and what business job it is doing	An inventory written in business terms, held in a system of record, with the date it was last reconciled against reality	Attested
2. What it returned	What moved, against what, since when	Paired metrics with baselines dated before the change and the measurement method stated. Unbaselined claims flagged, not dropped	Attested
3. What is in pilot	What is in flight, what decides next, and what was stopped	Portfolio with go or no-go dates, forecast benefit, and stopped pilots with the reason and the money released	Mixed
4. What depends on it	Not seat counts. What degrades if this stops on Monday	Usage drawn from the platform, plus a stated view of the fallback	Attested
5. Assurance	Whether the four items above can be believed, and whether we are lawful	Material systems mapped to principal risks, controls operated against bypassed, and the regulatory position with evidence held	Attested
6. Incidents	What went wrong, how fast it was found, what changed	Named incidents with detection and response times, including near misses that cost nothing	Attested
7. Capability	Who is competent to supervise this	Role-mapped competence records, not attendance at an awareness session	Artefact
8. Decisions	What the board needs to decide, and by when	Each decision paired with the evidence it rests on and what is still unknown	Mixed

The items people get wrong

Item 2 is where credibility is won or lost, and where most packs overreach. One result with a dated baseline and a stated method is worth more than three without, because the three get challenged together and none survives. Report the one and say the others are being baselined.

Item 3 is most often missing entirely, and boards notice its absence faster than anything else. They want the portfolio: what is in flight, what decides this quarter, what was killed. A stopped pilot reported with its reason is among the strongest things you can put in a pack, because it shows the measurement is real. A portfolio where nothing is ever stopped says the opposite.

Item 5 fails when the control reports on itself. If the evidence that a review gate operated comes from the team it would have blocked, the board is reading an assertion with extra steps. The record has to come from somewhere the reporter does not control.

Item 7 is where AI literacy obligations land. Attendance is an artefact. Competence to supervise is a different claim needing different evidence, and a board attesting to it wants more than a completion percentage from a learning platform.

THE SAME RECORDS, TWO FORUMS

The main board gets items 1 to 4 in full and item 5 compressed to a paragraph, because what it needs from assurance is permission to believe the first four. The audit or risk committee gets item 5 in full, with the materiality mapping, the exception register and the Provision 29 evidence. Same underlying records, different depth. Build them from one source, because if the two ever disagree it is the audit committee that will find it.

Regulated and listed callouts

The general structure above works anywhere. These are the additions that apply where a specific obligation bites. Obligations are described here; the dates attached to them sit in Appendix C, so that a change of timetable does not invalidate the guidance.

UK listed companies, and those adopting the Code voluntarily

Provision 29 requires a declaration of effectiveness, not a description of arrangements. The practical consequences for an AI pack are three. Material AI controls need identifying and mapping to principal risks. Their operation needs evidencing across the period, not at a point. And any that did not operate effectively need naming, with remediation, in the annual report.

The FRC's position is that information collected internally through the framework's own monitoring components can be relied on for the declaration. That is permissive and it is also demanding: it means your internal monitoring is the evidence, so its quality is now a reporting matter rather than an operational preference.

Organisations in scope of the EU AI Act

Three obligation families generate reportable evidence. AI literacy, which attaches to the people who deal with AI systems and which a capability section has to answer for. Transparency, where systems interact with people or generate synthetic content, which needs disclosure that holds across every touchpoint rather than at launch. And, for high-risk systems, the fuller regime of risk management, logging, human oversight and record keeping.

The obligations that were deferred were deferred largely because the technical standards underpinning them were unfinished. The ones that stayed are the ones where evidence was already achievable. That is a reasonable guide to where supervisory attention will land first.

Where an assurance standard helps, and where it stops

ISO/IEC 42001 specifies an AI management system and is certifiable by an accredited body. ISO/IEC 38507 addresses governance of AI by the governing body itself. The pairing is useful in a board context precisely because it names the split: 38507 describes what the board is doing, 42001 describes what management is running. A pack that uses that vocabulary makes the division of responsibility legible without having to argue it.

