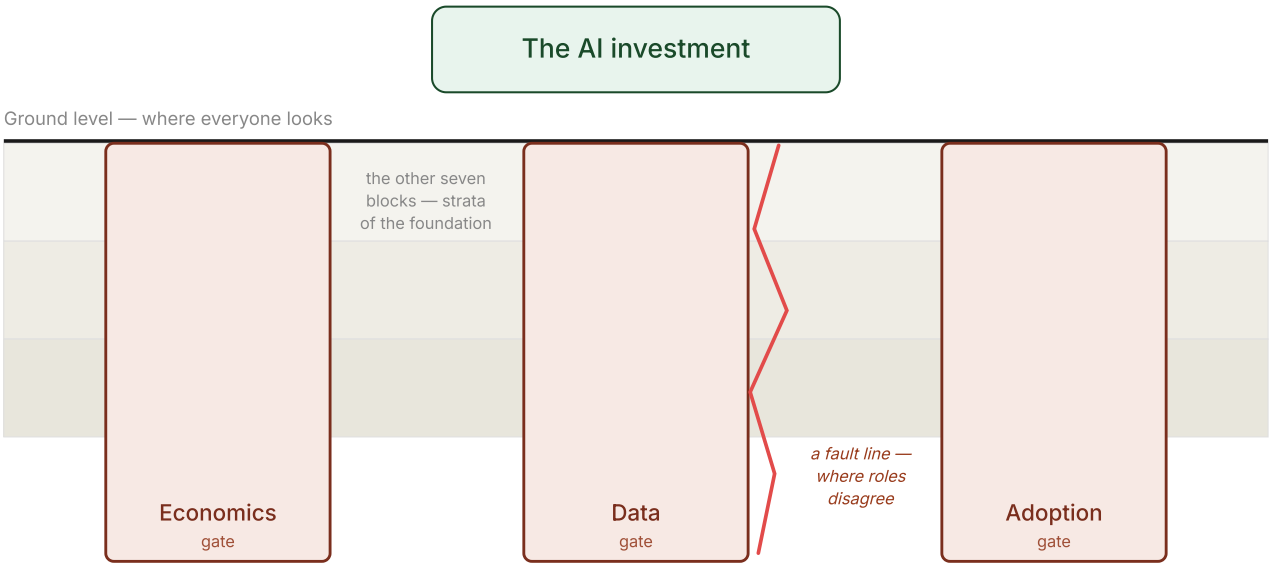
CONCEPT NOTE · VERSION 1.0 · PUBLISHED FOR COMMENT

Grounded

A framework for testing whether an organisation is ready for the AI investment in front of it

Most AI investments do not fail because the technology is weak. They fail because the foundation underneath them — the data, the economics at scale, the people whose behaviour must change, and the alignment of the leadership team itself — was never tested before the capital was committed. Grounded is an open diagnostic framework for running that test: ten blocks, thirty observable evidence questions, three gates that cannot be averaged away, answers collected privately by role, and an output that is a verdict with a priced register of gaps — not a score. This note sets out the reasoning, the method, and the design decisions behind it, including the ones we expect to be challenged on. It is published to be read, used, and pushed back against.



The argument in one page

Boards approve AI budgets expecting transformation. What they are often funding is a sophisticated analytics experiment with no clear line to the P&L — and the failure, when it comes, arrives eighteen months later wearing the costume of a technology problem. It almost never is one. The recurring causes are foundational: data that flows but cannot support a decision, pilot economics that quietly die at rollout, frontline people who were never going to change their behaviour, and — least visible of all — a leadership team that never actually agreed on what was true, because nobody asked them separately.

Existing tools do not catch this, for two structural reasons. They run in **reflection mode**: they ask people to assess themselves, which works only if you already know what you don't know. And they run in **consensus mode**: they are filled in by groups, where the highest paid person's opinion becomes everyone's answer. A canvas completed in a workshop records agreement with authority, not truth.

Grounded replaces reflection with diagnosis. Thirty observable evidence questions across ten blocks, governed by one rule: **a "Yes" only counts if you can name the artifact that proves it** — the document, the dashboard, the person, the date. Three blocks — unit economics, data, and adoption — are gates: fail one, and no amount of excellence elsewhere can average it away. Every role answers alone, before any meeting; where their answers split is a **fault line**, and the pattern of fault lines is often more predictive than any score. The output is a verdict — Grounded Conditionally Grounded Ungrounded — plus a priced register of every gap, and an expiry date, because readiness decays.

The framework is open, citable, and runnable by one person with a spreadsheet. This note explains each design decision and, deliberately, the places where the framework is most exposed to challenge. We would rather it be broken in public and improved than adopted politely and ignored.

