OPENING EDITION · 2 OCTOBER 2026

Someone Said AI. Now What?

A practical guide for people responsible for using, introducing or overseeing AI at work.

Mani Padisetti

Almost Magic Tech Lab

Edition note

Opening Edition, 2 October 2026. This edition contains the Prologue, the Introduction, Chapters 1 to 4, the Epilogue and Appendices A to D. Chapters 5 to 12 are listed by working title and will be added in later editions.

This book is free. You do not need to sign up or buy anything to use it. The latest edition, the update history and the four interactive fill-in tools are at book.almostmagic.io.

Update history

2 October 2026: Opening Edition published.

For those asked to act before anyone has properly named the problem.

And for the people who will live with the decision.

Contents

PROLOGUE

The Friday Paper

The continuing case in this book is invented. Its organisation, people and events represent recurring situations rather than a disguised account of one company.

At 9.07 on Monday morning, Maya Sen was asked for the company’s AI strategy.

The request came near the end of the executive meeting, after the financial forecast and before a discussion about office leases. It occupied less than four minutes.

Tom, the chair, had attended an industry dinner the previous Thursday. Two people at his table had spoken about AI agents. One said his company expected large productivity gains. The other said organisations without an AI strategy would soon struggle to compete.

Tom placed his coffee beside the meeting papers.

“We need to get ahead of this,” he said. “What’s our AI strategy?”

Maya, the chief operating officer, looked towards Daniel, who led technology. Daniel looked back at her.

The company did not have an AI strategy. It did, however, have AI.

A sales manager had built a tool that could analyse tender documents. It had impressed his team until it attributed a customer requirement to the wrong section of a contract. A person caught the error before the proposal left the building. The incident had not been recorded because nobody was sure where it belonged.

Customer service wanted an assistant that could search old cases. The team had spent months asking for its knowledge base to be cleaned. Some procedures contradicted one another. Product information sat in shared drives, old tickets and the memories of two experienced employees.

A large customer had also sent an assurance questionnaire. One question asked whether the company maintained a register of AI systems and their approved uses. The response was due on Friday, alongside the paper Tom now wanted for the board.

“Could we start with the service desk?” the chief financial officer asked. “If AI handles the repetitive enquiries, there should be a saving.”

Leila, who ran customer operations, closed the folder in front of her.

“Which enquiries are repetitive?” she asked.

The chief financial officer glanced at his notes. “You tell us. You run the team.”

“We record the same problem under six different names,” Leila said. “The reports make six problems look like one large category and one problem look like six small ones. If the assistant learns from that, it will answer quickly. I’m less sure it will answer correctly.”

The room went quiet for a moment.

Daniel said he could arrange demonstrations from three providers. Tom nodded.

“Good. Let’s move. Maya, can you give us a recommendation for the board paper by Friday? We have a budget decision next week.”

By 9.16, Maya owned an AI strategy, a customer assurance response and a deadline tied to funding.

FIG-P.1 · SOMEONE SAID AI

One request creates five decisions

Strategy
What direction can we defend?

Assurance
What can we truthfully evidence?

Existing use
What needs a boundary now?

Service failure
What work needs diagnosis?

Budget
What spending can we authorise?

FIG-P.1 Separate strategy, assurance, existing use, service failure and budget before choosing technology. Source: author’s method and fictional case.

Back at her desk, she opened a new document and typed:

AI Strategy

She began with a list of possible uses. Customer support. Tender analysis. Internal search. Invoice processing. Marketing content.

It was the sort of list the board expected. It was also easy to produce. She could add a coloured matrix, rank the ideas by value and feasibility, and recommend two pilots. Daniel could obtain vendor estimates. Finance could attach a savings range.

The paper would look finished.

FIG-P.2 · SOMEONE SAID AI

Two possible board papers

Expected paper
Use-case matrix; vendor estimates; forecast savings. Assumptions can disappear inside the presentation.

Maya’s paper
Current exposure; decision owners; knowns and unknowns; bounded next steps. Investment direction still has to follow.

FIG-P.2 A finished-looking paper can still conceal the decisions it asks others to make. Source: author’s method and fictional case.

Her first interruption came from the sales director.

“The tender tool is still a trial,” he said from the doorway. “It hasn’t touched customer data.”

“Where did the tender documents come from?”

“The sales folder.”

“Those are customer documents.”

He leaned against the doorframe. “They’re commercial documents. We didn’t upload personal information.”

“Do we know what the provider retains?”

“Daniel’s team can check. The point is, it saved hours.”

“How many?”

“We haven’t measured it formally.”

He wanted the tool included in the board paper as evidence that the company had already begun. He also wanted the assurance response to avoid language that might alarm the customer.

“We caught the contract error,” he said. “That’s what review is for.”

After he left, Maya stared at the list on her screen. One line now contained a claimed saving with no baseline, a customer document used under uncertain terms, an error that had not been recorded and a reviewer who happened to notice it.

The line still read: Tender analysis.

At 10.42, Leila called.

“I sounded difficult in there,” she said.

“You sounded as though you knew something the rest of us didn’t.”

“I know the knowledge base is a mess. I also know my team is tired. If a decent assistant could find the right procedure, they’d use it tomorrow.”

“So you want the pilot?”

“I want the problem fixed. If the pilot helps, good. But if we put a fluent answer on top of contradictory instructions, customers will get the contradiction faster.”

Leila described a case from the previous week. A customer had called twice about a service credit. The first employee followed an old procedure and declined it. The second found the revised rule and approved it. The customer had spent forty minutes correcting a decision the company should never have made.

“Which version would the assistant have found?” Maya asked.

“That depends on what we feed it. Nobody has decided which sources are current.”

Maya saved the use-case list in a draft email to Tom. Her cursor rested over Send. It would meet the request she had been given and buy her time, but it would also turn five guesses into the company's apparent direction.

She closed the email without sending it.

She did not yet know whether the service desk needed AI. She knew the customer had already paid for the organisation's confusion with time and frustration. She knew Leila's team wanted relief, while finance wanted a saving it could put into next week's budget. She knew a software demonstration would not settle either matter.

The customer questionnaire asked the company to identify each AI system, its approved purpose and the person accountable for it. Maya could not answer any of the three questions for the tender tool. She borrowed the first question for a blank page:

What has happened?

The chair had requested an AI strategy. A budget decision was approaching. A customer wanted assurance. Employees were already experimenting. At least one experiment had used customer material and produced an error.

Below that, she wrote:

What decision is needed by Friday?

The company could not responsibly decide its entire AI strategy in four days. It could decide which existing uses needed immediate boundaries, which proposed uses deserved diagnosis, and who would own that work.

Maya called Daniel and Leila. The three of them agreed on a provisional first response.

Until the provider terms had been checked, the sales team would stop uploading new customer tenders to the trial tool. The existing material would not be deleted until Daniel had established what records were needed and what deletion would actually remove. Customer service would not begin a vendor pilot that week. Leila would nominate one service problem for a short diagnosis, starting with the conflicting service-credit procedures. Daniel would identify AI features already enabled in the company's main software.

These were temporary boundaries, not a complete inventory. Maya wrote that distinction into the paper.

At 4.36, she sent Tom a one-page note.

It did not contain a roadmap or a savings forecast. It recorded the request, the pressure behind it, what was known, what remained uncertain and three decisions the board could responsibly make on Friday.

Tom replied six minutes later.

I expected options. This reads like a list of questions.

Maya read the message twice. She reopened the opportunity matrix and wondered whether she had made the wrong call. On Friday she would have to defend a paper that did not resemble the one the chair had asked for.

INTRODUCTION

Before the Technology Becomes the Answer

It rarely sounds like a reckless request.

It may come from the chief executive, the board or a trusted provider. Employees may already be experimenting. However it arrives, the expectation is familiar:

“We need AI.”

The problem is usually less clear.

The person who inherits it is expected to create movement from very little clarity. Sometimes experimentation has already begun, so delay carries consequences as well as caution.

This book is for that person.

You may lead a business function, a technology team or an entire organisation. In a smaller company, you may hold several of those responsibilities at once. You have enough authority to organise the next decision, but you cannot create certainty that the evidence does not support.

The purpose of this book is to help you sequence the work. It begins with the demand, examines the problem beneath it, and follows the resulting decisions through experiment, design, production and eventual retirement. AI is one possible outcome of that journey. A repaired process, a better source of information or a decision to wait may be more valuable.

The first decision is smaller than an AI strategy

Maya’s board did not need a complete AI strategy by Friday. It needed an honest response to a demand that had already created pressure and exposure.

Her one-page note could not tell the company where every use of AI belonged. It could establish what had happened, who needed to act and what should be contained while the organisation learned more.

For a new consequential use outside existing permission, the first useful decision is often:

Does this demand deserve diagnosis, and what must be bounded while that diagnosis occurs?

For those consequential demands, there are three first decisions: decline; contain and clarify; or authorise a bounded diagnosis. Routine use within existing approval does not enter this gate. Wider lifecycle decisions follow when an experiment or production use is proposed.

A request may be declined when it has no credible outcome, owner or organisational reason to continue. It may return to its sponsor with better questions.

The answer may be yes, with prerequisites. The underlying work may be real while the data, process or ownership is unfit for an AI intervention.

An experiment may follow diagnosis. It should answer a specific question under conditions that protect people and information, with agreed limits and a clear decision at the end. The full experiment contract comes later in the book.

Later, the organisation may decide to enable a feature, configure an existing product, buy a service or build a capability. Each route still requires a clear purpose and evidence proportionate to what the system may influence or do.

Production brings a different set of decisions. The organisation has to determine whether the capability should continue, change, lose authority or be retired. Approval is never the last decision.

FIG-I.4 Approval begins a new set of responsibilities; it is never the last decision. Source: author's method and fictional case.

Most AI use does not need the full method

At this company, most everyday use covered by the existing approval sits in the first row below. That is the working boundary of this example, not a claim about how AI is used in every organisation. A new tool, purpose, information category or permission requires the boundary to be checked again.

Consider a finance analyst summarising her own internal meeting notes with an approved office feature. These particular notes contain no personal or customer information. The feature is approved for this information, she checks the summary, and it stays inside the company. Daniel records the use in the provisional register. She does not need a Demand Note or an independent challenger for each summary.

The service-credit assistant is different: its sources disagree and its output could shape customer answers. It needs a bounded experiment. The later claims system is different again because its decisions would affect customer money. The evidence and authority must match that consequence.

FIG-I.1 · SOMEONE SAID AI

Match the process to the proposed use

Use and decision	Requirements and stopping conditions
Routine use | proceed within permission	Approved summarising feature; analyst's own non-personal internal notes; checked output. Record the use, not every summary. Stop if information or audience exceeds approval.
Bounded experiment | obtain a new decision	Service-credit retrieval. Diagnose source conflicts; agree limits, challenge and an end decision. No customer contact or record changes before approval.
Consequential system | withhold live authority until justified	Small claims decisions affect customer money. Test material assumptions; establish challenge and correction; name a pause owner. The board decides the proposed authority.

FIG-I.1 Routine approved use can proceed; a new experiment needs a decision; consequential authority needs stronger evidence. Source: author's method and fictional case.

Proceed under existing authority when the actual use remains inside its approved purpose, information, audience and checking requirements.

Obtain a new decision when the purpose, information or authority changes, or the existing permission is unclear.

Stop the affected use when its boundary is crossed or a required protection is absent. For the analyst's example, that includes personal or customer information entering this narrowly approved use, or output leaving unchecked. It is not a universal ban on every authorised use of such information.

The rest of the book concentrates on the middle and bottom rows. Routine permission should be easy to use; consequential decisions need the fuller method.

Authority changes the question

AI conversations often begin with capability. Can the model summarise this document? Can it answer a customer? Can it rank these cases? Can an agent complete the task without help?

Capability does not settle permission.

This book uses five authority modes: Inform, Assist, Recommend, Determine and Act. The figure below shows the practical distinction.

The word determine distinguishes an automated output that takes effect as a decision from a recommendation that a person considers. Terminology varies across organisations and jurisdictions; accountability does not. Named people still have to approve the purpose, establish the limits and provide a route for correction or challenge.

These modes are not a universal risk ranking. A determination affecting employment or access to an essential service may carry greater consequence than a tightly bounded, reversible action. Assess the combination of data access, audience, reach, autonomy, consequence and reversibility. One system may occupy several modes.

Additional authority requires evidence that matches the decision. The evidence falls into four broad families: the reality of the problem and comparison with simpler options; technical and operational performance in the intended workflow; human consequences, review and correction; and organisational fitness, including privacy, security, permissibility, cost, support and recovery. A high accuracy result in a test set cannot establish all four.

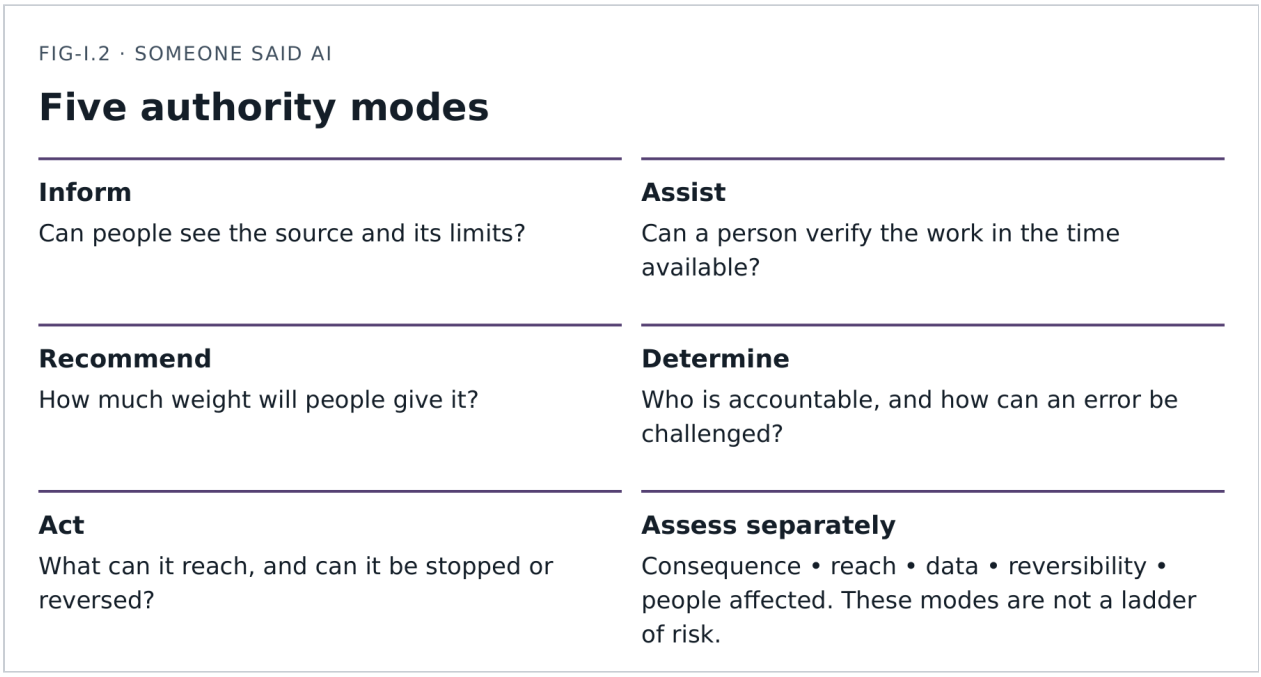


FIG-I.2 Authority describes permission; consequence, reach, data and reversibility determine the evidence needed. Source: author’s method and fictional case.

A model sits inside operating arrangements

The service assistant proposed in Maya’s meeting would not operate alone.

It would depend on source material that somebody had to approve and maintain. Employees would need to know when to trust an answer and when to check it. Customers would need a way to correct an error. Managers would need to understand whether the assistant reduced repeat contact or simply made first replies faster. Someone would have to respond when a provider changed the service.

A capable model can depend on poor operating arrangements.

The operating arrangements around an AI capability include the work, people, information, permissions, controls, suppliers and support that allow it to function. They also include the less visible parts: who can pause it, who pays for exceptions, which records are preserved and what happens when the original owner leaves.