THE LIMIT

Certification evidences that a management system exists and is audited on a cycle. It is a tier two artefact of high quality. It is not a tier three record of any control operating, and a declaration of effectiveness cannot rest on it alone. Report it as what it is and it strengthens the pack. Report it as proof of control and the first director who has been through an ISO audit will say so out loud.

Financial services

Financial services leads on responsible AI maturity in the McKinsey data, which raises the bar rather than lowering it. Supervisors here already expect evidence of operational resilience, outsourcing oversight and model governance, and AI systems inherit those expectations rather than sitting outside them. So an FS pack should map AI controls into the existing frameworks rather than building a parallel structure beside them.

The groundwork

Everything so far assumes you can reach the records. Most managers cannot, and that is the real starting position rather than an embarrassing exception to it.

Five things have to be true before a pack can carry attested evidence. None of them is a technology problem in the interesting sense. All of them are conversations that somebody has to have, usually with people who do not report to you.

1. Query access you do not have to ask for

A dashboard someone else maintains is not access. What you need is the ability to run the same extract next quarter without a favour, ideally with the query written down. If every number in your pack depends on one helpful person in platform engineering, the pack degrades the week they change role, and it will degrade silently.

2. A written definition for every recurring number

The most common way a pack loses credibility is not a wrong figure. It is a figure whose definition quietly changed between quarters, so the trend line describes two different things. Definitions belong in a dated, versioned dictionary, and a change to a definition is itself reportable. Say it in the pack when it happens.

3. At least one control that emits

Most controls are procedural. A rule exists, people are expected to follow it, and nothing is produced either way. A control that emits leaves a record when it operates and, more importantly, when it is bypassed.

If none of your controls emit, your first engineering request is not a new control. It is making an existing one leave a trace, which is usually a much smaller piece of work and considerably easier to get funded.

4. Somewhere to put the record that you do not administer

Evidence held by the team being assessed is weak evidence, and an experienced audit chair will say so. The fix is usually cheap and the conversation is usually hard: the log lands in the platform, security or audit estate rather than in delivery's own store. Ask for this early, because it is an ownership negotiation rather than a technical change.

5. A named accountable person, agreed above you

McKinsey's finding is that organisations with explicit ownership for responsible AI score markedly higher on governance maturity than those without. The practical version is blunter. If nobody senior to you has agreed in writing that they own this, your pack is a personal opinion with a company logo on it, and it will be treated that way the first time it says something unwelcome.

IF YOU HAVE NONE OF THE FIVE

Report that. A first pack that says here is what we can evidence, here is what we cannot, and here is the sequence for fixing it is more credible than a confident pack built on nothing, and it dates the baseline for everything that follows. The worst opening move is a green pack you cannot defend in the second meeting.

The five conversations, in this order

WHO	WHAT YOU NEED FROM THEM	WHY THIS ORDER
Your own director	Agreement on what the pack is for, and that it will contain bad news.	Without this you will be asked to soften the first uncomfortable finding, and the pack never recovers from that precedent.
Risk or compliance	Their working definition of materiality, and whether AI systems are already inside it.	Defining materiality alone is how a delivery manager ends up owning a judgement that belongs to the second line.
Internal audit	What they would test, and what evidence they would accept.	Getting this before you build is free. Getting it after is a finding.
Platform or security engineering	Where records can land outside your control, and what it costs.	This is the long-lead item. Start it before you need it.
The company secretary	Page limit, submission deadline, house template, and the annex convention.	They control the format and the cut-off, and almost nobody speaks to them early. Doing so is the cheapest advantage available.

In a federated group, multiply the middle three by the number of divisions and start earlier than feels necessary.

If you are starting from nothing

The minimum viable first move is far smaller than most people assume, and the instinct to do it properly across the estate is what kills it. Pick one material system. Establish one baseline, dated. Make one control emit. Get one record landing somewhere you do not administer. Get one name agreed above you.

That is roughly a quarter of work for one person, and it produces a pack containing exactly one attested claim. One is enough. It establishes the pattern, it proves the plumbing, and it gives you a working example to point at when you ask for the other thirteen.