> *AI is a force multiplier. What it multiplies is the quality of the foundation underneath it.*

IN THIS NOTE

SECTION 01

The problem — mispriced transformation

> *"Automation applied to an inefficient operation will magnify the inefficiency."*
>
> — Bill Gates, *The Road Ahead* (1995)

Every major technology wave has failed enterprises in the same way. ERP programmes in the 1990s did not fail because the software couldn't post a journal entry; they failed because processes weren't mapped, data wasn't clean, and nobody had planned for the people who would have to work differently. Cloud migrations in the 2010s did not fail because the infrastructure was unreliable; they failed because cost models built for a pilot collapsed at production scale and application dependencies nobody had documented came due all at once. The pattern is stable across three decades: **the technology is rarely the thing that breaks. The foundation is.** AI is now repeating this pattern at greater speed and greater cost — and with an extra hazard the earlier waves didn't have: an AI system can fail *silently*, degrading as data drifts while every dashboard stays green.

Four failure modes account for most of the damage:

The data is connected but not decision-grade. Sensors report, dashboards render, the volume is enormous — and yet nobody can reconstruct a single moment from it. Which batch was running, which material lot was loaded, whether the stoppage at 2am was planned or unplanned. A model trained on data that cannot answer those questions learns noise, and predicts noise with great confidence.

The economics were priced at pilot scale. The pilot's return was real. But integration cost, change management, data preparation and per-inference cost all multiply at rollout, and no one in finance ever signed their name to the number at ten times the usage. The business case that reached the board was, in the precise sense of the word, mispriced.

Nobody's behaviour was going to change. The most common failure mode in enterprise AI is not technical; it is adoption. The people whose daily judgement the system is meant to inform were not consulted in its design, cannot interpret its output in terms they trust, and — after the first false positive — quietly stop looking at it. The system keeps running. Nobody uses it. The dashboards, again, stay green.

The leadership team never agreed on what was true. The CFO believed the data was ready. The IT lead knew it wasn't. Both sat in the approval meeting; only one of them spoke, and it was the one furthest from the evidence. This misalignment is the least visible failure mode because ordinary meetings are structurally incapable of surfacing it — which is why Grounded is built, above all else, to make it visible.

> **The cost of the gap is real either way.** The organisation pays to close its foundation gaps whether or not it acknowledges them — the only choice is whether that cost appears in the business case before the capital is committed, or arrives eighteen months later disguised as "the AI didn't work here." A business case that omits its foundation cost is not optimistic. It is wrong.

SECTION 02

Why existing tools don't catch it

The instruments enterprises reach for were built for different questions. The **Business Model Canvas** evaluates how a venture creates and captures value — and it works, because founders generally do know their customer segments and revenue model. The **AI Canvas of 2017** evaluated whether a team could build a working ML feature — a sensible question in an era when the hard problem was capability: finding researchers, labelling data, training from scratch. That era is over. Foundation models made capability abundant; the hard problems moved to governance, adoption, and evaluation — the organisational wrapper around the model — and the canvas never followed them there. **Maturity models**, meanwhile, measure an organisation's general capability level and return a number on a five-point scale. What they do not do is attach that number to a decision. A maturity score of 3.2 tells a board precisely nothing about whether the €600K proposal in front of it should be funded.

Beneath these individual limitations sit two structural flaws that all self-assessment tools share:

Reflection mode

What it assumes: the person filling it in already knows what they don't know.

Why it fails here: a manufacturing VP who has never built an ML system has no frame of reference for scoring data readiness. Asked "is your data decision-grade?", they will honestly answer "we have lots of data" — and score themselves a four. The tool is only as good as the respondent's self-awareness, and AI readiness is exactly the domain where self-awareness is scarcest.

Diagnostic mode

What it demands: evidence you can point at, not opinions you hold.

How it works: instead of "is your data decision-grade?", ask "can you pull, right now, thirty days of the exact data this model needs — timestamped, no gaps, no IT ticket?" A non-technical leader can answer that by looking at what exists. The question carries the expertise, so the respondent doesn't have to.

The second flaw is **consensus mode**. Canvases and maturity assessments are filled in by groups — a workshop, a steering committee, a strategy offsite. Groups anchor to authority: the phenomenon is well enough known to have an acronym, HiPPO — the Highest Paid Person's Opinion. When the plant manager says the data is ready, the shift supervisor who knows the downtime codes are entered from memory at end of shift does not contradict them in the room. The workshop output records the plant manager's confidence and files the supervisor's knowledge under silence. Every role sees a different slice of the same truth; consensus tools systematically discard the slices that disagree with seniority — which are, reliably, the slices closest to the evidence.