These responsibilities exist in a small organisation too. One person may hold several roles, but the roles still need to be named. Where consequences are material, the person seeking approval should not be the only person challenging the evidence. The challenger should sit outside the expected benefit or delivery target and have enough knowledge to test the claim. Legal, safety or rights-sensitive uses may still require specialist advice.

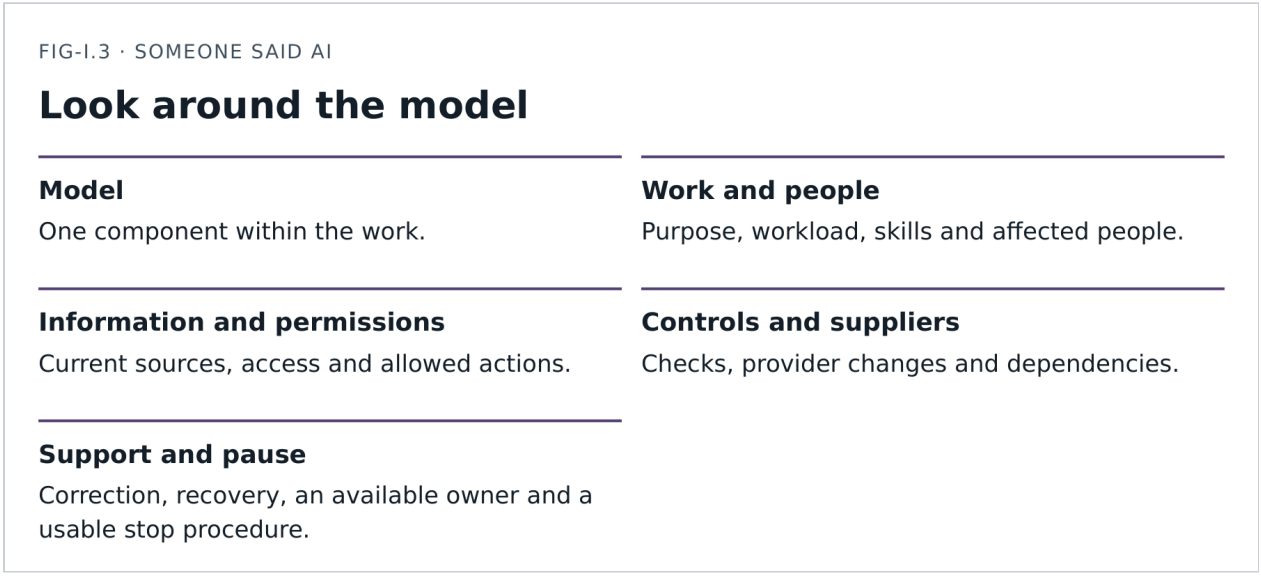


FIG-I.3 The model depends on people, information, permissions and operating responsibilities. Source: author's method and fictional case.

You can begin without buying another platform. Use documents, spreadsheets and shared systems your organisation already approves, with appropriate access, retention, confidentiality and backup arrangements. The discipline comes from naming decisions, owners, evidence and review dates.

What the journey will test

The chapters test the full operating cost, human review under real workload and whether people can correct an error. Chapter 5 supplies the experiment contract. Chapters 8 and 9 examine performance, incidents and recovery. Chapter 10 follows the arithmetic from baseline through experiment to live use. Chapter 11 shows how authority is renewed, restricted or retired; Chapter 12 returns to the strategy Tom requested.

How to use this book

You can read the chapters in sequence. They follow the path from an initial demand to the decision to continue, restrict or retire a capability.

If you already face a live issue, enter where the work is:

- A new consequential request has arrived: use Chapter 1 and the first-pass Demand Note. For routine use within existing permission, use the proportionality check in this Introduction.
- Employees or providers already use AI: start with Chapter 1 to record activity and apply necessary boundaries; Chapter 6 develops the risk screen.
- A pilot appears successful: Chapters 5 and 7 examine whether its evidence supports production authority.
- A capability is operating: Chapters 8 to 11 address outcomes, changed assumptions, recovery, cost and renewed authority.

A free companion website launches with this Opening Edition on 2 October 2026. It carries dated updates and the interactive fill-in tools, and will later add jurisdiction-specific guidance. The templates included here work without it. No website tool is required to use this edition.

If your demand requires a new decision, bring it with you. Preserve the words in which it was expressed. If you are uncertain whether existing approval covers the use, ask its owner to clarify that boundary first.

Write down who said it, what created the urgency and what they expect to change.

Chapter 1 begins with that record.

CHAPTER 1

The Demand Arrives

At 8.31 on Tuesday morning, Maya's phone lit up.

The board will expect a sense of direction. Questions are fine, but we need to show we are not standing still.

Tom wanted options.

By then, Maya had received suggestions from almost every member of the executive team. Finance proposed invoice matching. Sales wanted the tender tool restored. Marketing wanted access to a paid writing assistant. Daniel had booked two vendor demonstrations for Thursday afternoon.

Leila sent no new use cases. She sent the two service-credit procedures.

One was dated 11 June 2025, fourteen months earlier, and stored in the customer-service knowledge base. The other had been approved six months later, on 11 December 2025, and attached to an email. Both looked current. Neither contained a clear supersession notice.

Maya understood the concern. The company could not spend months studying every possibility while employees found their own tools and providers introduced features by default. Delay was not neutral.

But the long list of ideas concealed a more immediate fact: people were already using AI under different assumptions, and nobody had agreed what counted as authorised use.

She called Daniel.

“Can you cancel the demonstrations?”

“I can,” he said. “I don’t think we should.”

“Why?”

“Because the board wants movement, and we need to learn what the products can do. A demonstration doesn’t commit us to anything.”

“What will we ask them to demonstrate?”

There was a pause.

“Service support. Tender review. Internal knowledge.”

“Those are three different problems.”

“They are also three common products. We don’t need a finished business case before we look.”

Daniel was right about one thing. Seeing a product could help the team understand what was possible. The risk was allowing the product to define the problem. A polished demonstration could turn an undefined demand into a preferred solution before anyone had agreed what success meant.

“Keep one session,” Maya said. “Ask them to use a fictional scenario. No company or customer data. Tell them we are examining the service-credit problem, not shopping for an enterprise platform.”

“That will make the meeting very narrow. The provider has one opening this month, and procurement needs a price range before the budget meeting. If we cancel the wider session, we may lose both.”

Maya considered that. “Keep the opening. Ask for a price range separately, but make the demonstration about the service-credit problem. We will learn less about their platform and more about whether they can work with our evidence.”

“Tom may still call that slow.”

“Then the paper needs to explain the trade.”

Record why the request arrived now

An AI request is rarely only a request for technology. It carries a history.

The chair may fear that competitors are moving faster. A manager may need to reduce a queue. Employees may already be using tools because the approved process is slow. A vendor may be encouraging adoption before a renewal. A customer may be asking for assurance. An executive may have publicly promised savings.

These pressures explain why the demand has arrived now and what may distort the decision. They belong in the record.

If a sponsor has already promised a result, evidence that challenges the proposal may be unwelcome. If a team is exhausted, it may accept risks it would otherwise question. If a customer has asked for assurance, the organisation may feel tempted to describe intentions as existing controls.

The first task is to preserve the request before organisational language tidies it up.

Write down what was said as closely as you can. “We need an AI strategy” is different from “customers wait four days for a response” and different again from “reduce service cost by fifteen per cent”. Each statement carries an assumption that should remain visible.

Do not convert the request into a solution brief yet.

One request can carry five different pressures

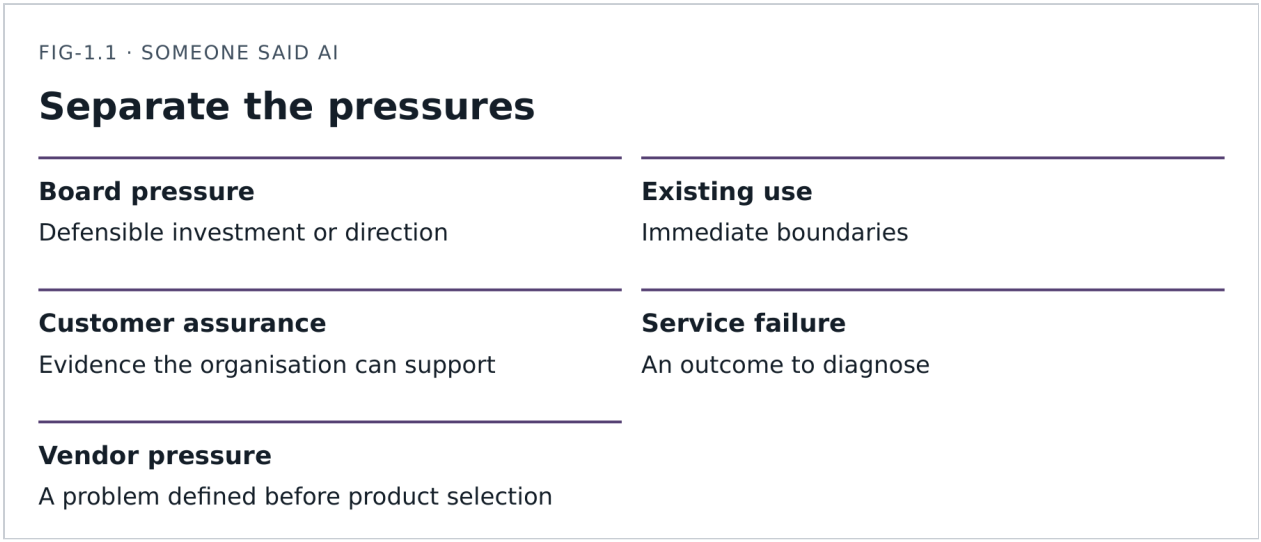


FIG-1.1 Different pressures create different decisions even when they arrive in one request. Source: author's method and fictional case.

Write the first useful note

At 4.10 on Monday 10 August, before sending the note described in the Prologue, Maya drew five headings on a blank page. She gave herself twenty minutes. Several answers were still missing. She recorded who would find them rather than filling the space with a confident guess.

What is being requested in the original words

Tom's words: "We need to get ahead of this. What's our AI strategy?"

What should improve and who owns it

Direction and assurance for the board. Maya owns this first record. Service and tender baselines: unknown. Leila and the sales director to clarify their intended outcomes by Friday 14 August.

What is already happening

Sales reports a tender trial and a caught error. Other use: unknown. Daniel to bring an initial list and supporting records by Friday 14 August.

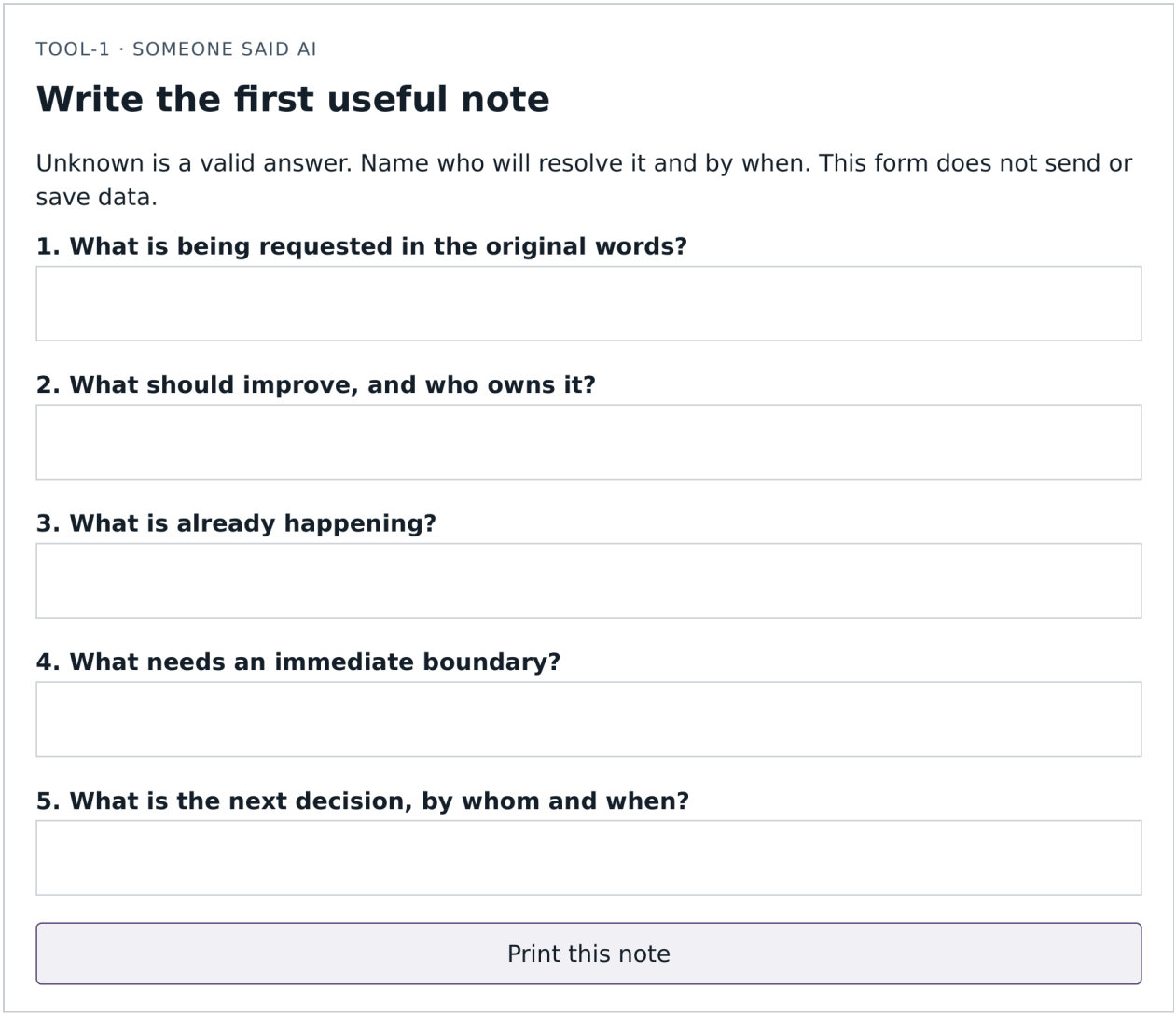
What needs an immediate boundary

No new customer tender uploads while handling is checked; Daniel applies the pause. No live data in demonstrations. No customer-facing service pilot yet. Maya reviews the first restrictions on Friday.

What is the next decision by whom and when

Tom and the board on Friday 14 August: decide what to contain and whether to authorise diagnosis. Implementation spending is not requested.

At 4.30 she stopped editing. At 4.36 she sent the page to Tom. The unknowns had owners and dates. They were work to do, not defects to hide.



TOOL-1 Answer five questions first; add supporting detail only where needed. Source: author's method and fictional case.

Layer 1 stays short. Layer 2 holds the records and questions that a particular demand needs. By Friday, Maya could update the first page and attach the supporting detail below. She had not known all of it on Monday.

Read the Friday first pass

What is being requested in the original words

"We need to get ahead of this. What's our AI strategy?" Tom Avery, chair. First recorded Monday 10 August 2026; updated Friday 14 August.

What should improve and who owns it

Tom needs an investment direction and a truthful assurance response. Maya owns the record. Leila owns the service outcome; the sales director owns tender work. Baselines remain unknown: each owner reports by Friday 28 August.

What is already happening

Tender use and a caught contract-reference error are evidenced in the saved trial history and review copy. Leila holds two conflicting procedures. Four of seven reported tools have evidence-backed entries; three reports and AI features in two existing systems remain unchecked.

What needs an immediate boundary

No new tender uploads to the trial; Daniel enforces the restriction. Demonstrations use fictional data. No customer-facing service pilot or experimental record changes. Maya decides exceptions after the required checks; reviews are dated in Layer 2.

What is the next decision by whom and when

Friday 14 August: Tom and the board decide on bounded diagnosis and interim controls. Requested end date: Friday 28 August. Combined ceiling: twelve staff-days and AUD 5,000 external advice. No implementation approval is requested.

Keep the supporting detail behind it

Layer 2 for this demand follows. These are fictional case records, included to demonstrate the method. A real note would link to the organisation’s access-controlled originals.

Record pressures and measures

Tom needs capital-allocation options and an honest customer response. Sales needs a route for next week’s tender. Leila needs fewer repeat contacts and reversals. No measured productivity or cash saving is yet established. The opportunity remains open while the exposure is contained.