THE FAILURE MODE TO AVOID

Attempting to instrument everything at once, discovering it needs a programme, requesting a programme, and reporting nothing for two quarters while it is considered. The board sees no change and concludes the problem was overstated. Meanwhile the first honest pack, produced in ten weeks with one real number in it, would have made the case for the programme by itself.

The first ninety days

Sequenced so that each step produces something usable if the next one stalls. None of it requires budget approval, which is deliberate: a plan that needs funding before it produces evidence tends not to produce evidence.

WEEK 1-2

Enumerate, including the awkward ones

List every place AI is in use. Include the systems that predate the AI programme, the ones bought on a departmental card, and the features that shipped before there was anyone to register them with. Record the date of enumeration, not the date of the document.

DONE LOOKS LIKE: a dated inventory with named owners, and an explicit list of areas not yet swept.

WEEK 3-4

Tier every number already in the pack

Take last quarter's board pack and mark each figure assertion, artefact or attested. Do not fix anything yet. The distribution is the finding, and it is usually the most persuasive slide you will produce all year.

DONE LOOKS LIKE: the existing pack, annotated, with a count per tier.

WEEK 5-6

Fix the baselines, or declare their absence

For every value claim, find the baseline. Where one exists, check it was measured the same way and dated before the change. Where none exists, stop claiming the saving and start the baseline now so that next quarter has one.

DONE LOOKS LIKE: each value claim either baselined or reclassified, with no silent reclassifications.

WEEK 7-9

Instrument the two controls that matter most

Not all of them. Pick the two whose failure would be hardest to explain, and make them produce a record when they operate and when they are bypassed. The record must land somewhere the delivery team does not administer.

DONE LOOKS LIKE: two controls producing exception records, with at least one real exception captured.

WEEK 10-12

Rewrite the pack, then have it attacked

Restructure to the eight standing items. Then give it to someone whose job is to disbelieve you: internal audit, second line, a non-executive with a forensic habit. Ask them to find the weakest number, and fix or footnote it before the meeting rather than after.

DONE LOOKS LIKE: a pack that has survived a hostile read, with the caveats written by you rather than discovered by them.

IF YOU ONLY DO ONE OF THESE

Do weeks 3 and 4. Tiering the existing pack costs almost nothing, requires no permission, and produces the single artefact most likely to get the rest of the work funded. It also tells you honestly where you are, which is harder to find out than it should be.

Communicating it upward

A director will repeat one sentence of your pack in a meeting you are not in. Decide which sentence that is, and make it the first one.

Constructions that survive, and the ones that do not

The difference is almost never tone. It is whether the sentence contains something a listener could check.

DOES NOT SURVIVE

SURVIVES

We have robust controls in place across the estate.	Eleven of fourteen material systems produced control records this period. The three that did not are named below.
Adoption is strong and continuing to grow.	Nine of fourteen units are using it, covering 71 percent of engineers. Two units have not started and one did not report.
We are compliant with the relevant requirements.	We meet the transparency and literacy obligations and can evidence both. We cannot yet evidence review substance, and that gap is scoped.
AI has saved approximately 4,000 hours this year.	Median cycle time fell from 9.2 to 6.4 days since March, measured the same way before and after. We have not converted that to a cash figure.
No significant issues to report this period.	Three exceptions, all detected within 48 hours, none reaching a customer. Detail in Annex B.
We are broadly in line with industry peers.	We have no reliable peer data. Here is our own trajectory across three periods, which is the comparison we can stand behind.

Delivering bad news

The order is fixed and it matters more than the wording.

1. **What happened**, in one sentence, with the number.
2. **Why**, as a diagnosis rather than a narrative. "Alerting was built around human-authored change patterns" beats "a combination of factors".
3. **What it means for the declaration**, said plainly, because that is the question forming in the room whether or not anyone asks it.
4. **What changes and by when**, with an owner who is present.

Do not open with context. Context before the finding reads as preparation of the ground, and experienced directors recognise the shape of it immediately.