What the AI investment question actually requires, then, is a tool with three properties that no existing instrument combines: it must **force evidence rather than invite opinion**; it must be **answered alone, before any group forms**; and it must **end in a decision** — not a maturity level, not a heat map, but an answer to the question the board is actually asking: *should this capital move?* Grounded is an attempt to build exactly that tool, and the rest of this note describes how.

SECTION 03

The insight, and five design principles

> *AI is a force multiplier. What it multiplies is the quality of the foundation underneath it.*

This is not an AI insight. It is a transformation insight, and the single idea everything in Grounded is built on. Put a capable model on decision-grade data, honest economics, and a workforce ready to act on its output, and it multiplies all of it. Put the same model on ambiguous data, pilot-scale economics, and unconsulted operators, and it multiplies that instead — faster, at scale, with the confidence of an authority. The technology is constant across both outcomes; the foundation is the variable. A readiness framework should spend almost all of its attention below the ground line. Five principles follow:

1 **Evidence over opinion**
Every question is answerable by pointing at something that exists — a document, a dashboard, a named person, a date. A "Yes" that cannot name its artifact is not a Yes. This single rule is what separates a diagnostic from a survey, and it is developed fully in Section 06.

2 **Gates over averages**
Some failures degrade a return; a few void it entirely. Averaging treats them the same, which is how a fatal data gap hides inside a respectable total score. Three blocks — economics, data, adoption — are gates: fail one and the verdict is capped, whatever the rest of the canvas says. Section 05 defends this choice.

3 **Disagreement as data**
Answers are collected privately, per role, before anyone is in a room together — then the *pattern* of disagreement is surfaced, never the individuals. Where the CFO's Yes meets the IT lead's No, something important is true and someone doesn't know it. Section 08 makes this the framework's boldest claim.

4 **A verdict, not a score**
Scores get rounded up in board decks; a 34/50 becomes "we scored well on readiness." A verdict — Grounded, Conditionally Grounded, Ungrounded — cannot be rounded. The score still exists underneath as working material; it is simply never the headline. Section 09 explains the logic.

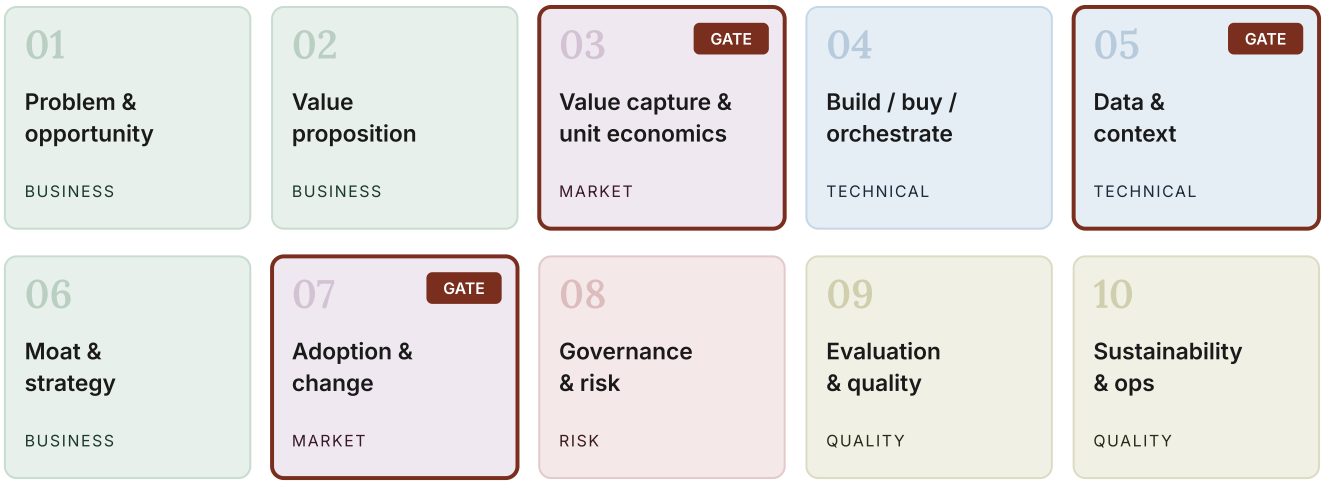
5 **Readiness decays**
Vendors deprecate models, accountable people change roles, data pipelines rot, regulation moves. A readiness assessment with no expiry date becomes actively misleading within months. Every Grounded verdict carries a 90-day validity and a defined re-grounding path.