Identify people and correction

Priya and other service advisers who would use, review and correct output under existing workload targets.

The customer who spent forty minutes obtaining correction of a service-credit decision, and customers in comparable cases.

Customers whose tenders, service records or communications may be processed.

Managers accountable for service quality, cost and correction.

Voice, correction and escalation during diagnosis:

Priya will represent frontline evidence in the service workflow session.

Existing complaints, reversals and repeat-contact records will be reviewed as provisional customer evidence.

Leila owns correction of live service records during diagnosis and escalation of any material customer impact to Maya.

Direct customer participation will be considered after the first case review; if it is not used, the reason will be recorded.

Name challenge and specialist input

Nadia Bell, head of internal assurance, will test whether evidence supports the proposed next step before either diagnosis closes.

Nadia does not own the expected service benefit, sales target or technical delivery. She will record any conflict that emerges.

The initial screen has identified customer confidentiality, provider handling, uncertain information content and contractual obligations in the tender trial. Daniel will obtain security and privacy review, and the commercial counsel will review the relevant customer and provider terms, before Maya decides whether any tender use may resume.

Record each boundary

Tender trial pause. No new customer tenders may be uploaded. Daniel owns enforcement; Maya may vary or lift the pause after provider terms, settings, retention, deletion and customer confidentiality obligations are checked. The sales director may request an exception in writing, but cannot approve it. Review: Friday, 14 August 2026.

Demonstration data. No live company or customer information may be used. Daniel owns enforcement and may stop the session. Any exception requires Maya’s prior approval after information classification and provider handling are checked. Review: after the Thursday demonstration.

Customer-service pilot hold. No customer-facing pilot may begin until the service-credit workflow, current sources, correction route and outcome owner are confirmed. Leila owns the hold; Maya approves any move to experiment after independent challenge. Review: Friday, 28 August 2026.

The board retains authority to approve the experiment budget and scope. Maya may implement that approval within its conditions; she cannot authorise an additional experiment or spending overrun herself. This distinction also applies when she varies a temporary boundary.

Attach sources to known facts

Board paper and assurance response due Friday. Source: Monday executive meeting record and the customer’s dated questionnaire, held by Maya. Checked against the originals on Friday 14 August.

Tender tool used and an incorrect reference caught. Source: saved trial history, uploaded tender and marked review copy, compared by Daniel on Thursday 13 August. Establishes this recorded incident, not the overall error rate.

Two service-credit procedures conflict. Source: knowledge-base article dated 11 June 2025 and approved email attachment dated 11 December 2025, held and compared by Leila on Tuesday 11 August.

A reversal is not explained in the live contact history. Source: the service-credit record inspected by Leila with Priya on Wednesday 12 August. The credit entry exists; the reason for reversal is absent.

Keep reports assumptions and unknowns distinct

Reported saving: sales says the trial saved hours; source is the sales director’s account. Baseline not measured. Working assumption: faster replies reduce total service cost; finance must test review and repeat-contact work. Permission and handling for resumed tender use remain unresolved despite the check of the retained document set.

By Friday, the original “no personal information” report has become a checked statement about the retained trial set only. Other tools remain outside that check. Daniel owns the remaining tool and feature checks for Friday 21 August. Leila and the sales director own their baselines and alternatives for Friday 28 August. Maya records which matters remain open at each decision.

The proposal asks for diagnoses within twelve staff-days and AUD 5,000 for external advice. Their conclusions must distinguish closing the request, repairing prerequisites and proposing a bounded comparison. This note is not an approved AI strategy or a complete inventory.

In a small organisation, one person may hold several of these responsibilities. Keep the responsibilities separate on the page even when the names repeat. Where the decision could materially affect a person, a contract, safety or confidential information, find a challenger outside the proposed benefit and obtain specialist advice where needed.

Keep evidence in its proper place

FIG-1.2 · SOMEONE SAID AI

Keep evidence in its proper place

Evidence class	Example and source
Verified fact	Conflicting procedures | Source: Leila’s knowledge-base copy dated 11 June 2025 and approved email attachment dated 11 December 2025.
Reported information	Hours saved | Source: sales director’s account; no measured baseline yet.
Assumption	Faster replies reduce total cost | Source: finance’s working proposition; test checking and repeat-contact work.
Unanswered question	Other enabled AI features | Source: Daniel’s discovery list; checks in two systems still outstanding.

FIG-1.2 State what each claim rests on; a report or assumption is not a verified fact. Source: author’s method and fictional case.

One request, several demands

On Tuesday afternoon, Maya met with Tom.

He read the first-pass Demand Note in silence. Maya kept the supporting records beside it.

“This is useful,” he said, “but it still doesn’t tell the board where we should invest.”

“We don’t have enough evidence to recommend an investment.”

“Then what do we have?”

“A customer assurance problem, an existing-use problem and at least one operational problem. They need different responses.”

Tom turned back to the first page.

“The board asked for a strategy.”

“The words were ‘What’s our AI strategy?’ I recorded them exactly.”

“You know what I meant.”

“I may have misunderstood part of it,” Maya said. “I heard a request for an investment direction. The notes suggest a customer-assurance issue, uncontrolled adoption and a service problem as well. Which of those does the board need resolved first?”

“Capital allocation,” Tom said. “And confidence that unmanaged use is not growing while we deliberate. I need to explain why we are reserving money or why we are not.”

“Then I can give you a bounded discovery request and the controls we need now. The investment options will follow the evidence.”

Tom tapped the table with his pen.

“What can the board approve?”

“A short diagnosis. Immediate boundaries. Named owners. We can return with options supported by evidence.”

“How long?”

“Two weeks for the first two demands.”

“Why two?”

“Because that is enough time to map one service problem and verify the tender trial without pretending we have completed an enterprise review.”

“And the budget?”

“Reserve a small discovery amount. Do not approve implementation money yet.”

“Give me a page I can defend,” he said eventually. “No governance essay.”

Name the owner of the outcome

A sponsor can create urgency and provide authority. That does not make the sponsor the owner of every outcome.

If the proposed use concerns customer service, the owner should understand the service process and be accountable for what improves or worsens. Technology can operate the system and own its technical components. Leila remains accountable for service quality and customer consequences. Risk, privacy and security specialists challenge the proposed approach within their expertise.

Benefits also need owners.

“Save time” is not an owned outcome. Whose time? How much? What will happen with it? Will customers receive faster service, will the team handle more work, or will the organisation reduce cost? Each answer changes the baseline and the people who should participate.

Maya asked Leila to own the diagnosis of the service-credit problem.

Leila hesitated.

“If I own it, finance will expect a saving.”

“Finance can test the economics. You own whether the service outcome improves.”

“What outcome?”

They agreed on two initial measures: the number of contacts required to resolve a service-credit request and the rate at which employees reversed an earlier answer because they found a different procedure.

Neither measure mentioned AI.

That mattered. If cleaning the procedures solved most of the problem, the diagnosis needed to make that visible.

See who lives with the result

The people affected by a consequential AI use extend beyond its sponsor and users. In this case, customers could receive the wrong answer while advisers inherited the work of checking and correcting it.

The first-pass note need not contain a complete impact assessment. Where the request raises consequences for people, use Layer 2 to identify who may experience them and how they can raise a problem during diagnosis.

Leila asked Priya, one of the senior service advisers, to join the first workflow session. Priya had handled the customer whose service credit was initially declined.

“The old procedure wasn’t the only problem,” Priya said. “The account screen shows the date but not the reason for the last decision. I thought another adviser had already checked eligibility.”

“Would an assistant have helped?” Maya asked.

“If it showed me the current rule and the earlier adviser’s reasoning, yes. If it wrote a confident reply without either, no.”

The distinction changed the proposed work. The team had begun with “an assistant for repetitive enquiries”. It now had a narrower question about current rules and decision memory.

Priya also asked who would correct the customer record if an AI-assisted answer was wrong. Leila checked the earlier case. The credit had been applied, but the original refusal remained in the contact history without a note explaining the reversal. A future adviser could still interpret it as a valid decision.

Leila assigned herself responsibility for correcting live service records during the diagnosis. Maya added the voice, correction and escalation details to Layer 2 of the Demand Note.

Establish provisional boundaries

Diagnosis can take time. Existing exposure may require action before the diagnosis is complete.

Provisional boundaries are temporary rules that prevent a demand or experiment from expanding while the organisation establishes what is happening. They should be narrow enough to enforce and short enough to review.

For this case, three boundaries mattered most: keep customer documents out of the unapproved tender tool; use fictional data in the demonstration; and prevent experimental systems from contacting customers or changing records. The supporting prompts are in Appendix A.

A boundary is not an enduring policy merely because it is sensible. Record its owner, reason and review date. Explain how employees can ask for an exception or report an existing use without being punished for surfacing it.

FIG-1.4 · SOMEONE SAID AI

Make a boundary usable

Boundary and reason

No new customer tenders in the unapproved trial; provider handling and obligations remain unresolved.

Owner and enforcer

Daniel applies and checks the restriction.

Request and approval

Sales director may request an exception; Maya decides after the relevant checks.

Review and lifting

Review on Friday 14 August 2026. Record evidence, decision and any continuing restriction.

FIG-1.4 A boundary needs enforcement, an exception route and a review date. Source: author’s method and fictional case.

Maya wanted employees to tell Daniel what they were already using. She made clear that reporting an existing use would not itself attract a penalty. Deliberate misuse would still be handled through the company’s ordinary processes.

In this invented case, Daniel checked the saved trial history against the uploaded documents. He and the commercial counsel found no personal information in that retained set, but the documents contained customer commercial information subject to confidentiality terms. The provider’s written reply showed different deletion schedules for the account and service logs. Daniel recorded the scope of the check; it did not establish what employees might have entered into other tools.

The company had not approved this use or confirmed that the provider terms and customer confidentiality obligations permitted it.

Daniel disabled the trial account after preserving the available history and obtaining confirmation of the provider’s deletion process. The sales director objected.

“You’re shutting down the only real example we have,” he said. “We have another tender next week. The team will lose the time saving, and the board will see a closed account instead of progress.”

“I’ll put that cost in the paper,” Maya said. “I’ll also put in the customer obligation we had not checked. Bring me the baseline and we can test another route.”

Decide whether diagnosis is deserved

Routine use within an existing approval does not need a new project. This triage is for proposed consequential uses outside that approval, or uncertain current use that needs clarification.

First check current or uncertain exposure on every route, even if you intend to decline the proposal. Containment can remain necessary after the proposal ends. Then ask:

1. Is there a real outcome, failure or opportunity to examine?

2. Is someone accountable for that outcome?

3. Is there a credible initial map of affected people and a plan to find indirect or missing groups, including people who may struggle to raise concerns?

4. Does the underlying problem deserve examination independently of AI?

5. Is it worth comparing process repair, existing software, deterministic automation and AI?

6. Can the organisation establish safe boundaries, time limits and meaningful independent challenge for diagnosis?

Run an initial specialist screen before authorising diagnosis where the proposed use may affect rights, employment, safety, essential services, vulnerable people, personal information, security, contractual obligations or regulated decisions. Seek advice proportionate to the context; do not wait for a polished evidence pack if an immediate restriction may be needed.

For demands entering this gate, the result should be one of three first decisions. Necessary containment remains active whichever decision is made.

Decline

The request has no clear outcome, owner or strategic reason to continue. Record why it was declined and what would need to change before it could return.

Contain and clarify

Existing use or provider functionality creates exposure, but the underlying demand is not ready for diagnosis. Establish boundaries, identify ownership and collect the missing facts.

Diagnose

A real problem or opportunity exists, an owner is named and a short examination can produce a useful decision. Approve the scope, time limit and evidence required.

At this stage, “diagnose” is not approval to buy, build or pilot.

The first decision gate

FIG-1.3 Declining a proposal never removes the need to handle existing exposure. Source: author's method and fictional case.

What the board received

On Friday morning, Maya replaced the requested AI strategy with a paper titled:

AI Demand: First Decisions

The paper contained three recommendations.

First, the board would approve provisional boundaries for unapproved or uncertain AI use. Approved routine use could continue within its existing information and audience limits. A new purpose or data category required review of the relevant provider terms, settings and obligations. Customer communications or record changes required an approved workflow with verification, correction and monitoring. A named person had to be able to pause it.

Second, the company would run two diagnoses ending Friday, 28 August. Leila would own the service-credit problem and return with a current-work baseline, source conflicts, customer consequences and process-repair options. The sales director would own a review of tender-analysis work, including baseline time, types of error, document obligations and non-AI alternatives. The two-week limit belonged to this case; future work would be sized by scope, consequence, evidence access and participation needs.

Third, Daniel would maintain a provisional register. He had checked four of seven reported tools against account or configuration records and recorded each approved purpose or unresolved permission. Three tool reports and AI features in two existing systems remained unchecked. Those two counts were separate:

reports of tools and checks inside established systems. Each outstanding check had Daniel as owner and Friday 21 August as its next review date.

The paper included no enterprise platform purchase and no headcount-saving target. The two diagnoses shared a ceiling of twelve staff-days and AUD 5,000 for external specialist advice. Maya could move effort between the two diagnoses but could not approve an overrun. Any additional discovery spending, experiment or implementation required a separate board decision.

Tom presented it to the board.

The first question came from a non-executive director.

“Are we being too cautious?”

Tom looked towards Maya.

“A customer document went into a trial before we reviewed the provider terms,” she said. “The service assistant would also draw from two procedures that give different answers. I cannot recommend expanding either use until we resolve those facts.”

The chief financial officer asked when the board would receive numbers.

“After we establish a baseline and identify where the work goes,” Maya said. “If time moves from advisers to reviewers, we will show that. If a process repair creates most of the benefit, we will show that too.”

The board approved the boundaries and the service-credit diagnosis. It authorised the tender review only far enough to establish the document obligations, baseline and alternatives; any new trial would return for approval. One director dissented, arguing that the company should run a low-cost productivity pilot immediately. The request for an enterprise AI investment returned to the agenda without a decision.

Nadia read the assurance response before it left. “Four entries supported by records is not a verified inventory,” she said. The response stated exactly what Daniel had checked, listed the three unverified reports and the two systems awaiting feature checks, and described the interim restrictions actually applied. It made no claim to a complete AI management system.

The tender trial remained closed.

Leila left the meeting with responsibility for a problem she genuinely owned and no promise that AI would solve it. Priya had a place in the diagnosis. Finance had to wait for a number it could defend.

After the meeting, Leila opened the two service-credit procedures beside the earlier customer record. The later procedure looked clearer, but it still did not explain who could approve an exception or what an adviser should record after reversing a decision. Her two-week clock had begun.

Create your AI Demand Note

Use the five-question first pass in Appendix A for a consequential demand needing a new decision. Preserve the original words. Allow unknowns, but assign each an owner and a date.

Add Layer 2 only for the questions the first pass exposes. Keep the decision page short and the supporting evidence accessible. Do not finish with “explore AI opportunities”. Name the next decision and who can make it.

Chapter 2 follows Leila’s diagnosis. It asks whether the conflicting procedures explain the customer’s experience or hide a deeper failure in recording and correcting decisions.

CHAPTER 2

The Work Beneath the Request

At 8.15 on Monday 17 August, Leila Morgan moved the service queue to one side of her screen and opened twenty completed service-credit cases. Priya had forty minutes before she was due back on the phones. The board had allowed two weeks for the diagnosis. Leila could already see how the investigation might consume the people it was meant to help.

They had taken the last twenty cases completed in July from the case register, rather than choosing memorable complaints. Each had a full seven-day follow-up window. Beside the register sat the contact history, the credit ledger and both versions of the procedure. The old article was dated 11 June 2025. The approved replacement had arrived by email on 11 December. Neither told an adviser what to do with the other.

“Shall we time how long it takes to find the rule?” Leila asked.

“First tell me what we’re trying to find,” Priya said.

She opened a case in which a credit had eventually been paid. The first adviser had declined it. The second had approved it. The ledger showed the money. The contact history still showed the original refusal, followed by a note saying only that the customer had called again.

Leila recognised the pattern from the case they had discussed the previous week. She had thought it was an untidy exception. Now another record was on the screen.

Priya put a finger against the second contact. “If this comes back to me, which decision am I meant to believe?”