The three questions you will be asked

THE QUESTION

THE TRAP

THE ANSWER THAT HOLDS

Is this a problem?	Answering yes or no. Both are wrong and both will be quoted back.	Name the exposure and the trigger. "It becomes a problem if X, and we would know within Y."
--------------------	---	---

THE QUESTION	THE TRAP	THE ANSWER THAT HOLDS
How do we compare to our peers?	Reaching for a vendor benchmark whose peer set you cannot see.	Decline the comparison and substitute your own trajectory. Say why the available benchmarks are not usable.
Can you guarantee this will not happen?	Any form of yes.	No. Offer detection instead. "I cannot prevent it. I can tell you we would detect it within thirty-one hours, and that number is falling."

WHEN YOU DO NOT KNOW

Say so, then say how you would find out and by when. Never estimate a number aloud in a board meeting. A guessed figure becomes a commitment the moment it is minuted, and you will spend the next quarter defending an answer you invented under pressure.

The board is not one audience

Three people will read your page one for three different things, and all three should find theirs inside a minute.

WHO WANTS WHAT

- **The chair** wants the position, in a sentence they can repeat
- **The audit or risk chair** wants the exceptions and the things you cannot evidence
- **The commercially-minded NED** wants what moved and what it cost
- **The executive who sponsored the spend** wants to know whether they are about to be embarrassed

IN THE MEETING ITSELF

- Do not present the pack. Reading it aloud punishes everyone who read it
- Open with what changed since the pack was written
- Go to the asks early, while attention is highest
- Take the hardest question first if you can choose
- Write down anything you could not answer, and send it within 48 hours

The last of those is the one that builds standing over cycles. A manager who reliably closes their own open questions before being chased becomes the person whose numbers are not interrogated line by line, which is the only real reward available in this job.

The evidence ladder

One page. Print it and keep it next to the pack while you write.

TIER	WHAT IT IS	HOW A DIRECTOR VERIFIES IT	WHAT IT CAN CARRY
Assertion	Someone says it. Status colours, confidence ratings, unmethodded percentages.	They ask you again and get the same answer.	Judgement and forward view, clearly labelled as such. Never effectiveness.
Artefact	A document exists. Policy, register, assessment, certificate, standard.	They request the document and read it.	Design and intent. Evidence that you thought about it and wrote it down.
Attested	A record made by the system at the moment of action, outside your control.	They query a system you do not administer, and see failures as well as passes.	Operation. Effectiveness. Anything a declaration depends on.

The four questions that sort any line

1. **What would someone have to do to check this?** Ask you, request a document, or query a system. That answer is the tier.
2. **Does the decision it supports need a higher tier than it has?** If yes, either raise the evidence or lower the claim. Both are acceptable. Neither is optional.
3. **Would this show a failure if one had occurred?** Evidence that can only come out positive is not evidence. It is a filter.
4. **Who is the person, not the team?** If nobody's name attaches, nobody checked it.

THE DISTINCTION THE WHOLE PAPER RESTS ON

Documented is not the same as effective. A control that exists on paper and a control that demonstrably changes behaviour at the point of action are different objects, and only the second one can be declared.

RAISE THE EVIDENCE

- Move the record out of the reporting team's control
- Make the control emit on bypass, not only on pass
- Fix the denominator so coverage means something
- Date the baseline before the change

OR LOWER THE CLAIM

- "We assess" rather than "we have assured"
- "Reported by teams" rather than "measured"
- "Indicative, no baseline" rather than a percentage
- "Not currently instrumented" rather than silence

The board pack, drawn up

You are not writing into empty space. You are adding to a document that is already too long, and the thing making it too long is reporting exactly like yours.

B.1 The constraint you are writing into

Research by Board Intelligence with Cambridge Judge Business School, published January 2026, found more than half of board packs now run past 200 pages, with some approaching a thousand. Directors spend just under four hours reading, about 30 percent more than the same research found in 2011, while meeting frequency and duration have barely moved. On the researchers' own benchmark of 30 pages an hour read for comprehension, close to half of the average pack goes unread.

The cause is the uncomfortable part. Among packs that had grown, 92 percent carried more compliance and regulation reporting and 88 percent more risk reporting. Meanwhile only a third of meeting time goes on strategy, and directors whose packs had not grown were the ones reporting more strategic focus.