None of these principles is exotic on its own — evidence-based assessment, independent elicitation, gating criteria, and time-boxed validity all exist in due diligence, intelligence analysis, and safety engineering. What Grounded does is combine them into one instrument aimed at one question, and make the combination open, so it can be tested, criticised, and improved in public.

SECTION 04

The ten blocks

The canvas covers ten blocks across five themes — business, market, technical, risk, and quality. Each block carries three observable evidence questions (thirty in total), and each block has a core question that states what it exists to test. Three blocks are gates, marked below and defended on the next page.



▭ = gate block — cannot be averaged away; two or more No answers fail the gate and cap the verdict

BLOCK	CORE QUESTION
1 · Problem & opportunity	What decision is slow, expensive, or inconsistent — and is AI load-bearing here, or decorative?
2 · Value proposition	What will a specific user be able to do that no non-AI alternative allows — not faster, but at all?
3 · Unit economics GATE	How does value reach the P&L — and does the margin survive at ten times the usage?
4 · Build / buy / orchestrate	What is owned versus rented — and what breaks when a rented component is discontinued?
5 · Data & context GATE	Is the data decision-grade — or merely connected? Who owns its quality, by name?
6 · Moat & strategy	What sustains advantage when every competitor can call the same API?
7 · Adoption & change GATE	Whose behaviour must change, why would they, and what happens the first time the AI is wrong?
8 · Governance & risk	Who is liable when it's wrong — and what is the worst plausible failure?
9 · Evaluation & quality	How is silent degradation caught in production — and by whom?
10 · Sustainability & ops	Who owns the system when it degrades — and what happens when the model is deprecated?

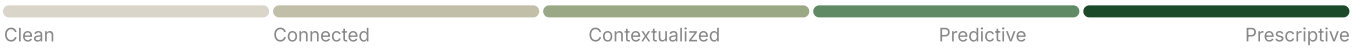
SECTION 05

The three gates

A total score treats every weakness as tradeable: strong strategy can offset weak governance, a brilliant moat can offset thin evaluation. For seven of the ten blocks, that trading logic is roughly right — weakness there degrades the return. For three blocks it is dangerously wrong, because failure there does not degrade the return; it **voids** it. These three are gates. A gate fails when two or more of its three evidence questions come back No, and a failed gate caps the verdict at Conditionally Grounded no matter how high the total score climbs. You cannot average your way past a foundation that isn't there.

Gate 3 — Unit economics. Pilot economics almost never survive rollout: integration, change-management and data-preparation costs multiply, per-inference costs compound with usage, and the benefit curve flattens as the system meets the messy middle of the organisation. The gate demands three artifacts: a model connecting the investment to a named P&L line, that same model run at full deployment scale, and a named person in finance who has signed the result. Engineering teams underestimate cost and overestimate benefit with great reliability; an independent finance signature is the structural check on optimism.

Gate 5 — Data & context. The block most investments fail on, and the one most vulnerable to self-assessment inflation — "we have lots of data" is the four-word sentence that has sunk more AI budgets than any other. Data maturity runs through five stages:



Most organisations sit between Connected and Contextualized: data flows, but shift boundaries aren't encoded, planned downtime looks identical to unplanned, and traceability breaks partway through the process. Predictive AI requires Contextualized as a floor — below it, the model learns noise. The gate's questions test retrievability now ("pull thirty days, no gaps, no IT ticket"), named ownership of quality, and whether the data has ever carried a consequential decision that turned out correct. The cost of moving from Connected to Contextualized belongs in the business case — not in a future phase.

Gate 7 — Adoption & change. The number-one failure mode in enterprise AI is not technical. An unexplainable alert is ignored after its first false positive regardless of the model's accuracy; the first conflict between the AI and an experienced person's judgement, if unresolved by written protocol, is settled informally — almost always in favour of ignoring the AI permanently. The gate tests whether the people who must change were consulted, whether the first-conflict protocol exists in writing, and whether the organisation has ever actually changed frontline behaviour through technology before.

> **Why these three — and not others?** They are the three places where failure is simultaneously common, fatal rather than merely costly, and invisible in a pitch deck. This is a design judgement, not a law of nature, and it is deliberately falsifiable: field data may show that governance (Block 8) deserves gate status in regulated industries, and industry editions of the framework are free to promote or demote gates with evidence. We hold the specific gates loosely. We hold the principle — that some blocks must not be averageable — tightly.