The answer was not in either procedure. They could search faster and still arrive at an incomplete account of what the company had already done.

Follow the work through one complete case

They stopped the search timer. On a blank page, Priya wrote what had actually happened: request received, rule found, eligibility judged, refusal recorded, customer called again, different rule used, credit paid. Underneath, she added the missing step: explain the changed decision in the record that the next adviser would read.

The written procedure ended at the credit transaction. The customer’s experience did not. A later adviser could repeat the old refusal because the first decision remained easier to see than its correction. The company had completed a payment without completing its memory of the decision.

Leila asked whether they should delete the refusal. Priya shook her head. Then they would lose the reason for the customer’s second call. They needed a visible correction linked to the original entry, with the person, time, reason and applicable rule. History should show what changed, rather than pretend the first decision never happened.

They followed the next case from start to finish. This time both advisers used the current rule. The first recorded “outside policy”. The second found an exception that a supervisor had approved. Nothing on the first screen revealed that approval. The customer had to supply the fact that the company itself had failed to carry forward.

By the sixth case, Leila had crossed out “advisers cannot find procedures” as the diagnosis. She left it on the page, legible, and wrote underneath: “Some advisers cannot establish which decision is current or why.” The earlier wording mattered. It explained why a search assistant had seemed like the obvious solution.

Mapping the work begins with an instance, not an ideal flow. Follow the request until the person affected can rely on the outcome. Include the return visit, the handover and the correction. Ask which screen the next person actually opens. A process that ends when one team completes its task may leave another team, or the customer, to repair the join.

For each step, record the input used, the judgement made and the record left behind. Notice where someone supplies knowledge from memory. Then ask what happens when that person is away. Do not turn every informal workaround into a defect. Some are sensible responses to an incomplete system. Find what the workaround protects before removing it.

Priya’s handwritten sequence was enough for this decision. There was no need to model the whole service department. The boundary was service-credit requests and the information required to resolve them. When a case raised another problem, Leila recorded it for its owner instead of expanding the diagnosis without limit.

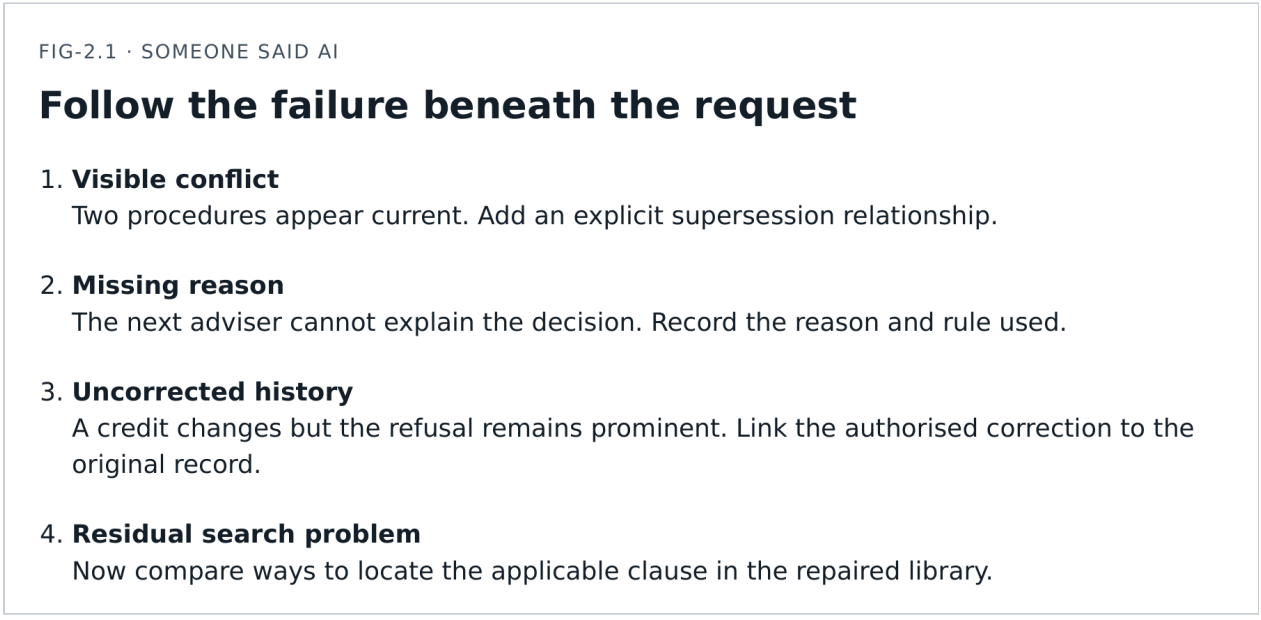


FIG-2.1 Faster search alone cannot repair a missing decision history. Source: Author’s method and fictional case.

Count the failure without pretending to know its cause

By 11.30, they had reconciled the twenty cases. There had been forty-eight contacts in total. Thirteen cases involved more than one contact. In ten of those thirteen, the reviewers found a missing reason or an unresolved procedure version that plausibly explained an avoidable return. They assigned one principal explanation per case: six to missing decision reasons, four to unresolved versions. Other contributing problems stayed in the notes.

Two of the twenty cases contained a reversal of the initial eligibility decision. A second contact was not automatically a reversal. One customer had called to check when an already approved credit would appear. Another had supplied information that had not been available at the first call. Leila would not label either a failed first decision merely to make the proposed repair look useful.

These were fictional company records, not research findings. The counts described this review set. They did not establish that ten cases would certainly have avoided a return if a field had existed. Someone might have entered an empty explanation in a mandatory box. A customer might still have wanted to speak to a person.

Priya kept a column for disagreement. Where the two reviewers could not reconstruct why the customer returned, they wrote “reason unclear”. They did not ask a language model to infer the missing reason and then treat the inference as history. A plausible reconstruction would have made the sheet tidier while making the evidence weaker.

At 2.20 that afternoon, Nadia read Leila’s proposed conclusion: “Most repeat contacts are caused by missing reasons and conflicting procedures.”

“You’ve found a likely mechanism,” she said. “You haven’t established the cause of most contacts.”

Leila pointed to the ten cases.

“Ten of thirteen in this set,” Nadia said. “And we should still test whether the repair changes what happens.”

The claim did not survive unchanged. Leila narrowed it to what the records supported and proposed a short observation of the repaired process. She had enough evidence to correct an obvious record defect under her existing authority. She did not yet have enough to promise the resulting saving.

A useful diagnosis separates three questions: what happened, what might explain it, and what changes when the suspected condition is repaired. Record-based observation can answer the first. It can suggest the second. The third needs further observation, and sometimes a comparison designed to distinguish competing explanations. The degree of investigation should match the decision being made.

For this repair, the team did not need to prove that every repeat contact had the same cause. They needed to establish that the procedure conflict was real, that the correction record was incomplete, and that the proposed fix would not erase legitimate exceptions. The saving would be measured rather than used as permission to fix a known defect.

Make the baseline answer the question

The twenty cases gave the team a mechanism to examine. They needed a wider record for the financial and service baseline. On Tuesday 18 August at 9.10, finance brought the July extract. It contained 1,200 service-credit requests and 2,880 linked contacts within the agreed observation window. The average was 2.4 contacts per request. There were 132 recorded reversals of the initial eligibility decision, or 11 per cent.

The observation window was seven days after the initial recorded resolution. A request was counted once, even if it had several contacts. A reversal meant an eligibility decision changed after that initial decision; it did not include a customer supplying a genuinely new request. The extract was a reconstruction of the July cohort with follow-up complete, not a simple count of everything that entered the contact centre during July.

Finance had previously used a different number. It divided all service contacts by tickets closed in the same month. That mixed requests that started earlier with requests still open at month end. The old measure remained useful for workload planning, but it could not tell them how often one request returned. They kept it, renamed it, and stopped comparing the two ratios as if they were the same.

Leila sampled the links between request IDs, contact records and credit entries. Priya checked cases where an adviser had opened a new ticket for the same problem. Finance retained the reconciliation and the exclusions. They could now reproduce the denominator instead of relying on a dashboard label.

The time estimate was less firm. The available activity records suggested around twelve minutes of adviser handling per contact, including the recorded after-call work. A practical planning range was eleven to thirteen minutes. The team had not measured every interruption or every supervisor consultation. They recorded those limits beside the estimate.

At the central estimate, 2,880 contacts multiplied by twelve minutes gave 34,560 minutes, or 576 adviser hours for the monthly cohort. The lower and upper time assumptions gave 528 to 624 hours. Those were handling estimates, not a claim that the company could remove that many paid hours from its roster.

The CFO asked which figure would appear in the board paper.

“All three,” Leila said. “The middle one explains the calculation. The range tells you how much we know.”

He agreed, then asked for the source and the repeat-contact definition to remain attached. He did not want the apparent precision of 576 to travel into next year’s budget without its assumptions.

Choose a baseline that could change the decision. Contacts per request helped this team see whether work returned. Reversal rate showed one kind of correction. Handling time helped estimate capacity. None of them alone measured whether customers understood the outcome, whether an initially wrong decision went unchallenged, or whether advisers could maintain the process under pressure.

Keep those gaps visible. A low reversal rate can mean better first decisions, but a case review is still needed to understand what the number represents here. If customers stop challenging an outcome, the recorded reversal rate could fall without service improving. The team therefore kept a small review of

completed cases and invited customer evidence instead of declaring the metric self-explanatory.

Avoid making the baseline larger than the decision. Leila did not commission a company-wide time study. She asked finance to preserve the extract, define the measures and follow the repaired workflow. The unanswered questions had owners. The diagnosis still had an end date.

Preserve the cases that do not fit

Finance found requests that could not be reconciled cleanly because two tickets appeared to concern the same event. Those records were investigated before the July cohort was finalised. The reconciliation file retained the link decisions and their reasons, so a later analyst could challenge them without rebuilding the whole extract.

Leila also kept a separate list of unresolved requests outside the completed-case review. A measure built only from finished work could otherwise make a growing unresolved queue disappear. That list was not added to the denominator of a completed-case ratio. It was shown beside the ratio, with the oldest unresolved request and its owner available for follow-up.

The board did not need every line of this working file. It did need to know that the twenty reviewed cases were not the entire service, that requests could be mislinked, and that unfinished work was still being watched. These qualifications could change the interpretation of the result. They were therefore part of the decision, rather than detail removed to make the paper shorter.

For another organisation, the appropriate observation window might be longer than seven days. The team chose it here to make the initial comparison feasible and consistent. It left later returns as an explicit blind spot. Leila assigned continued review of those returns to her service reporting routine. If later contacts changed the picture, the baseline and the claimed improvement would both be revised.

Repair what already has an owner

At 10.40 on Tuesday, Leila approved three changes within her existing service authority. The old procedure acquired a prominent supersession notice linking to the approved version. The current procedure acquired an effective date, an owner and a record of what it replaced. Changing a service-credit decision now required a reason and a link to the earlier decision in the contact record.

These changes were made in the existing knowledge base and case system. There was no new model and no new external data transfer. Daniel checked the configuration with Priya before it was enabled. A corrected decision remained visible at the top of the case, while the original entry stayed available in its history.

The reason field accepted a short explanation. It also allowed “awaiting supervisor decision” with a named owner, so an adviser would not invent a reason merely to get past the screen. An unresolved item could not be displayed as a final approval. Leila remained accountable for the procedure library, and assigned a service supervisor time to maintain this portion of it.

Priya tried the repaired screen with the case they had opened on Monday. She could see the refusal, the correction, the current rule and the reason for the change without opening the credit ledger separately. She asked Daniel to move the correction above the free-text contact history. The first configuration had put it below several old notes, where an adviser in a hurry might still miss it.

“Now I can tell the next person what happened,” she said.

That was the immediate test. They were not asking whether the screen looked modern or whether the company had adopted AI. They were asking whether the next adviser could reconstruct the decision without requiring the customer to do it again.

The old cases were not silently rewritten. Leila set a queue for reviewing incomplete records when they resurfaced, and for identifying records that needed proactive correction. Priya flagged possible customer consequences separately. The small repair did not provide authority to change an earlier credit outcome without checking the case.

Process repair should not wait for the technology proposal when the repair is already justified and authorised. Equally, “ordinary process work” is not permission to make any change casually. Keep the owner, the reason and the effect visible. Here, recording a corrected decision was within Leila’s remit. Changing the eligibility rule itself would have required the rule owner’s decision.

The board would hear about the repair on 28 August. It did not have to approve a supersession notice before the service team could stop presenting two contradictory procedures as current. The distinction preserved both speed and accountability.

Leila explained the reason field to the advisers before the change reached their queue. It was there so the next person could reconstruct the decision, not so a supervisor could count words. Priya suggested reviewing a few explanations together rather than setting a minimum length. A short reference to an applicable exception could be more useful than a long account that never said why the decision changed.

They kept a way for advisers to report an unsuitable field or missing option. If the new screen forced people to misdescribe the work, that would be evidence against the configuration. The form was part of the repair and could itself need repair.

Hear what the measures leave out

On Wednesday 19 August at 12.15, Leila joined a short call with three customers who had agreed to discuss service-credit experiences. The invitation made clear that this was optional and would not affect an individual decision. The team offered a written response and a separate call for anyone who did not want to speak in a group.

They did not begin with a demonstration. Leila showed a plain explanation of a corrected decision, with names and account details replaced by fictional information. One participant read it twice. She could see that the refusal had changed, but not whether she had to call again to obtain the credit.

Priya added a sentence saying what would happen next and when the customer should contact the company if it did not. Another participant wanted a way to challenge the explanation without retelling the whole story. The existing reference number was not visible on the example. It was added beside the contact route.

The woman who had questioned the next step agreed to return for future discussions. Leila proposed a standing customer panel, with other participants rotating as different services were examined. It would meet again before wider service changes. Its membership and limits would be recorded. One cooperative customer, or three, could not speak for everyone who found the process difficult.

The panel’s early contribution concerned the human correction process. It said nothing about accepting an automated refusal, or about a system deciding eligibility without a person. Leila wrote that boundary in the meeting note. Much later, someone would be tempted to use “customers consulted” as a broader claim than this conversation supported.

Customer evidence was also gathered outside the panel. Complaints, corrected records and people who had abandoned a request might reveal different experiences. Leila asked the service team to note where the invitation method excluded people, including anyone unable to use the offered channels. They did not treat silence as agreement.

Use participation to improve a decision, not to decorate it. State which question people were asked and what changed because of their answer. If the same meeting is cited later, check whether the later decision is actually the one they discussed. The existence of a panel is not evidence that every proposed use has been understood or accepted.

Let the simpler change earn its result

At 3.30 on Thursday 27 August, Priya brought the follow-up sheet. The team had tracked the first 100 eligible service-credit requests initially resolved in the repaired workflow between 18 and 20 August. Each now had the same seven-day window used for the baseline. The set involved ordinary requests handled by the participating service group; unusual cases and other teams were listed as limits, not quietly counted as successes.

Those requests had generated 180 contacts, or 1.8 per request. Four initial eligibility decisions had been reversed. The source was the linked case and contact extract, reconciled to the correction record. Priya and finance had checked the classification of each reversal. None of the cases had used an AI retrieval feature.

The shift was large enough to matter and too early to call permanent. They had changed several things together: the procedure notices, the reason field and where corrections appeared. Staff also knew the workflow was being watched. The result supported continuing the repair. It did not isolate the contribution of each change or prove that the same rates would hold across the full service group.

Leila’s first draft said “AI readiness work reduces repeat contacts”. She deleted the first two words. The improvement belonged to the process repair. A later tool should have to improve on this new starting point, rather than inherit credit for a change completed before it was used.

At unchanged volume and twelve minutes per contact, 1.8 contacts across 1,200 requests implied 432 handling hours. That was 144 fewer than the old central estimate of 576. Using the eleven-to-thirteen-minute range gave 396 to 468 hours. These were volume-normalised projections from the early cohort, not hours already removed from payroll.

The reversal change from 11 per cent to 4 per cent was also provisional. The team would continue the same definition and review the cases behind the rate. If the mix changed, the comparison would be revisited. The CFO accepted the potential capacity gain but crossed out “saving” until the use of the released time was known.

The repair had removed the dominant failure mechanism found in the initial review and delivered a substantial early improvement without AI. The remaining technology question was correspondingly smaller. Nobody could now justify the original service-assistant proposal by pointing to all the old repeat work.