WEIGH THIS BEFORE YOU QUOTE IT

The sample is 50 people and 72 percent of them are company secretaries rather than directors, so it is largely secretaries reporting on behalf of boards. The 30-pages-an-hour figure is the firm's own advisory rule of thumb, not an independent measurement. Board Intelligence sells board reporting software and report-writing training, so the finding that packs are too long and badly written sits with its commercial interest. The Cambridge Judge collaboration raises the standing without removing that. Use the direction, not the decimal places.

The direction is enough to set the rule. Your AI reporting is, structurally, more compliance and risk reporting arriving in a pack already drowning in it. Adding forty pages does not demonstrate rigour. It demonstrates that you have not understood the problem you are describing.

THE DISCIPLINE

One page in the pack. Everything else in an annex that no director is obliged to read, and which exists so that the one who wants to check something can.

B.2 What the summary page must carry

This page is read by everyone. Most of the rest is read by one or two people. Write it last, put it first, and give it these six elements in this order.

#	ELEMENT	WHAT IT SAYS	BUDGET
1	The position	One sentence a director could repeat accurately to a regulator without having read anything else.	1 sentence
2	What changed	Since the last pack. Movements, not levels. If nothing changed, say that, because it is itself a finding.	3 bullets

#	ELEMENT	WHAT IT SAYS	BUDGET
3	The three numbers	Not eight. Three, each shown as a pair or with its denominator, each with its source system named.	3 lines
4	What is worse	The deterioration, the exception, the thing you would rather bury. Named, with a diagnosis and a date.	2 bullets
5	What we cannot evidence	The claims still sitting at assertion, and when they will be instrumented. Boards forgive gaps. They do not forgive discovering them.	2 bullets
6	The asks	Decisions required, each with a date, an owner and the evidence it rests on. A pack with no asks trains the board to treat AI as reporting.	1–3 items

TEST FOR THE SUMMARY PAGE

If the bad news is not on this page, it is not in the pack. An annex is where a director goes to verify something they already know about, not where they discover it.

B.3 Six visuals that earn their page

A chart earns its place when it shows something prose cannot say in the same space, or when it makes a gap visible that a sentence would let you round off. These six do.

VISUAL	WHAT IT SHOWS	WHY IT BEATS A PARAGRAPH
Evidence tier mix, this period against last	Stacked bar. How many claims in the pack sit at assertion, artefact and attested.	It reports on the reporting. Nothing else shows a board whether its own visibility is improving, separately from whether the programme is.
Exceptions over time	Columns by period. Control bypasses, overrides, break-glass uses, near misses.	It reframes zero. A flat line at nothing reads as clean until it is plotted, at which point it reads as unmonitored.
The coverage funnel	Systems known, then mapped to risk, then instrumented, then evidenced this period.	It puts the denominator on the page. Coverage claims survive prose and do not survive a funnel.
Unit by item grid	Business units down, the eight standing items across, shaded by evidence tier.	It stops the weakest unit disappearing into a group average. In a complex group this is the single most valuable chart in the pack.
Paired trend	Two series, one axis. Throughput with instability. Volume with rework. Speed with escaped defects.	It makes the trade visible rather than arguable, and stops one half of a pair being quoted on its own later.
Obligation and decision timeline	Horizontal. Regulatory dates above the line, decisions the board must take below it, with lead times shown.	It converts compliance from a status into a schedule, which is the only form a board can actually act on.

B.4 Five that should be refused

DO NOT PUT THESE IN

- **RAG grids with no stated threshold.** A colour with no rule behind it is an assertion in costume.
- **Radar and spider maturity charts.** The shape averages the weakest dimension into invisibility, and the weakest dimension is the one that fails.
- **Adoption curves with no denominator.** A rising line of users tells a board nothing about penetration or dependence.
- **Trend lines through three points.** Three points is a shape, not a trend.
- **Vendor benchmark comparisons** where the peer set is not visible and the vendor sells the remedy.

RULES FOR EVERY CHART THAT STAYS

- Denominator on the chart, not in a footnote
- Source system named in the caption
- Axis starting at zero, or an explicit note saying why not
- The period covered, stated, not implied
- Exceptions plotted, not filtered out to make the line clean
- One chart per point. A chart carrying three arguments carries none

B.5 The aggregation trap in a complex group

Organisations with boards are rarely simple. A group with fourteen business units, several legal entities, two regulators and a shared services function has fourteen different answers to every question in the pack, and the act of combining them is where the useful information usually dies.