SECTION 06

The Evidence Rule

> *"Yes" is not an answer. "Yes" plus the name of the thing is an answer.*

Every one of the thirty questions is answered Yes Partially or No, under one rule: a Yes only counts if the respondent can name the artifact that proves it — the document, the dashboard, the named person, the date — within a few seconds. A Yes with no named artifact is recorded as Partially. No exceptions, including for the most senior person involved. The rule works because it changes what kind of claim an answer is. You can be optimistic about an opinion; optimism about a document that does not exist is simply a false statement, and people are far more reluctant to make one — especially knowing the artifact may be asked for on screen in the group session.

The rule also repairs the most abused answer in every three-point scale: the middle. In conventional assessments, "Partially" is the socially safe option that absorbs most responses while admitting nothing. Under the Evidence Rule it acquires exactly one meaning: **the artifact exists but is incomplete, stale, or could not be named on the spot.** It stops being a hedge and becomes a finding — usually pointing at a retrievability or freshness gap worth pricing.

Q: "Is there a named person in finance who has signed off the running cost of this system at ten times pilot scale?"

A: "Yes."

Q: "Who — and when did they sign?"

A: "...I'd have to check."

Recorded: Partially. Nothing accusatory has happened — but the answer is now honest.

Three supporting mechanics complete the rule. **Freshness:** named evidence older than twelve months records as Partially — a legal opinion from two years ago, a cost model built before prices moved. Foundations decay; the diagnostic must not credit the past. **The live demonstration:** one question per assessment — the *anchor question* — is never answered on a form. It is demonstrated, on screen, while the facilitator watches. In manufacturing the anchor is the *2am Question*: "If production stopped unexpectedly at 2am last Tuesday, could you tell within one hour, without calling anyone, exactly which batch was running, which material lot was loaded, and whether the stoppage was planned?" A Yes produced live is real; a Yes that "could be pulled together by Thursday" is a Partially, and everyone has just learned why. **The calibration flag:** any single role answering twenty-seven or more of thirty questions Yes is flagged before aggregation — honest organisations do not score that way; the facilitator spot-checks two named artifacts aloud. Quality control, not accusation.

Together these mechanics make the diagnostic expensive to game and cheap to answer honestly — which is the precise inversion of how self-assessment normally works, and the reason a non-technical leader can complete it without a consultant translating. The expertise lives in the questions, not the respondent.

SECTION 07

The method — from one sentence to a verdict

It begins with one sentence. Before anyone answers anything, the sponsor completes the Investment Statement: *"We are investing — so that — can —, measured by —."* Every subsequent question is answered against it. A landed statement looks like: *"We are investing €600K so that the maintenance team can predict bearing failures on Line 4 before they cause unplanned stops, measured by unplanned downtime hours per month, from 40 to under 15."* The amount need not be precise — a ceiling, a magnitude, even "not yet costed" all complete the sentence honestly; the last of these is logged as Debt Item Zero, because an organisation that cannot bound its own spend has already told you something real. And if the leadership team cannot agree the sentence at all, that is finding number one — a scoping failure discovered in five minutes rather than eighteen months.

Then each role answers alone. Four to six roles — finance, technical, operations or frontline, risk, sponsor — each receive the questions for the blocks they own, plus all three gate blocks, which everyone answers. Twenty minutes, individually invited, no conferring; that deliberate overlap on the gates is what creates fault lines. The independence is not a procedural nicety — it is the entire defence against the HiPPO dynamics described in Section 02. The shift supervisor who knows the downtime codes are fiction will say so on a private form. They will not say so across a table from the plant manager.

Aggregation happens before the room. A facilitator — one person, a spreadsheet, an hour — turns the independent answers into block scores, gate results, a fault-line report and a provisional verdict. When answers are combined, **proximity beats seniority**: on each gate block, the role closest to the evidence carries double weight — finance on economics, the technical lead on data, the frontline lead on adoption. A CFO's Yes cannot outvote a data engineer's No on data readiness, by design.

The session opens on disagreement, never on a blank canvas. Ninety minutes, all roles present, the analysis already done. The facilitator shows the pattern of divergence by role category — never by name; that protection, the *facilitator's contract*, is what made the intake honest and is non-negotiable — asks for artifacts rather than opinions, lets silence do its work when no artifact appears, converts each confirmed gap into a debt item with a cost, an owner and a deadline, and only then reveals the verdict.