Keep the residual problem in view

The repaired process still had a slow part. During the observation period, Priya and another adviser logged clause searches across the much larger procedure library. The version conflict for service credits had been resolved, but customers described the same condition in different words. An adviser who knew the library could find the clause. Someone covering another product often needed help.

Priya showed Leila a case where the decision history was complete and the current rule was unambiguous. Finding the relevant exception had still taken several searches. A supervisor then located it using the library’s internal product term, a phrase the customer had never used.

This was a different problem from the one that had produced the repeat contacts. It was about locating applicable text, not deciding eligibility or repairing a record. The team did not claim a new baseline for search time from a few observed instances. They kept the examples and proposed to measure the residual task directly.

The demonstration Daniel had arranged on 13 August had shown a way to display a source clause beside an answer. Its limitations were now more useful than its sales promise. Leila wanted to compare a source-finding feature with the existing search, using the repaired library. A fluent customer reply was no longer part of the immediate question.

Write the residual problem after the repair, not before. Otherwise the proposed technology can continue solving the original presentation long after the organisation has discovered a different cause. A smaller problem may need a smaller product, or no new product at all.

Give each alternative the same problem

Leila wrote four routes against the remaining work. Process repair would continue: ownership, supersession and correction were necessary whichever technology they chose. Better use of the existing software might improve navigation and search. Deterministic automation could reject publication without

required fields and direct known product codes to known clauses. AI retrieval might help when ordinary language and library terminology differed.

They did not score “AI” against “do nothing”. They compared each route against a defined task: help an adviser locate the applicable clause and its effective date, without changing the customer record or deciding eligibility. The routes could be combined. A required effective-date field was useful even if the retrieval feature never proceeded.

FIG-2.2 · SOMEONE SAID AI

Compare routes against the diagnosed work

Route	Work it addresses	Limit to preserve
Process repair	Current procedures, decision reasons, visible corrections.	Does not remove every clause search.
Existing software	Navigation, saved searches and available features.	Existing ownership does not establish permission for a new use.
Deterministic automation	Require metadata; route known codes to known clauses.	Someone must maintain rules and exceptions.
AI retrieval	Find candidate clauses from varied wording.	Applicability and source checking remain necessary.

FIG-2.2 Keep process repair in every route; test AI only against the residual search problem. Source: Author’s method and fictional case.

A simple route should not win merely because it is familiar. If deterministic rules required a growing list of exceptions that nobody could maintain, that burden belonged in the comparison. Equally, a new product should not win because its demonstration included more functions than the team needed. Compare the work required to maintain each answer, including what happens when the procedure changes.

The short comparison would use the same repaired source set, the same questions and the same definition of a useful result. Advisers would need to find and inspect the source, not simply obtain an answer. Cases where no applicable rule existed would remain in the set. A tool that always produced something should not receive credit for inventing certainty in those cases.

Take the tender owner seriously

At 11.00 on Thursday 27 August, the sales director put a separate sheet in front of Maya. It covered six completed tenders from before the trial. He had reconstructed requirement-mapping and checking time from work records and saved revisions. The six totals were eight, nine, ten, eleven, eleven and eleven hours: sixty hours altogether, an average of ten.

He excluded waiting for customer clarification and commercial negotiation, because the proposed assistance would not remove those activities. Three mapping defects appeared in the saved drafts: two omitted requirements and one incorrect clause link. Review had caught them before submission. They were defects per six reviewed tenders, not a claim that half of all future tenders would fail.

Nadia asked whether the time included checking. He showed the reconciliation. It did. That claim held. The earlier claim that the trial “saved hours” remained reported information: no comparable timed run existed for the trial, and the paused account was not reopened to manufacture one.

“I still think we should have another route,” he said. “But I can’t give you the saving yet.”

Maya had expected an argument about lost momentum. He had brought a better account of the work than the one in her first paper. He also wanted a clause check because he did not want a plausible summary to conceal an omitted obligation.

The tender diagnosis would proceed towards reviewed terms and an approved tool. It would not establish that the disabled original account could resume. Daniel and commercial counsel still had to determine which documents could be processed, where, for what purpose and under which retention conditions. The sales director accepted that boundary and asked for a decision date. Maya put 4 September against it.

Close the diagnosis with a decision

At 9.20 on Friday 28 August, Tom asked which of the two diagnoses should receive money. Maya began with what had already improved. The service repair was operating under Leila’s authority. Its early result justified continuing it and measuring it across the wider workflow. No AI purchase was required to keep that benefit.

She asked the board to authorise a bounded comparison of source retrieval against existing search. It would use the repaired procedure set and fictional questions, with no live customer contact or record changes. The board also allowed a focused demonstration of the configured option within that comparison. The vendor session of 13 August had already occurred; this permission did not retrospectively authorise it or create a second version of the same event.

The board allowed four staff-days and AUD 1,500 of external assistance for this comparison, ending Friday 11 September. These were new limits after diagnosis. They were not an extension silently absorbed into the original ceiling. The completed diagnoses had used ten and a half of the twelve authorised staff-days and AUD 3,800 of the AUD 5,000 advice allowance, according to the work log and invoices.

The dissenting director wanted an immediate service rollout. He was right that customers should not wait for a technology experiment to receive the record repair. Leila confirmed that they were not waiting. The retrieval question remained open because it would introduce a different dependency and had not yet shown a benefit over the repaired process.

Tender assistance received a route to a decision, not permission to restart. Reviewed customer and provider terms, a usable citation check and an accountable reviewer were the remaining conditions. Maya would decide within the board’s stated delegation once Daniel and counsel had supplied the evidence.

Tom asked what would happen if ordinary search won. “We use it,” Leila said. The answer finally sounded like a service decision.

Complete the Diagnosis Record

The worked record at the end of the paper fitted on a page. Its question was whether service-credit repeat work required an AI assistant. Its evidence was the twenty-case review, the reconciled July cohort and the hundred-request observation of the repaired workflow. Its conclusion separated the proven record defects from the provisional estimate of improvement.

The chosen route was to keep the repair and compare source retrieval only. Leila owned the service outcome; Daniel owned the technical comparison; Nadia would challenge the interpretation. The baseline remained 2.4 contacts and 11 per cent reversals in the July cohort. The early repaired-workflow observation was 1.8 contacts and 4 per cent reversals, with scope and follow-up limits attached.

The record named what remained unknown: search benefit over the repaired process, performance on less common products, and whether the early service improvement would persist. It named what could proceed now: current procedures, visible correction and the bounded comparison. Tender resumption remained conditional on the separate route decision.

Use Appendix B to make the same distinction in your own work. Record the original question, the observed failure, the sources and limits, the simpler change tried, what happened afterwards and the residual question. End with close, repair prerequisites or proceed to a bounded comparison. If the answer is to repair, name who can act now. A diagnosis has earned its time when it changes the work that follows.

CHAPTER 3

The Demonstration

At 2.00 on Thursday 13 August, four days before the case review began, Daniel checked the shared folder one last time. It held a fictional customer, a fictional service interruption and two procedure extracts copied into a test pack with identifying details removed. Nothing in it described a live customer. The supplier had agreed to use that pack instead of connecting to company systems.

This was the demonstration booked in Chapter 1. Maya had nearly cancelled it. Daniel had argued that they could learn from seeing the product without committing to buy it. As he opened the meeting, the service diagnosis and its later repair were still ahead of them.

The presenter entered the fictional customer’s question. A polished answer appeared. It recognised the service interruption, expressed regret and explained that the customer was not eligible for a credit. The sentences were clear. The answer sounded like something an experienced adviser might send after checking the policy.

Priya looked at the test pack. “That’s the old rule.”

The presenter said the system could be instructed to prefer current documents. Daniel asked him to leave the answer on the screen. Before changing the instruction, they wanted to see what it had used.

A small reference opened the older procedure. Daniel leaned forward. “Can you keep that visible beside the answer?”

The presenter could. Priya then asked the question that would shape the later design: “Does it show me which rule it used and when it changed?”

Separate producing an answer from establishing it

The product had done something useful and something wrong in the same response. It had turned a short question into readable text and offered a route to the source. It had also answered from a superseded procedure. Calling the whole demonstration a success would hide the error. Calling it useless would discard the source display that had made the error easier to find.

Daniel asked the presenter to expand the reference. The clause matched the answer. That established a narrower fact than anyone first assumed. The output was supported by a document in the pack, but the document was not the applicable rule. A valid citation to an obsolete instruction remained the wrong basis for the customer’s decision.

Leila wrote three questions separately: did the output follow the cited clause; was that clause current for this case; and did the complete customer record justify using it? The demonstration could help examine the first two. Its fictional customer could not establish the third for any real case.

This distinction would matter after the diagnosis. The immediate design idea became a visible clause and effective date, with an adviser deciding applicability. The team had started the meeting expecting to see an assistant compose replies. They left interested in a narrower capability that helped a person inspect the evidence.

Generative text can be fluent while containing false statements or failing to follow its source. A confident tone is part of the output, not proof that the underlying claim has been checked. Research on generated language examines this problem across tasks, including summarisation and question answering.[1] In this demonstration, the team did not have to debate how often that happened elsewhere. They had a saved answer and the wrong procedure beside it.

Ask what the system is producing

On the shared whiteboard, Daniel drew two company examples. One was a proposed forecast of next week’s service demand from earlier volumes and calendar information. Its output would be a number, perhaps accompanied by a range. The other was the demonstration’s written response to a customer’s question. Its output was newly composed text.

He used “predictive” for the first purpose and “generative” for the second. These were practical distinctions for this conversation, not sealed categories. A language generator also predicts the next part of its output. The important question was what the organisation would use the result to do.[2]

If the forecast were wrong, Leila might roster too few advisers. If the written explanation were wrong, a customer might receive an incorrect account of entitlement. Both required evaluation, but an attractive paragraph could not validate a staffing forecast, and a good forecast could not validate the paragraph.

The CFO, joining for the commercial part of the call, asked whether the supplier’s “AI platform” covered both. The presenter said it could support several applications. Daniel brought the discussion back to the proposed task. What evidence existed for finding a source clause in their procedure library? A broad product category did not answer that question.

For your own proposal, complete one ordinary sentence: the system will produce this output for this person to use in this decision. If the sentence contains several decisions, separate them. A system that forecasts demand, drafts replies and approves credits has several jobs, even if one interface presents them together.

For the fictional company, the demand forecast remained an example on a whiteboard. Nobody added it to the register as a deployed forecasting tool or counted its benefit in the service case. No one had authorised it to run. Discussing a possibility did not mean the company had deployed it.

Notice where variation enters

The presenter repeated the same question. The second answer was shorter and used different wording. It still relied on the old procedure. Priya asked whether an adviser could assume the answer would always be the same after the pack was corrected.

“No,” Daniel said. “We have to specify and test the behaviour we need.”

Language generation involves selecting successive pieces of text from model-produced possibilities. Generation settings can change how those selections are made; sampling can produce different wording on repeated runs.[3] A repeated prompt can also reach different source material if the library or retrieval configuration changes. Keep the model, settings and source version in the test record instead of treating the visible question as the whole test.

The presenter offered to make the output more consistent. Leila asked what consistency would prove. If a system repeated the same wrong rule every time, the service problem would remain. If it offered several correct phrasings of the same applicable clause, some variation might be harmless. They needed to distinguish variation in expression from variation in the decision-relevant content.

They saved both outputs. Priya marked the rule used, the date displayed and whether the answer omitted the exception. Those were the features they would later compare. A screenshot of one good run would not reveal whether another run selected an incompatible rule.

Do not insist that every AI output be identical merely because ordinary software sometimes is. Decide what must remain stable for the task. For this use, the cited source identity, applicability information and absence of unauthorised action mattered more than whether the explanation began with “under the procedure” or “according to the procedure”.

Where reproducibility matters, retain enough of the test context to investigate a difference. Do not assume a provider can reconstruct it later from a vague description of what the adviser remembers seeing. That is a practical record requirement, not a promise of perfect repeatability.

Follow the sources through the system

The presenter described the product as retrieval-augmented generation. Daniel translated the phrase: it found passages in a connected collection, then supplied them to a language model to help compose the answer. That combination has a research basis, but connecting retrieval does not by itself establish that every returned answer is current or suitable for a particular decision.[4]

The test pack made the stages visible. Both procedures entered the collection. A retrieval step selected material. The model produced a response using what it received. The interface displayed an answer and a reference. The organisation still had to decide which procedure applied and whether the reference supported the claim being made.

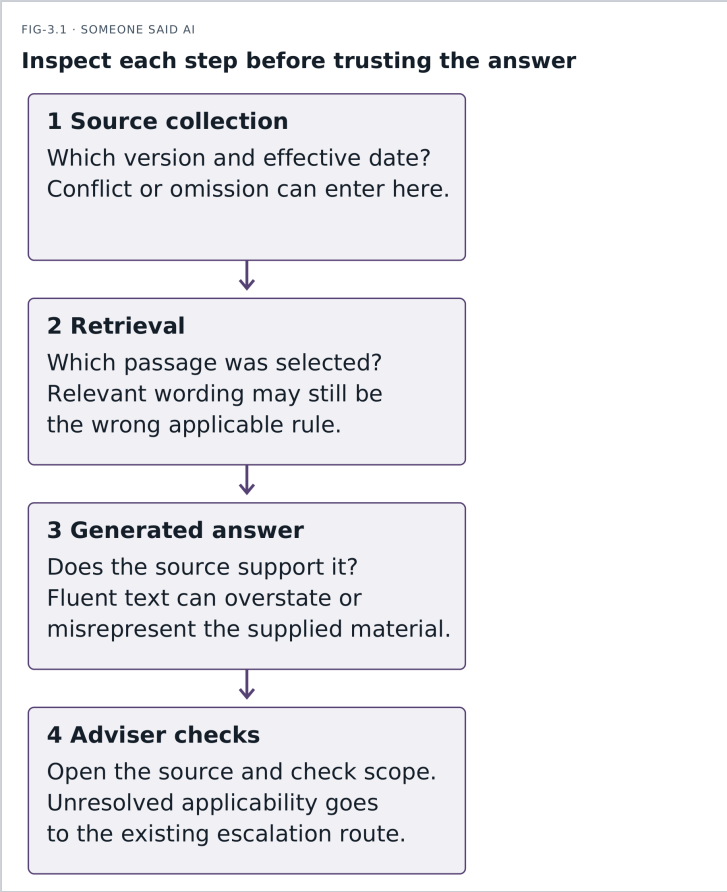


FIG-3.1 A citation can point to a real document that is wrong for the decision. Source: Fictional case and author’s design questions; retrieval mechanism: Lewis et al. (2020), reference R21.

Priya asked the presenter to remove the older procedure from the active test collection. On the next run, the system used the newer clause. That was a useful observation about this configuration. It was not evidence that the product would identify superseded instructions in the company’s full library without someone maintaining their status.

Leila asked what happened when a document had a new upload date but an old effective date. The presenter paused. The product could display supplied metadata, but someone had to define which field meant what. The newest file was not necessarily the rule that governed an earlier event.

In their example, a customer’s entitlement could depend on the date of the service interruption. The applicable procedure might therefore be an earlier version retained for a legitimate reason. Removing every old file from view would make current search look cleaner while making historical decisions harder to explain.

The eventual requirement was more precise than “use the latest document”. The system should reveal the source, its effective period and its relationship to other versions. If applicability could not be established, the adviser should see the uncertainty rather than a silently chosen winner. The procedure owner, not the model’s prose, would determine which rule was authoritative.

The team also separated publication from indexing. A corrected document in the library did not prove that the retrieval collection had already been refreshed. Daniel asked for a way to check which version the configured feature actually returned. That check would become part of maintaining the service, rather than a one-time purchase question.

These were requirements generated by the fictional test, not claims that the supplier’s product already met them. Daniel placed each beside an observed result or an unanswered question. The distinction prevented the requirements list from becoming an accidental assurance statement.

Make an accuracy figure describe its test

The presenter moved to a slide reporting 96 per cent accuracy on a supplier test set. The figure belongs to this invented demonstration; it is not a statistic about a real provider. Maya asked what had been counted as correct.