RULE	WHY
Aggregate numerators and denominators, never percentages	Averaging percentages across units of unequal size produces a figure that describes no unit and no group. It is arithmetic that looks like analysis.
Report the worst unit by name beside the group figure	The group figure is the position. The worst unit is the risk, and the risk is what the declaration turns on.
Show n per unit, and name units that did not report	A non-responding unit is not a zero and it is not the average. Silently excluding it is the most common way a group figure becomes untrue.
Separate designed centrally from operated locally	A control designed by group and run by fourteen units has one design record and fourteen operating records. Effectiveness lives in the fourteen.
Get third-party evidence contractually or admit you have none	Where a joint venture, outsourcer or managed provider runs the system, evidence arrives because the contract requires it. If it does not, you have a relationship, not assurance.
Reconcile entity packs with the group pack before either goes out	Two packs describing the same system with different numbers is the most damaging single thing an audit committee can find, because it puts every other number in question at once.

In a federated group, expect the first honest pack to show wide variance between units. That variance is the finding. Smoothing it is how a group ends up declaring effectiveness on the strength of its best division.

The skeleton

One page in the pack. Four pages of tabled detail behind it. Everything else in an annex.

PAGE	PURPOSE	WHAT GOES ON IT	VISUAL
1. Position and asks IN THE PACK	The whole story for a director who reads nothing else. Ninety seconds.	The six elements from B.2, in that order, on one side of paper.	Evidence tier mix, small, top right.
2. Exposure and controls IN THE PACK	What is material, and did the controls over it operate.	Material systems mapped to principal risks. Controls operated against bypassed, with the worst unit named.	Coverage funnel. Unit by item grid if the group is federated.
3. Value and adoption IN THE PACK	What moved, against what, and what now depends on it.	Paired metrics with baseline dates. Dependence, not seat counts. Unbaselined claims flagged as such.	Paired trend, two series on one axis.
4. Incidents and capability IN THE PACK	What went wrong, how fast it was found, and who is competent to supervise.	Named incidents and near misses with detection times. Competence by role, not attendance.	Exceptions over time.
5. Decisions IN THE PACK	What the board must decide, by when, and on what evidence.	Each ask with owner, date, evidence tier, and what remains unknown.	Obligation and decision timeline.
A. Method ANNEX	How every figure was produced.	Per metric: definition, source system, query or extract, owner, date.	None.
B. By unit ANNEX	The eight items for every business unit and entity.	Full detail, including non-responders and the reason.	None.
C. Register ANNEX	The inventory and materiality mapping in full.	Systems, owners, classification, last reconciliation date.	None.
D. Obligations ANNEX	The regulatory position, with dates isolated.	Obligation, status, evidence held, gap, owner.	None.

WHY THE ANNEX MATTERS MORE THAN IT LOOKS

The annex is not filler and it is not defensive padding. It is the thing that lets the front five pages be short. A director who wants to test a number can find it; everyone else is spared. Written the other way round, the pack grows every quarter and gets read less every quarter.

Working sheet: fill this before you write page 1

The position, in one sentence a director could repeat to a regulator.

.....

The three numbers, each with its pair or denominator and source system.

1 SRC
2 SRC
3 SRC

What got worse this period, and the diagnosis.

.....

Worst-performing unit, named, and its figure against the group figure.

Unit Unit figure Group Non-responders
.....

What we still cannot evidence, and the date it will be instrumented.

Claim Instrumented by

The asks. Decision, owner, date, evidence tier it rests on.

.....

Three checks before it goes in the papers

READY

- Page 1 contains the worst news in the pack
- Every number reduces to figure, method, system, date, owner
- No annex figure contradicts a front-page figure
- The pack is shorter than last quarter, or you can say why not

NOT READY

- A section with no exceptions and no explanation of why
- A group percentage built by averaging unit percentages
- A borrowed external statistic doing load-bearing work
- Any chart whose denominator lives in a footnote

A worked example

A specimen page one. Composite and illustrative, drawn from recurring patterns rather than any single organisation.