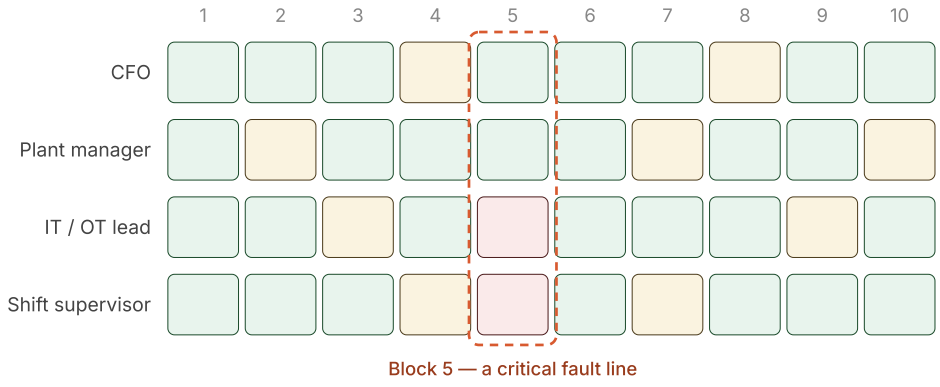
SECTION 08

Fault lines — disagreement as diagnostic

> *Two organisations can return the same score and face opposite prospects. The score does not predict failure. The disagreement does.*

When two roles answer the same evidence question differently, only two things can be true: either the artifact does not exist and someone believes it does, or it exists and someone who needs it does not know. Both are foundation failures. Both are invisible in ordinary meetings, where the answer of the most senior person quietly becomes the answer of the room. Grounded's private intake exists to capture these splits before the room can erase them, and the **Fault Line Report** — the pattern of disagreement across roles and blocks — is arguably the most valuable single output of the assessment. Fault lines come in three severities:

SEVERITY	PATTERN	MEANING
Critical	The role closest to the evidence says No; a senior role says Yes — on a gate block.	The investment case rests on evidence its owners say does not exist. Caps the verdict on its own.
Significant	A clean Yes / No split between any two roles on the same question.	The artifact is invisible to half the leadership or imagined by the other half. Resolved by producing it, not by voting.
Minor	Yes / Partially or Partially / No splits.	Shared awareness, different confidence — usually a freshness or retrievability gap.

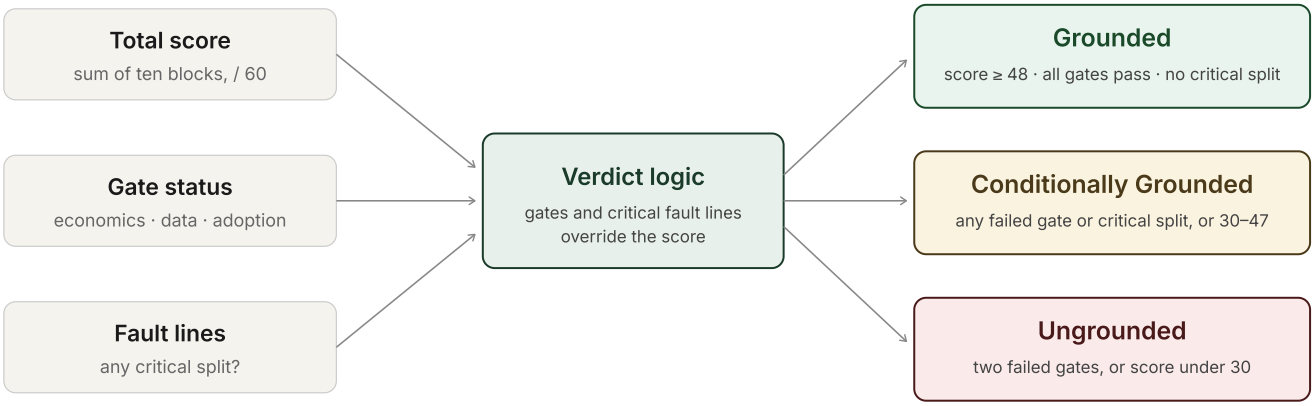


Green = Yes (named), amber = Partially, red = No. The roles closest to the data say No; the roles who approved the budget say Yes. The average score barely moves. The prognosis moves entirely.

An honest note on the claim above. "The disagreement predicts failure better than the score" is the framework's boldest statement, and today it is a design hypothesis — grounded in the well-documented HiPPO dynamics of group assessment and in practitioner experience, but not yet a validated statistical finding. Field assessments will test it, and we will publish what we find, including if it is wrong. Note, though, what does not depend on it: private intake is justified on anti-HiPPO grounds alone, and a critical fault line is worth surfacing even if its predictive power turns out to be weaker than we believe.