A response counted as correct if the evaluator judged that it answered the supplied question using the reference answer. Daniel asked whether the set included conflicting procedures, missing exceptions or a question for which no answer existed. The presenter did not have that breakdown in the meeting. He agreed to provide the test description rather than improvise one.

Nadia would later classify the 96 per cent as a reported claim. It could help them formulate questions. It could not be copied into the board paper as the company’s expected performance. The denominator, task mix, judging method, product version and consequences of the remaining errors were still unknown to them.

Priya asked whether a correct refusal to answer counted as a success. For her work, it might be the safest and most useful response when the source did not settle the question. A score that rewarded answering everything could favour behaviour the company did not want.

The supplier was not wrong to measure its product. The team was asking whether that measurement addressed their decision. A benchmark may demonstrate capability on its stated tasks. The company still needs evidence for its own material conditions and for the work surrounding the output. Success on one task need not transfer to a similar-looking task.[5]

Use the demonstration to find what should enter the later evaluation. Include ordinary work, important exceptions, absent information and cases where the appropriate answer is to seek help. Separate errors that are inconvenient from errors that change a person’s entitlement. Do not average them into a single reassuring number before deciding how the system may be used.

Distinguish a tested score from a persuasive sentence

The presenter showed another feature: a confidence indicator. It was described as a score for the result. The CFO asked whether the company could automatically accept results above a threshold and send the rest to a person.

Daniel asked what the score measured. Was it similarity between a question and a passage, a classifier's estimate, or another calculation? The label did not tell them. A retrieval similarity score could be useful for ranking passages without being a probability that an eligibility decision was correct.

A tested decision score, as this book uses the term, is a numerical output whose relationship to a specified outcome has been evaluated on relevant cases. "Tested" does not mean infallible or automatically transferable. The record must say what was scored, against which outcomes, on which cases and for which proposed action.

Research on classification models has found that their confidence estimates can be poorly calibrated: the number reported need not match the observed likelihood of correctness.[6] That finding concerns numerical model estimates. It does not make a generated sentence such as "I am very confident" a measured probability.

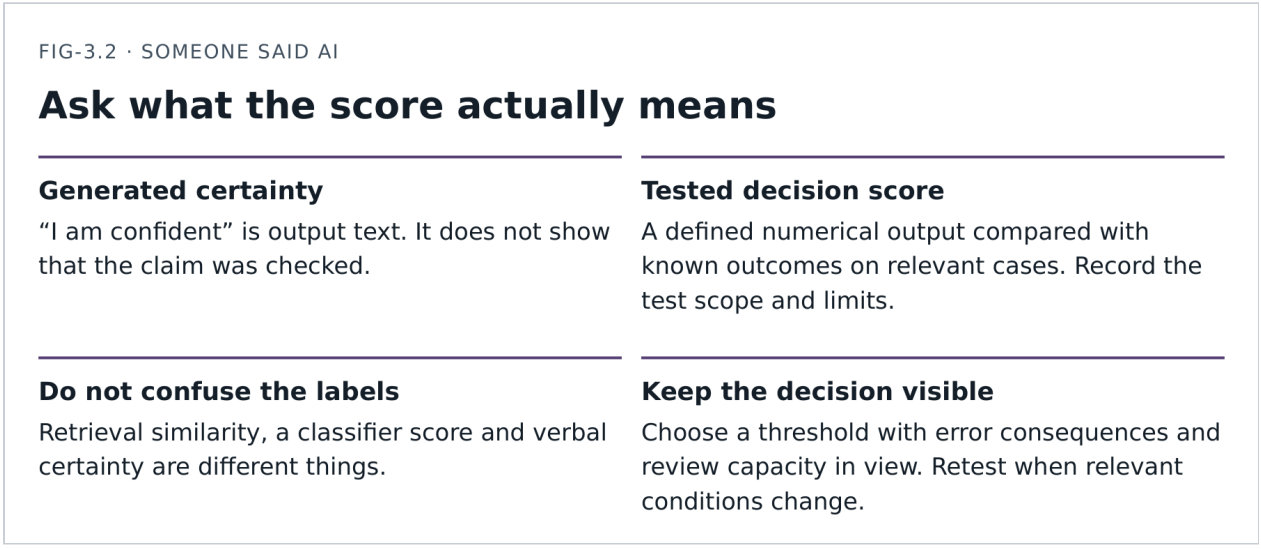


FIG-3.2 A confident sentence is not a measured probability; a numerical score needs task-specific evaluation. Source: Author's explanation; calibration distinction: Guo et al. (2017), reference R15.

Daniel offered a separate illustration. Suppose a classifier assigned scores between zero and one to a defined prediction. Group its predictions by score and compare them with known outcomes on cases not used to fit the model. If a group near 0.8 is correct only about half the time, the number should not be presented as an 80 per cent chance of correctness. This is an illustrative explanation, not a result from the company's tool.

Even a well-calibrated score would leave a decision to make. A threshold trades different mistakes and different amounts of work for people. A credit wrongly withheld may matter differently from a case unnecessarily sent for review. The organisation has to judge those consequences and the capacity of the review route, rather than ask a number to make that judgement on its own.

Maya wrote "no automatic eligibility decision" beside the demonstration record. They had not defined or evaluated a score for that use. The immediate retrieval question did not require one. An adviser could inspect a clause without the company granting the product authority to approve or decline money.

Chapter 8 will return to score behaviour and monitoring. The point here is sufficient for a buying decision: ask what the number means and where that meaning was established. If the answer is missing, preserve the number as an unexplained output, not a control.

Trace what an agent could reach

The next screen offered an agent workflow. The presenter showed a fictional sequence that could read a case, draft a response and call a function to update a record. The company’s test environment had no live connections, so the demonstration could not alter an actual customer account.

An agent can use a model within a sequence that selects and invokes tools, observes results and continues working. Research systems have demonstrated such combinations of model-generated reasoning and actions with external interfaces.[7] The business consequence depends on the tools and permissions supplied. A text-only suggestion and a connected record update therefore need different operating boundaries.

Leila asked what happened if the first interpretation was wrong and the workflow continued. Daniel drew the proposed connections: procedure library, customer record, messaging service and credit transaction. He marked where the system would read, where it could write and where a result would leave the company.

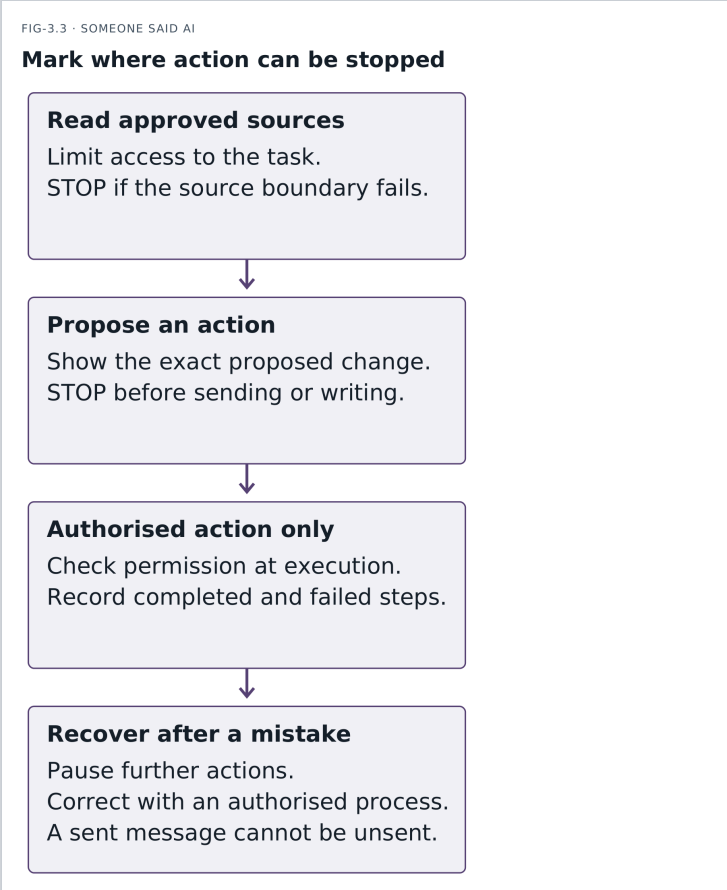


FIG-3.3 Stop before the consequential action; reversal afterwards needs its own authority and may be incomplete. Source: Author’s proposed boundaries; model and tool interaction: Yao et al. (2023), reference R22.

For this project, there was no reason to connect all four. Reading an approved procedure collection could answer the immediate question. The customer system and payment function added authority without helping establish whether source retrieval was useful.

The presenter said those connections could remain disabled. Daniel asked how they would verify that in the selected configuration and who could enable them later. An assurance that the team “would not use” a connection was weaker than a configuration in which it was unavailable to the account doing the work.

They also distinguished stopping from reversing. Preventing the next action would not undo a message already sent. Correcting a record might require preserving the mistaken entry and adding an authorised correction. Reversing a credit could itself harm a customer and require a separate decision. “Undo” on a product slide did not settle those responsibilities.

If a later proposal introduced spending, the team would need to specify the amount, purpose, permitted recipients and person who could stop further commitments. For this comparison, the simpler rule was no purchasing connection and no spending authority. A capability they did not need was not included merely because it came with the platform.

An agent’s action history would also need to be understandable to someone investigating a failure. The team would want to distinguish a proposed action, an approved action, a completed action and an attempt that had failed. That requirement came from their own need to reconstruct a decision. A long stream of generated reasoning would not replace the record of what actually happened.

Check the software you already have

As the meeting ended, Leila asked whether they were discussing a new product because the company lacked the capability or because this supplier had offered the first demonstration. Daniel opened the existing software inventory. Its knowledge-base system had a retrieval feature that had not yet been assessed for this purpose.

He did not enable it in the meeting. Being included in an existing subscription would not establish permission for customer data or prove that its settings matched the proposed use. But it belonged in the route comparison. The demonstration had taught them what to ask of it: source visibility, effective dates and a way to decline an unsupported answer.

This was separate from the approved summarising feature used by the finance analyst in the Introduction. Her permitted activity could continue. There was no reason to bring every checked internal summary into the new project because another feature needed investigation.

The provisional register still contained reports awaiting verification. The outstanding feature checks in two existing systems were not magically completed by noticing an option on a screen. Daniel added the knowledge-base feature to the discovery record with its status: identified, configuration and terms not yet checked for service retrieval. He would reconcile it against the existing outstanding items rather than increase the count without checking overlap.

Inventory work should describe actual features, settings, purposes and permissions. A product name alone cannot distinguish an approved use from a new connection. Conversely, a new marketing label should not make a familiar, authorised use appear to require an entirely new governance project when its actual scope has not changed.

Save the demonstration as evidence of limited things

At 3.25, Maya asked Daniel to send a short record rather than the supplier’s slide deck. He kept the original wrong answer, the second run, the source extracts and the altered test-pack version. The supplier’s proposed requirements and its reported accuracy figure were recorded separately.

The observed facts were narrow: this configuration generated a fluent answer from the older procedure; it could expose the cited clause; removing the old procedure from the active pack changed the source used on the observed rerun. These facts had sources in the saved session record. They did not prove performance across the live library.

Reported information included the 96 per cent claim and the supplier’s account of features not exercised. The working assumption was that seeing the clause and date could make an adviser’s check easier. That assumption deserved a comparison, not immediate treatment as a benefit. Unanswered questions included applicability, permissions, update behaviour, failure handling and performance against existing search.

Daniel had been right to retain the session. The error had taught them more than a clean sales demonstration would have. It exposed an information requirement, and the source display offered a practical design idea. Maya wrote that in the record too. A colleague should not have to be entirely right about the product to be right about the value of learning from it.

Before closing the file, Priya checked the session notes against the saved outputs. One sentence said the product had found the wrong rule because it could not understand dates. The observed run did not establish that cause. The old rule might have been retrieved because of the test configuration, metadata or wording. They replaced the sentence with the observed behaviour and left the cause open for investigation.

That correction protected both sides. It prevented a weak explanation from becoming a permanent judgement about the supplier, and it stopped the team assuming that a date instruction alone would fix the problem. A demonstration can reveal a failure without revealing every reason for it. Preserve that difference when turning observations into requirements.

Complete the Demonstration Question Sheet

The worked sheet named the task as locating the applicable service-credit clause for an adviser. Its test material was the fictional pack used on 13 August. Its requested outputs were a clause, its identity, effective-date information and an explicit indication when applicability remained unresolved. No live records, customer messages or financial actions were permitted.

The sheet asked the supplier to show the source and the wrong case before explaining the headline score. It asked what changed between runs, which functions could act outside the interface, and what remained untested. Each answer received an evidence class and a reference to the saved observation or supplier document.

The next step was to carry the source-display idea into diagnosis and the later route comparison. It was not to purchase the demonstrated product. When the team completed its diagnosis on 28 August, this earlier record would help it specify the narrower retrieval question.

Use Appendix C before the next demonstration you attend. Write the job you want to observe and the actions the session may take. Bring a case that could expose a material mistake. Leave with a record of what happened, what was only claimed and what the next decision still needs. A useful demonstration changes the question you ask afterwards.

CHAPTER 4

Choosing the Route

At 8.40 on Tuesday 1 September, Daniel had three windows open and no recommendation written. One showed the retrieval feature in the company’s existing knowledge-base software. Another held the proposal from the supplier who had demonstrated its product. The third was an internal estimate for a small build using a model service and the company’s own interface.

The sales director was waiting for a separate answer by Friday. His team had a tender to prepare. Leila had offered staff time for the retrieval comparison only if it stayed within the board’s four-day effort allowance. Daniel could see how easily a limited decision might become a procurement exercise larger than the problem.

Maya asked which route was best.

“For what we know today, or for everything we might want later?” he asked.

“For the next decision.”

The question was no longer whether the company should buy an AI assistant. It was whether advisers could locate and inspect the applicable clause more readily than with the repaired library’s ordinary search. The comparison could not contact customers, change case records or make eligibility decisions. It had to end on 11 September.

Daniel closed the supplier’s future-roadmap tab. It was interesting. It did not answer the question in front of them.

Choose the smallest route that answers the question

The existing feature already operated inside the knowledge-base environment. Its configuration could return source passages with document identity and effective-date fields. That did not make it approved for the proposed work. It did mean Daniel could examine the feature without first creating a new document store and a second administration process.

The demonstrated product had a clearer conversational interface and more options for future workflows. It would also introduce a new supplier relationship, a separate source collection and another place to maintain permissions. The small internal build offered control over the interface, but the team would own the integration, testing and support work that the estimate had only begun to describe.

Daniel’s first comparison sheet favoured the internal build because its initial licence estimate was low. Leila asked who would fix it during an incident. The proposed developer was already committed to a finance-system change. The estimate did not contain cover for leave or a maintained fallback. Daniel changed the entry from “low cost” to “initial cost estimated; support commitment unresolved”.

The discovery did not make internal development a bad route in general. It made the route incomplete for this decision. A prototype that one person could produce quickly was not yet a service the company had agreed to support. The missing work would have to be funded or the scope reduced before the route could fairly be compared.

They chose the existing feature as the first candidate for the bounded comparison. It added the fewest new arrangements while exposing the clause and its date. The demonstrated product remained an alternative if the existing feature failed the material requirements. The internal build was deferred, with its missing support commitment recorded rather than dismissed as technically inferior.

Choose the least new thing that can answer the actual question. “Least new” includes sources, permissions, skills, suppliers and obligations to maintain the service. It does not simply mean the lowest licence price or whatever the organisation already owns. An existing product with unsuitable data handling can be the wrong choice; a new product with a justified advantage can be the right one.

The current decision also had a deliberately short horizon. Selecting a route for comparison did not commit the company to a future agent platform. Daniel left the later possibilities in a separate note. They could return when a diagnosed problem required them.

Compare routes against the same requirements

Leila wrote the requirements in the language an adviser would use. Show the passage. Show which document it came from. Show when it applies. Let me open the full source. Tell me when the source set does not settle the question. Do not change a customer’s record while I am looking.

The team then added the operating requirements: approved access, a source owner, evidence of the configured data handling, a way to disable the feature and a manual search route that remained available. Each route was assessed against that same set. A long feature list could not compensate for a missing material requirement.