A listed group, fourteen business units, shared engineering function, December year end. Nothing below is a real company and no figure should be reused.

AI – BOARD SUMMARY

Progress, value and assurance · Q2 FY26 · Owner: Director of Engineering · Prepared 14 August
· Page 1 of 5, annexes A to D

POSITION

Eleven AI deployments are live across nine of fourteen business units. Two return measured improvements against dated baselines, one is too early to judge. Four pilots are in flight, two deciding this quarter. We can evidence control operation for eleven of the fourteen systems classified as material; the three we cannot are in Retail Operations, instrumented by 30 November.

WHAT IS LIVE, AND WHAT IT RETURNED

- **Claims triage**, six units, live since March. Median handling time **14.2 to 9.1 minutes**, measured the same way before and after. Source: case management system.
- **Code assistance**, engineering, live since January. Median cycle time **9.2 to 6.4 days**; change failure rate **11 to 14 percent** over the same period. Source: deployment pipeline. Both moved; we report the pair.
- **Contract review**, Legal, live since May. Too early for a baselined result. First comparison due Q3.

PILOTS IN FLIGHT

- Four running. Two decide this quarter: supplier onboarding and forecast reconciliation. Criteria recorded in Annex B.
- One stopped in May. Forecast benefit did not survive baselining; budget returned to the portfolio rather than reallocated inside the programme.

ASSURANCE POSITION

- Material systems producing control records: **11 of 14**. The three sit in Retail Operations, which has not reported for two periods.
- Reviews recorded against merges: **1,190 of 1,240**. Fifty under break-glass, each with a named approver. Source: merge log.
- Transparency and literacy obligations met and evidenced. Detail in Annex D.

WHAT IS WORSE

- Mean time to detect AI-related incidents **31 hours** against 9 for non-AI. Diagnosis: alerting was built around human-authored change patterns. Remediation funded, delivery 28 February.
- Retail Operations non-reporting. Escalated to the divisional managing director on 14 May. Unresolved.

WHAT WE CANNOT EVIDENCE

- Whether break-glass approvals were substantive. We can show they happened and by whom, not what was assessed. Instrumentation scoped, not funded.
- Third-party model provider controls. The contract does not require evidence. Renewal is March, clause drafted.

ASKS

1. Approve funding for review-substance instrumentation. Decision by 30 September to land before year end.
2. Direct Retail Operations to report from Q3. Board authority needed; two escalations have not moved it.
3. Note the intended Provision 29 disclosure wording for the Retail exception. Annex D.

Why it is written that way

CHOICE	REASON
It opens with deployments and returns, not with controls	The board's agenda item is AI, not AI governance. A pack that opens on control coverage answers a question the board did not ask and buries the one it did.
The assurance block is fourth and short	It licenses the first three blocks rather than being the subject. Three lines is enough at main board; the audit committee gets the full version.
Every result carries its baseline and method	14.2 to 9.1 minutes, measured the same way before and after. Without the method clause it is an assertion with a decimal point on it.
Code assistance reports a pair that disagrees	Cycle time improved and change failure rate worsened. Reporting only the first would be the single most quotable sentence in the pack, and the most damaging when somebody later found the second.
The stopped pilot is reported, with the money	A portfolio where nothing is ever killed tells a board the measurement is decorative. Naming the stop, and returning the budget rather than quietly reallocating it, is what makes the surviving forecasts credible.
Denominators are on the page	1,190 of 1,240, not 95.9 percent. The ratio survives being quoted. The percentage detaches from its base within one retelling.
The non-reporting unit is named twice, and nothing is green	In a federated group the unit that does not report is the risk, and an annex is where it goes to die. There are no status colours anywhere on the page: every claim is a count, a ratio or a date, which removes the threshold argument before it starts.

The same period, written the usual way

WHAT MOST PACKS WOULD HAVE SAID

Our AI programme continues to make strong progress. Several deployments are now live and delivering significant efficiencies, with positive feedback from teams. A number of promising pilots are underway. Controls are in place across the estate and operating effectively, and we remain on track for compliance with upcoming regulatory requirements. No significant issues to report this period.