SECTION 09

The verdict, foundation debt, and expiry



Why a verdict and not a score. Scores invite negotiation and get rounded up on the way to the board — a 34/50 becomes "we did well on the readiness assessment" by the third retelling. A verdict cannot be rounded. It is also more honest about what the diagnostic knows: the difference between 41 and 44 is noise; the difference between a passed and a failed data gate is everything. The score survives underneath as working material for the team — it is simply never the headline. Conditionally Grounded, it is worth stressing, is not a soft no: it means *proceed, with the gaps priced in and the gate debt closed before capital is released.*

Foundation debt. Every No in the assessment becomes a debt item — a piece of foundation the investment is borrowing against — with an estimated cost to close, a named owner, and a deadline, compiled live in the session into the Foundation Debt Register. The framing is deliberate: boards do not intuit "readiness gaps," but they understand debt intimately, including the part where unacknowledged debt compounds. Debt items on gate blocks are senior debt, scheduled for closure before capital release rather than in parallel with the build. The register's total is added to the investment case before it goes to the board — which is the whole point: **an AI business case that does not carry its foundation debt is mispriced.**

Grounding expires. A verdict is a snapshot, not a certificate. Vendors deprecate models, accountable people rotate, pipelines rot, a supplier change silently invalidates a trained model. Every verdict therefore carries a ninety-day validity; if capital has not moved within the window, or a gate-block artifact has materially changed, the assessment is re-run. Re-grounding costs a fraction of the original effort because the artifacts are already named — it is closer to an audit than a fresh diagnosis — and for Conditionally Grounded verdicts it is where each debt owner reports closure. The expiry is what turns Grounded from a one-off workshop into a standing instrument of capital discipline.

SECTION 10 · ILLUSTRATIVE — A COMPOSITE, NOT A CLIENT ENGAGEMENT

A worked example

A mid-size discrete manufacturer proposes a predictive-maintenance investment. The sponsor lands the Investment Statement on the first call: *"We are investing €600K so that the maintenance team can predict bearing failures on Line 4 before they cause unplanned stops, measured by unplanned downtime hours per month, from 40 to under 15."* Specific beneficiary, specific outcome, falsifiable metric — a strong start, and already better scoped than most proposals that reach a board.

Five roles complete the private intake: the CFO, the plant manager, the IT/OT lead, the quality lead, and a shift supervisor. Most of the canvas comes back strong *with named artifacts*. The economics gate passes cleanly — the cost model exists as a file, it has been run at three-line rollout scale, and the CFO names her own sign-off with a date. The value proposition is crisp: detect the failure signature roughly 72 hours before the current vibration alarm, something no threshold-based system can do. The vendor comparison is documented; a fallback plan exists.

Block 5 is where the picture splits. Asked whether thirty days of the model's input data can be pulled right now — timestamped, no gaps, no IT ticket — the CFO and plant manager both answer `Yes`: they have seen the dashboards, and the dashboards are genuinely impressive. The IT/OT lead and the shift supervisor both answer `No`. The historian has gaps — and the gaps cluster around stoppages, precisely the events the model must learn from, because the data collector drops samples when the line electronics brown out. Worse, stoppage cause codes are entered from memory at end of shift; in the data, planned and unplanned downtime are indistinguishable. Two senior Yeses against two evidence-adjacent Noes, on a gate block: a **critical fault line**.

The facilitator runs the anchor question as a live demonstration before the session. Asked to reconstruct one specific stoppage from ten days prior, the IT/OT lead can produce the work order within minutes — but not the material lot, and nothing in the record says whether the stop was planned. The 2am Question fails on screen, in front of no one, which is exactly where it should fail first.

The arithmetic makes the design argument by itself. The total comes to **51 out of 60** — a score that, in any conventional framework, reads as a green light. The verdict is `Conditionally Grounded`: the data gate has failed, and no amount of excellence in the other nine blocks can average it back to health. The session produces two senior debt items — encode stoppage classification at source, and backfill a labelled event history: an estimated €80K and four months, with the IT/OT lead as named owner — and the business case is re-priced from €600K to €680K, with capital release staged behind the debt closure and a re-grounding booked for day 90.

> **What just didn't happen.** Without private intake, the approval meeting would have heard the CFO's Yes and the supervisor's silence. The €600K would have trained a model on data that cannot tell planned from unplanned downtime — a model that would have alerted on scheduled maintenance and missed real failures, been ignored by operators within a month, and surfaced eighteen months later in the post-mortem as "predictive maintenance doesn't work here." The fault line cost four weeks and €80K to fix. Its absence from the business case would have cost €600K and the organisation's appetite for the next attempt.

SECTION 11

Where to push back

A framework published for comment should name its own weak points. These are the objections we consider most serious, with our current answers — several of which are openly incomplete.