Their comparison did not use a weighted score to conceal a failed condition. Customer confidentiality and the absence of unauthorised actions were conditions of entry. Convenience and configuration effort mattered after those conditions could be met. If a route could not meet the intended boundary, it would be narrowed or rejected rather than awarded enough points elsewhere to average the defect away.

The existing feature was promising because it could expose the required fields and use the repaired library. The supplier’s product might be more capable at composing replies, but reply composition was outside the immediate comparison. The internal build might offer a better interface later, but nobody had established that the standard interface prevented an adviser from doing the present task.

Daniel kept one route table in his working paper. It recorded requirement, evidence, unresolved question and consequence for the choice. He did not colour every untested claim green because a provider had said the feature was possible. Configuration evidence and a feature description received different statuses.

A fair comparison also gives the ordinary process a chance to improve. Priya showed the team a saved search in the existing library that located a common clause quickly. That case stayed in the comparison. It was not removed because it reduced the apparent benefit of AI. If ordinary search already did a task well, that was information the company could use immediately.

For less familiar product terms, the new retrieval feature might have an advantage. That was the claim to test. The comparison would examine whether the adviser could reach an applicable source, inspect it and know when to stop. It would not reward a polished paragraph for work the adviser still had to undo.

Ask the provider about the actual configuration

At 10.30 on Wednesday 2 September, Daniel met commercial counsel with the existing provider’s order documents, configuration guide and written answers. He had highlighted the relevant feature, not the whole product family. Counsel asked which information the comparison would contain and whether the proposed settings were available under the company’s account.

The comparison would use the repaired procedure extracts and fictional questions. Customer records were excluded. That narrowed the immediate data question, but the procedure material still belonged to the company and needed appropriate handling. A later experiment with other information would require its own scope to be checked.

They traced where the feature processed a question, where the source material was indexed and what was retained in the configured service. They asked who could access it for support, which other organisations were involved in processing, what changes could occur without notice, and what records would be available if something went wrong. Each answer was tied to a document or a configuration check.

These were ordinary provider questions made specific to the use. There was no need to turn them into an encyclopaedic questionnaire about every possible AI risk. Nor was there a reason to accept “enterprise grade” as an answer to a question about retained content.

Daniel found an optional history setting enabled in the assessment account. The feature could keep user questions for convenience. The comparison did not need that history. He disabled it and recorded the setting, while counsel kept the provider’s separate operational-log terms in view. Switching off visible history did not, by itself, prove that no information existed anywhere else.

The team’s record distinguished content used for the feature, permitted operational metadata and information retained by the company for its own evaluation. Where a term remained unclear, they asked a narrower question. They did not silently convert “not used for training” into “not retained”, or “deleted from the screen” into “removed from every copy”.

For this fictional comparison, counsel accepted the documented handling of the restricted source set and Daniel verified the selected settings. That conclusion applied to this configuration and purpose. It was not an assurance about all products from the provider or a general rule that an existing contract covers every newly enabled feature.

They retained the evidence in the company’s controlled procurement record. The board paper contained the conclusion, the scope and the unresolved issues, with links to the records. It did not reproduce confidential contract terms in a widely circulated appendix.

Check the promise against the screen

At 1.15 that afternoon, Priya tried the configured feature with Daniel. It returned the correct current clause. Beside it was a date. At first glance the requirement appeared satisfied.

Priya opened the source. The date in the result was the file’s last modified date, not the procedure’s effective date. The underlying field was available, but the standard result layout had selected the wrong one. A pass against “shows a date” would have accepted a misleading display.

Daniel changed the mapping to the explicit effective-date field. He also labelled it in words. They reran the case and retained the before-and-after result. The new configuration showed what the team had actually asked to know. It still did not decide which period applied to every possible customer event.

A second test used an extract without an effective-date value. The feature initially presented it alongside complete records without making the omission prominent. Daniel changed the result layout to display “effective date not supplied” and prevented that source from appearing as an unqualified current instruction in the comparison collection.

Leila asked whether they could simply discard the incomplete file. For the comparison they could exclude it from the active source set, with the exclusion recorded. For the library itself, the owner still had to resolve the missing information. Hiding the defect from a search result was not the same as repairing the source.

The team kept the route but changed the configuration and its entry conditions. This was the discovery that made the comparison useful before any performance number was available. The existing feature could meet the narrow requirement, but the default display had not done so. Buying a new product would not have relieved the company of deciding what the date meant.

Ask someone who does the work to test the promise as it appears on the screen. “Source shown” can mean a document title without the relevant clause. “Date shown” can mean the upload date. “Human approval” can mean a person who never sees the proposed action. Replace these labels with an observable behaviour and a record of whether it occurred.

Do not make a configuration screenshot carry more than it can prove. It can show a setting or a result at a time. It does not prove that the setting cannot change, that every source has complete metadata, or that a provider will preserve the display after an update. Those become operating questions for the next stage.

Give the tender thread a bounded yes

At 9.00 on Friday 4 September, the sales director came to Maya with Daniel and commercial counsel. He had brought the first tender proposed for the new route and the requirement-mapping sheet his team already used. The original trial account remained disabled.

Counsel had reviewed this customer’s terms and the selected enterprise tool’s agreement. In this fictional case, the customer terms permitted the proposed processing in that approved environment. The provider agreement and the activated configuration allowed processing without retaining the submitted tender content or generated output after the request, and without using either for model training.

Daniel showed the account settings and the written confirmation for the specific service path. Counsel had checked the treatment of content across processing and support, rather than relying only on a “no retention” label. Limited non-content operational metadata was described separately. The company would retain its own tender and review records under its existing arrangements.

This was not a universal conclusion about tender documents or enterprise tools. It applied to the reviewed customer, document set, purpose and configuration. A different customer’s restrictions could lead to a different answer. Documents with unresolved permissions would stay outside the approved route.

Maya asked who would approve the answer sent to the customer. The sales director would remain accountable. The tool could assist with identifying requirements and drafting a mapping. A named bid reviewer would check each requirement against the exact source clause before the work entered the proposal. The tool could not send the tender, commit a price or accept an obligation.

The sales director pointed to the checking time already included in his baseline. “We’re not inventing review because there’s a model,” he said. “We’re making sure this review catches the mistake it can introduce.”

He was right. The new check had to fit the actual workflow rather than appear as an additional signature at the end. A reviewer would mark each mapped requirement as supported, disputed or missing. A link to a nearby clause would not count as support merely because it looked plausible.

Maya authorised the first reviewed tender and a limited continuation for documents meeting the same recorded conditions, under the board’s delegation of 28 August. The review date was Friday 18 September. Any new customer terms, changed processing path or failure of the citation check would require clarification or a pause before further use.

The sales director left with a yes he could explain to his team. He also left with a boundary more precise than “use AI carefully”. The original incident had not proved that tender assistance should never be used. It had shown that permission, source checking and a record of the work could not be assumed.

Put the check where the error becomes visible

At 2.10 that Friday, the bid reviewer found a requirement that the approved tool’s first draft had linked to a general service clause rather than the customer’s more specific exception. The linked text existed. It did not establish the proposed response.

The reviewer marked the mapping disputed and returned it to the sales director. They corrected the reference before the proposal was used. The discrepancy and its correction stayed in the tender review record. They did not use this caught defect to claim a reliable error rate or to conclude that every future error would be equally visible.

Maya asked whether this meant the authorisation had failed. The sales director said the bounded route had produced a defect that its intended review caught. The result supported retaining the source check. It did not establish a net time benefit, because the team had not yet completed a comparable timed tender with the new workflow.

That answer was more useful than either celebration or an automatic shutdown. The team examined the defect’s consequence, whether the required check had worked, and whether the boundary still held. No unsupported requirement had reached the customer, and no unauthorised processing had been found in this instance. A missing review, an external release or a changed data-handling condition would have called for a different response.

The reviewer also compared the original requirements list with the mapping, so the check could find an omission as well as a wrong citation. Looking only at the tool’s selected rows would not reveal a requirement it had left out. This completeness check came from the two omissions found in the pre-trial baseline.

The authority mode was Assist. People prepared, checked and approved the tender. The model output was material they worked with, not an accepted statement of what the company owed or could promise. The practical boundary was expressed through access, review and release permissions rather than through the mode’s name alone.

A checked first tender would be evidence about that tender. It would not justify changing the register entry to “safe” without qualification. Daniel recorded the approved purpose, configuration, owner and review date. Future observations would accumulate beside that decision.

Daniel did not describe the whole provisional register as verified. The wider discovery record still contained overdue checks from 21 August. He named them as overdue in the status note and gave Maya the remaining owners and follow-up actions. Approval of the reviewed tender route did not settle those unrelated reports.

He also kept the disabled trial account separate from the newly approved enterprise route. Its available history remained preserved, and complete disposal had not been established. Formal retirement would still require checking retained records, provider obligations and access closure. The new yes did not erase that unfinished work or make the earlier unrecorded incident disappear.

Write down why the remaining uncertainty is acceptable

For the retrieval route, Maya drafted a decision record before the comparison was complete. She wanted to distinguish choosing what to examine from deciding that it should be used with customers. The first decision could be made with less evidence because the scope was narrower and the consequences were contained.

She wrote the decision in one sentence: configure the existing knowledge-base retrieval feature for a bounded comparison with ordinary search, using the repaired source set and fictional questions, within the board’s four staff-days and AUD 1,500 ceiling, ending 11 September.

Then she added the three fields Nadia had asked her to make explicit. What was still uncertain? Why was that acceptable for this decision? What would make the company reconsider? The questions forced her to explain permission in relation to evidence rather than write “risk accepted” and leave the reasoning invisible.

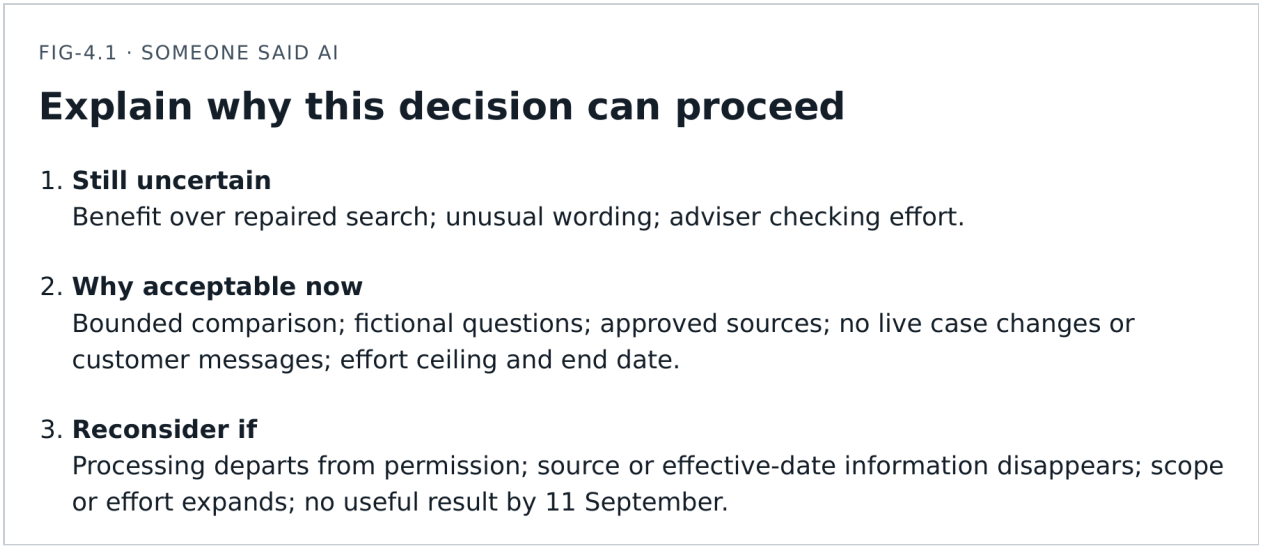


FIG-4.1 State what remains uncertain, why it fits this limited decision, and what would require reconsideration. Source: Author’s method and fictional case.

Still uncertain: whether retrieval improves an adviser’s source-finding work over the repaired library’s search; whether unusual product language produces misleading matches; and how much checking the displayed sources require. The available tests did not establish broad performance, live customer benefit or sustained operating effort.

Why acceptable: the current activity was a time-limited comparison with fictional questions and an approved procedure subset. It had no customer communication, financial authority or connection that could change a live case. The existing search remained available. The comparison could answer the uncertainty without first exposing customers to the proposed use.

Reconsider if: the selected configuration processed information outside the approved route; source identity or effective-date information disappeared; the comparison required live customer data or record access; effort exceeded the ceiling; or no useful distinction from ordinary search emerged by the end date. Daniel could disable the feature and inform Maya. A broader proposal would return for a new decision.

The three fields did not replace the evidence. They explained why the evidence was enough for this action and not for a larger one. A later reader could see what the decision depended on, instead of guessing from an approval signature.

The same fields appeared in the tender record with different answers. Remaining uncertainty concerned benefit and error frequency across eligible tenders. Limited use was acceptable because the specific processing permission and configuration were evidenced, source checking covered every requirement, and people retained release authority. A permission change, failed check or unauthorised disclosure would interrupt that route. The retrieval decision could not be copied across unchanged.

Stop asking questions that cannot change this decision

By Monday 7 September at 11.20, the team had collected another page of possible provider questions. Some concerned functions they had disabled. Others concerned a future deployment across the company. Daniel worried that he would spend the remaining comparison time describing possibilities rather than testing the task.

Nadia asked him to put each question beside a consequence. Would the answer change permission for the current comparison, change how it was configured, change what it measured, or change the response to a failure? If so, it belonged now. If it concerned a later expansion, it belonged with that later decision.

They retained the unresolved question about source refresh because it affected whether the comparison used the intended procedure version. They deferred a question about bulk customer-message dispatch because there was no dispatch connection in the proposed route. If a later proposal requested one, the question would return with the new authority.

They also dropped a request for a broad certification claim that would not answer their particular configuration question. Counsel already had the applicable handling terms. Daniel still needed to show that those terms and settings covered the actual feature path. More general assurance material would not substitute for that missing link.

Knowing when to stop gathering evidence requires a stated decision boundary. Without it, every possible future use can become a reason to delay the present one. With it, the team can say which uncertainty it will learn through a contained activity and which uncertainty prevents that activity from beginning.

This is not permission to stop because the paper is long or the deadline is near. If a material question remains unanswered, narrow the action, change the route or defer it. The reason to proceed is that the remaining uncertainty fits within the permitted scope and its protections, not that the organisation has become tired of asking.

Maya kept the outstanding questions with their future triggers. They were neither forgotten nor treated as prerequisites for work they could not affect. Daniel could return to the comparison.

Compare the whole act of finding a source

On Thursday 10 September at 2.30, Priya and another adviser completed the final comparison session. Daniel had prepared twelve fictional questions against the repaired source set. Eight represented ordinary cases, two involved exceptions and two could not be settled by the supplied material. The categories were chosen to expose specific questions, not to reproduce the company’s workload proportions.

Each adviser used both ordinary search and the configured retrieval feature on matched variants, with the order alternated. The team still expected learning effects and did not claim a controlled trial. They recorded time until the adviser had inspected an applicable source, not time until the first text appeared on the screen.

The feature helped on some less familiar wording. It offered little advantage where a saved search already led directly to the clause. In an unsupported case it returned related material, which the adviser had to recognise as insufficient. Priya refused to call that a resolved question merely because the display was populated.

Daniel retained the timings, outputs and source checks as an exploratory comparison. The set was too small and deliberately constructed to support a company-wide productivity claim. It was enough to show that the configured feature could expose the required source information and that some of the residual search difficulty remained worth examining with more realistic work.

The result changed the next proposal. The team would not ask for a general customer-answering assistant. It would ask for a bounded experiment on source retrieval, with reply drafting considered separately. The common saved searches would remain available, and an unsupported result would require an adviser to use the existing escalation route.

The provider’s original 96 per cent figure did not appear in the recommendation. The team now had evidence closer to its actual task, albeit limited. It also had a clearer question for the experiment: under ordinary service conditions, does this feature help advisers reach and check the applicable source without creating disproportionate review work or misleading certainty?

Keep the cost of waiting visible

The sales director’s Friday authorisation had removed one avoidable delay. The service comparison had a different end point. Leila wanted advisers to benefit from the source display soon, but she also needed the team to define the experiment’s participants, information and stopping conditions before introducing it into customer work.