Five sentences, no checkable claim, and the same quarter underneath. It even leads with value, which is the right instinct. But "significant efficiencies" is not a number, "several deployments" is not a count, and "promising pilots" is not a decision date. It takes one question to unravel: which ones, and how do you know.

Note what the bad version costs beyond credibility. It does not surface the detection-time problem, so nothing gets funded. It does not name Retail, so the board never exercises the authority that would fix it. It reports no issues, so the exception data that took a quarter to produce buys nothing at all. The weak pack is not merely less honest. It is less useful to the person writing it.

THE TEST, ONE LAST TIME

Read your own page one as though you were a director who intends to ask exactly one question. If you can predict the question and you would rather they did not ask it, that is the sentence to rewrite.

Sources, dates and how to weigh them

Every source used, with its age and its interest. Dates are isolated here deliberately, so a change of timetable does not invalidate the guidance in the body. Verified 3 August 2026.

SOURCE	DATE	HOW TO WEIGH IT
UK Corporate Governance Code 2024, Provision 29. Financial Reporting Council.	Code 2024; Provision 29 effective for financial years beginning on or after 1 Jan 2026	Primary standard-setter text. Premium-listed companies, with wider voluntary adoption encouraged. Read the FRC text rather than consultancy summaries, which vary on materiality.
State of AI trust in 2026. McKinsey AI Trust Maturity Survey, approx. 500 organisations.	Fielded Dec 2025 to Jan 2026; published 25 Mar 2026	Recent and directly relevant. Self-reported by respondents who own AI governance and scored on McKinsey's own maturity model. McKinsey sells governance advisory.
The Board's AI Moment. Protiviti and BoardProspects, 772 directors and executives.	Fielded Q4 2025; published 18 Mar 2026	Correlational, single-respondent, halo-prone. Figures here taken from the press release rather than the full report. Protiviti sells board advisory.
The state of AI in 2025. McKinsey Global Survey.	Published 5 Nov 2025	Nine months old. Best available on the adoption-to-value gap; treat exact percentages as indicative.
State of AI-assisted Software Development. DORA, Google Cloud, approx. 5,000 professionals.	Fielded Jun to Jul 2025; published Sep 2025	Eleven months old, predating current agentic tooling. Retained because DORA is the established delivery measurement reference. Directions likely hold, magnitudes dated.
ISO/IEC 42001:2023 and ISO/IEC 38507.	42001 published Dec 2023	Standards texts. Note that almost every accessible commentary is published by a certification body or GRC vendor with a direct commercial interest; treat cost and timeline claims accordingly.
The GenAI Divide. Project NANDA, MIT Media Lab.	Jul 2025	Used only as a worked example of an unverifiable statistic, not relied on for any claim. Preliminary working paper; the critique draws on secondary commentary, itself a stated limitation.
How AI Impacts Skill Formation. Shen and Tamkin, randomised controlled trial, n=52.	29 Jan 2026	Small-sample RCT, preprint. Relevant to the capability item: adoption is not competence, and the widest gap measured was in debugging, the skill supervision depends on.

A note on method

Primary sources under three months old were preferred, and where the best available source is older its age is flagged inline. Correlational findings are described as correlational, and where an author has a commercial interest in their own conclusion that interest is named. This is the standard the paper asks a board pack to meet.

The paper in five lines

1. A board report on AI is now an evidence artefact, because a declaration hangs off it.
2. Every line should answer one question: how do you know, and at which tier.
3. The value case and the assurance case draw on the same instrumentation, so they are one report rather than two.

4. Report pairs, date your baselines, name your exceptions, and own your caveats before somebody else finds them.
5. Documented is not effective. Only the second one can be declared.

Jason Cruz Falzon is the creator of the Cognitive Delivery Framework and the Cognitive Delivery Index. He writes on governance, provenance and accountability in AI-assisted software delivery.

cognitivedelivery.co.uk · jason@cognitivedelivery.co.uk

Reporting AI to the board, version 1.0, August 2026. This paper is general guidance and not legal, audit or financial advice. Obligations differ by jurisdiction, listing status and sector. Take advice on your own position.