"Any self-assessment can be gamed. Why is this different?"
Named artifacts can be checked, and several are — the live demonstration, the calibration flag, the facilitator's spot-checks in session. Gaming Grounded requires a leadership team lying *in concert* about artifacts that can be asked for on screen, which is a far higher bar than one optimist ticking boxes. Residual truth: no diagnostic survives a coordinated deception. Grounded's claim is narrower — it raises the cost of the everyday, uncoordinated optimism that actually sinks these investments.

"This slows us down while competitors ship."
A full assessment takes roughly two weeks elapsed, mostly waiting on calendars, and a Conditionally Grounded verdict does not stop the investment — it prices the gap and stages the capital. The honest comparison is not two weeks versus zero; it is two weeks versus the eighteen-month detour through a failed deployment. The fastest route through is knowing what you are standing on.

"Why these ten blocks? Why these three gates?"
Derived from the recurring failure modes of three technology waves, not from first principles — which means they are empirical claims and should be treated as such. Field data may promote governance to a gate in regulated industries or merge blocks that always move together. We hold the specific blocks loosely and the method — evidence, private intake, gates-not-averages — tightly. Challenges to the block structure, with reasoning, are exactly the feedback we most want.

"The disagreement-predicts-failure claim is unproven."
Correct, and Section 08 says so in the framework's own text. It is a hypothesis with a plausible mechanism, awaiting field validation, and we will publish results either way. The private-intake design does not depend on it: suppressing HiPPO dynamics is sufficient justification on its own.

"A three-value verdict is too blunt for a nuanced decision."
The nuance lives one layer down — in the block scores, the fault-line report, and the priced debt register. The verdict is blunt on purpose, because board communication is a lossy channel and anything roundable gets rounded. Bluntness at the headline, precision underneath.

"Ninety days is arbitrary."
Within a range, yes. The defensible claim is categorical: readiness decays, and an assessment without an expiry becomes misleading. Ninety days is an opening calibration — long enough to act on, short enough that a deprecated model or a departed data owner is caught. We expect field use to adjust it.

"This is consultant-ware — a framework built to sell engagements."
The framework, the questions, the method and the scoring are open and free, and the whole assessment is runnable by one internal person with a spreadsheet — this note is sufficient instruction. If facilitation services grow around it, they compete on facilitation quality, not on access to the framework. An open standard that is also useful to practitioners is a feature, not a confession.

SECTION 12

Beyond AI — and an invitation

Strip the AI vocabulary out of the thirty questions and what remains is technology-agnostic: What outcome are we actually trying to change? Is the foundation ready? Whose behaviour must change, and why would it? What does failure cost, and who owns it? Does the economics survive scale? These are the questions ERP should have been asked in 1996 and cloud should have been asked in 2014. Grounded is therefore built as a general transformation framework with AI as its first — and currently most urgent — instantiation, because AI is where boards are committing capital today and where the mispricing is largest. The framework is designed to outlive the wave that launched it.

Editions. The framework ships in an industry-neutral edition and industry editions that translate every question into local operational language — the manufacturing edition speaks in shift boundaries, material traceability, MES and historian integration. Each edition carries its own anchor question for live demonstration; in manufacturing it is the 2am Question. All editions share the method — the Evidence Rule, the gates, private intake, fault lines, the verdict — so results are comparable across sites, industries and portfolios.

Upstream: the Scan. Grounded assumes a candidate investment exists — a sentence can be written, a number (or an honest "not yet costed") attached. Organisations that arrive earlier, asking *"where, if anywhere, should AI touch what we already have?"*, are served by a deliberately lighter precursor: the Grounded Scan, which runs Block 1's core question in breadth across the product or operation and returns a shortlist of candidates, each with a rough Investment Statement. The Scan finds where to look; only Grounded tests whether what you found is ready to fund. They are kept separate on purpose — bending the full framework's discipline to accommodate discovery would quietly destroy the discipline.

What Grounded is, and is not. It does not evaluate the quality of the AI technology, the capability of a vendor, or the brilliance of a solution — a technically excellent investment can be Ungrounded, and that is the point. Nor does a Grounded verdict guarantee success; it eliminates the most common and most preventable causes of failure, which is the most any diagnostic can honestly claim. It is open, citable, and versioned like a standard, because a readiness framework that enterprises must trust with capital decisions should itself be inspectable down to every question and every scoring rule.

THE INVITATION

This note is version 1.0, published to be argued with. Three requests, in ascending order of value to us: **run it** — one person, one live investment decision, one spreadsheet, and tell us where the method creaked. **Break it** — challenge the gates, the block structure, the ninety days, the disagreement hypothesis; the objections in Section 11 are a starting list, not a boundary. **Extend it** — translate an edition into your industry's language and its own anchor question, and publish it. The framework improves in public or it doesn't improve at all.

> *Know what you're building on before you build.*