Maya recorded the cost of that remaining delay as a question for the next decision. The process repair was already helping customers; the residual source-search work continued manually. They would measure the additional review work required by the proposed experiment rather than assume that more checking could only help.

This did not settle the later choice about blanket second checks. That judgement still lay ahead. It did establish that waiting and review consumed something real: staff attention, service capacity and customer time. Those costs belonged beside the harms the team hoped to prevent.

A fair decision record should let someone disagree intelligently. The dissenting director could point to the reversible nature of internal retrieval. Nadia could point to unanswered questions about use under pressure. Neither argument had to be caricatured for the organisation to choose a limited next step.

Close the comparison and preserve the authority boundary

At 9.15 on Friday 11 September, Maya reviewed the comparison record with Leila, Daniel and Nadia. The work had used three and a half staff-days and AUD 900 of the authorised allowance. The saved work log and external-support invoice were attached. The unused allowance did not become permission to extend

the trial indefinitely.

They selected the existing feature as the route to take into the experiment proposal. They did not authorise customer-facing use in that meeting. The board had authorised a comparison; the next proposal would specify the experiment budget, scope and participants for the board’s decision. Maya could implement an approved experiment within delegation, but could not create that authority herself.

Daniel disabled the comparison access pending the next approved stage, preserved the evaluation records and confirmed that ordinary search remained available. The repaired library continued operating. The tender route also continued within its separate authorisation and review date. Stopping one temporary activity did not mean stopping all useful work.

Leila asked for the source-maintenance time to appear in the experiment proposal. The feature had not removed the need to maintain effective dates or resolve contradictory material. She would not accept an experiment budget that treated her team’s source work as free simply because it did not arrive on a supplier invoice.

Maya agreed. The next paper would carry the residual question, the selected route and the proposed controls. It would also carry the cost of those controls. Choosing the technology had made the decision more concrete, not completed it.

Before the meeting ended, Nadia asked who would receive the decision. Daniel needed the configuration boundary, advisers needed to know that the comparison had ended, and the sales team needed confirmation that its separate permission still stood. Maya sent a short operational message to each group and retained the fuller record for review.

An approval held only in a meeting note would have left people guessing. The message named the permitted activity, its owner and the next review, using the same words as the decision record. It also told users where to raise an unexpected result. Communication was part of making the decision effective, rather than an announcement saved for a later launch.

Complete the Decision Record

The worked retrieval record stated the decision and its limits first. It named the existing feature, the comparison purpose, the approved source set and the absence of live customer actions. It linked to the board’s 28 August authority, the configuration checks, the provider review and the saved comparison results.

It then recorded the alternatives. Ordinary search remained the fallback and comparator. The demonstrated product was retained as an alternative if a material requirement could not be met. The internal build was deferred because its support commitment was unresolved. Those conclusions concerned the present problem; none was a universal ranking of the three routes.

The record’s three evidence fields explained remaining uncertainty, why it was acceptable for the limited comparison, and what would trigger reconsideration. Daniel owned configuration and disabling the feature; Leila owned the service outcome and procedure sources; Nadia challenged the interpretation;

Maya owned the decision within the authority granted. The next decision was whether to approve the bounded experiment, not whether the company had “adopted AI”.

Appendix D provides the blank form and a compact worked version. Use it when the evidence has become sufficient for a specific action, even if it remains insufficient for a larger ambition. Write the boundary so the next person can distinguish the two. Then allow the authorised work to proceed, and stop gathering evidence that belongs to a different decision.

EPILOGUE

The Empty Chair

> Set after the planned Chapter 12. Chapters 5 to 12 arrive in later editions.

At 9.05 on a Thursday morning, after the strategy discussion in Chapter 12, another proposal arrived with the word AI in its title.

The small-claims system was one of the priorities Tom’s strategy had selected for investigation. It could compare claims with company rules and approve eligible amounts beneath a financial limit. The proposal now also asked it to decline claims it judged ineligible. That additional authority was written into the decision requested from the board.

One chair at the table was empty.

It belonged to a member of the standing customer panel introduced during the service-credit work. Her train had been cancelled. She had helped challenge the earlier human correction process. She could not speak for every customer, and the board could meet without her.

The sponsor finished his presentation. The demonstration had shown what the system could do on selected cases. Its performance evidence sat in an appendix. The question now was which live decisions that evidence could support.

Leila, attending from her regional role, asked what an employee would see when the system declined a claim. Daniel asked whether the source rule and its effective date would travel with the decision. Finance asked how many cases would require review near the threshold of the tested decision score. Nadia asked who could suspend approval authority without waiting for the next board meeting.

The score had been checked against known outcomes in the evaluation set. It was not the model announcing how sure it felt, and its earlier test did not establish how it would perform under every future condition. Finance wanted the observed workload around the proposed threshold, not a confident phrase in a generated answer.

Then Maya turned back to the business case. One line read: “Customers accept an automated decline when the reason is shown.” Beside it was the word confirmed.

“Which claim depends on customers,” she asked, “and who tested it?”

The sponsor pointed to the complaints figures and the notes from the customer panel.

“The panel asked us to explain decisions clearly. Complaints fell after we changed the correction process.”

“That was a person explaining and correcting a decision,” Leila said. “Did we ask about receiving a decline from this system?”

“No. We treated it as the same requirement.”

Nadia opened the complaints appendix. It described customers who had complained. It did not establish how many had misunderstood a refusal, abandoned a claim or found the challenge route too difficult to use.

“The empty chair isn’t the missing test,” Maya said. “We would need this evidence if she were here. One person could not confirm that sentence for everyone.”

Daniel suggested a remote conversation with the panel member that afternoon. The sponsor offered to send her a written question set. Leila proposed reviewing recorded calls about declined claims, within the permissions governing those recordings, to find what customers had struggled to understand.

Those steps could improve the test questions. None would settle whether customers could understand and challenge this proposed automated decline. The team agreed to try the proposed explanation and challenge route with a small, deliberately varied sample of customers. They would record whom the test did not represent and what uncertainty remained.

“What changes if it fails?” Finance asked.

“Declines go to a person,” Daniel said. “We consider automated approvals separately. The cost case changes because people still examine the refusals.”

“And approvals still require their own evidence,” Nadia said. “Removing one decision from the system does not approve the other.”

The board authorised a bounded comparison using appropriately authorised, minimised or de-identified historical records in a controlled environment, and the customer test of explanation and challenge. Access, retention and disposal had to be agreed before records were used. No live claim could be changed. The sponsor would return with separate proposals for automated approval and any automated decline, with the revised workload and cost case.

One director voted against the delay. He believed a limited release with a staffed correction route would teach the company more quickly. The team had evidence gaps, he said, but customers also waited while the company sought greater certainty. His objection and the proposed alternative were recorded beside the decision.

After the meeting, the sponsor caught Maya near the lift.

“You know she might tell us customers want a faster answer.”

“She might. We should hear that too.”

“So what changed?”

“We had counted an assumption as evidence,” Maya said. “Now we know which decision depends on it.”

The lift arrived. The sponsor stayed beside the open paper.

He moved the live-authority decision to follow the customer test and the revised comparison. Beside the specific claim that customers accepted an automated decline when a reason was shown, he removed confirmed and wrote not yet heard.

FIG-E.1 · SOMEONE SAID AI

Correct the claim in the paper

Claim	Earlier label
Customers accept an automated decline when the reason is shown.	Confirmed
Corrected label	**Decision consequence**
Not yet heard	Test comprehension and challenge. If unsupported, refer declines to a person; any automated approval still needs its own evidence.

FIG-E.1 The customer claim remains untested; an absent representative is not the reason to withhold authority. Source: author’s method and fictional case.

APPENDIX A

The two layer Demand Note

Use the two layer Demand Note

Begin with the first pass below. A routine use already inside its approval does not need this note. Where permission is uncertain, clarify it with the approval owner. Each unknown needs an owner and a review date.

What is being requested in the original words

Answer:

Unknown to resolve, owner and date:

What should improve and who owns it

Answer:

Unknown to resolve, owner and date:

What is already happening

Answer:

Unknown to resolve, owner and date:

What needs an immediate boundary

Answer:

Unknown to resolve, owner and date:

What is the next decision by whom and when

Answer:

Unknown to resolve, owner and date:

Add only the supporting detail needed

Pressures and commitments

Why now? Record deadlines, promises, supplier influence and the source of each reported pressure.

Record:

Baseline and measures

What should change, for whom? Name the owner, current measure, source, measurement period and unknowns.

Record:

Affected people

Identify users, customers, indirect effects and people who may struggle to raise concerns. Record how their evidence will be obtained and its limits.

Record:

Voice correction and escalation

Which complaints, reversals and service records inform the decision? Who corrects live records, receives concerns and can pause the work?

Record:

Challenge and specialist screen

Who tests the evidence? Record conflicts, relevant expertise and any specialist review required by the consequence, information or obligations.

Record:

Boundaries

For each restriction record its reason, supporting evidence, owner, enforcer, exception requester, approver, review date and lifting conditions. Separate delegated boundary changes from new spending or experiment authority.

Record:

Evidence record

Record each claim, status, source, checker, check date, scope and remaining limits. Reports and assumptions remain labelled. Attach the source or a controlled link; do not copy sensitive material into an unrestricted note.

Record:

Decision and follow up

State decline, contain and clarify, or bounded diagnosis. Preserve containment on any route with existing exposure. Record decision maker, rationale, scope, effort and spending ceilings, dissent, dates and evidence required next.

Record:

APPENDIX B

Diagnosis Record

Worked example

Question and owner

Does service-credit repeat work require an AI assistant? Leila owns the outcome.

Observed evidence

Twenty July cases: 48 contacts, 13 with repeats; ten have missing reasons or unresolved procedure versions as the principal suspected explanation. Source: linked case, contact and credit records, reviewed 17 August.

Baseline and repair

July cohort: 1,200 requests, 2,880 contacts, 132 reversals, with seven-day follow-up. Repair: supersession notice, mandatory decision reason, visible correction and library ownership. First 100 repaired requests: 180 contacts and four reversals; early, non-randomised observation only.

Decision and remaining uncertainty

Continue the repair. Compare source retrieval with ordinary search on the repaired library; no live customer contact or case changes. Board decision 28 August; four staff-days and AUD 1,500; end 11 September. Broader performance and persistent benefit remain unestablished.

Complete your diagnosis record

Keep the first record short. Attach evidence or controlled links where needed. An unknown needs an owner and a date. Expand these fields on paper, or use the matching tool at book.almostmagic.io.

Original question and owner

What outcome is sought and who owns it?

Record:

Observed work and sources

Follow one complete case, including return contacts and correction.

Record:

Baseline and limits

Define cohort, denominator, follow-up, exclusions and data quality.

Record:

Repair and observed result

What simpler change was tried and what happened?

Record:

Residual question and alternatives

What remains; compare process, existing software, rules and AI.

Record:

Decision and next boundary

Close, repair or bounded comparison; owner, authority, effort and review date.

Record:

Record owner:

Date:

Review:

APPENDIX C

Demonstration Question Sheet

Worked example

Task and scope

13 August demonstration, told as a flashback. Find a service-credit clause using a fictional pack; no live customer data, system connections or external actions.

Observed facts

Saved output follows an older rule; the source can be displayed; a rerun after changing the active pack uses the newer source. The underlying cause of the original selection is not established.

Other evidence classes

Reported: supplier’s fictional 96 per cent accuracy claim, not locally validated. Assumption: source display may help checking. Unknown: applicability across the library, relevant score meaning, live permissions and update behaviour.

Next decision

Carry source identity and effective-date visibility into the diagnosis and route comparison. The session does not authorise procurement or customer-facing use.

Complete your demonstration question sheet

Keep the first record short. Attach evidence or controlled links where needed. An unknown needs an owner and a date. Expand these fields on paper, or use the matching tool at book.almostmagic.io.

Task and permitted session

Who will use the output; what may the demonstration access or do?

Record:

Material test cases

Ordinary case, exception, obsolete source and absent answer.

Record:

Observed behaviour and record

Save inputs, source versions, outputs and configuration.

Record:

Claims and uncertainty

Separate observed facts, supplier reports, assumptions and unknowns.

Record:

Action reach and recovery

Which tools can read, write, spend or send; who can stop them?

Record:

Next decision

What did the session change; what must the next comparison establish?

Record:

Record owner:

Date:

Review:

APPENDIX D

Decision Record

Worked example

Decision and authority

Configure the existing knowledge-base feature for a bounded comparison. Authority: board decision of 28 August. Ceiling: four staff-days and AUD 1,500, ending 11 September. Fictional questions, repaired approved sources; no live customer actions.

Still uncertain

Benefit over ordinary search, less common wording and source-checking effort. Limited exploratory cases do not establish production performance.

Why acceptable now

The comparison can examine these uncertainties without connecting to live cases, communicating with customers or making financial decisions. Ordinary search remains available; Daniel can disable the feature.

Reconsider if

Processing leaves the permitted route; source or effective-date information disappears; live data or extra authority becomes necessary; effort exceeds the ceiling; no useful result by the end date.

Outcome and next decision

11 September: three and a half staff-days and AUD 900 used. Select the route for an experiment proposal; disable temporary comparison access. The board still decides the experiment budget and scope. Tender assistance has a separate authority and 18 September review.

Complete your decision record

Keep the first record short. Attach evidence or controlled links where needed. An unknown needs an owner and a date. Expand these fields on paper, or use the matching tool at book.almostmagic.io.

Decision and authority

Exactly what is approved; whose authority; for how long and within what budget?

Record:

Alternatives and evidence

Why this route; supporting records, checks and dissent.

Record:

Still uncertain

Which material questions remain unanswered?

Record:

Why acceptable for this decision

Connect each uncertainty to the permitted scope and protections.

Record:

What would make us reconsider

Name triggers, pause owner and action required.

Record:

Communication and next review

Who must know; date; next decision and records to retain.

Record:

Record owner:

Date:

Review:

Notes and references

The company, its people, supplier demonstration, internal records and numerical results are fictional. Case figures illustrate decision-making and are not reported research findings. The technical claims below are supported by external sources. The remaining chapters and their references will be added in later editions.

Numbered notes

1. Generated language can be fluent yet contain hallucinated or unsupported content. Ji et al., abstract and survey overview. The fictional wrong-rule example is the author’s case, not a result reported in that paper. Reference R19.

2. Generative models produce synthetic content; language models generate successive tokens and can produce false or inconsistent output. Generative AI Profile, Section 2.2 and its introductory definition. Forecasting demand versus composing a reply is the author’s practical company example. Reference R03.

3. Decoding and sampling choices affect generated text and its diversity. Holtzman et al., abstract. The chapter’s saved-run and version-recording advice is the author’s method. Reference R20.

4. Retrieval-augmented generation combines retrieved external material with generation. Lewis et al., abstract. The company’s source-date requirements are the author’s design judgement, not a claim that the paper validates this service. Reference R21.

5. AI assistance can improve some tasks and worsen others within a knowledge workflow. Dell’Acqua et al., journal abstract. The study does not establish this fictional product’s accuracy or the company’s expected benefit. Reference R14.

6. Classifier confidence estimates can be poorly calibrated. Guo et al., abstract. The definition of tested decision score and the 0.8 illustration are explanatory method, not a guarantee or a result from the case. Reference R15.

7. Language-model systems can combine reasoning and action steps and interface with external tools or environments. Yao et al., abstract. The proposed stopping and recovery boundaries are the author’s operating recommendations. Reference R22.

Reference records

R19

Ziwei Ji et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, DOI 10.1145/3571730; author version arXiv:2202.03629v7, 2024.

Type: Paper

Live source: https://arxiv.org/abs/2202.03629

Claim supported: Chapter 3, note 1: fluency does not establish factual or source faithfulness.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260909044219/https://arxiv.org/abs/2202.03629

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.

R03

NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, July 2024. DOI 10.6028/NIST.AI.600-1.

Type: Government guidance

Live source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Claim supported: Chapter 3, note 2: generated content and token prediction; confident erroneous output (Section 2.2).

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260921040857/https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.

R20

Ari Holtzman et al. The Curious Case of Neural Text Degeneration. ICLR 2020; arXiv:1904.09751v2.

Type: Paper

Live source: https://arxiv.org/abs/1904.09751

Claim supported: Chapter 3, note 3: decoding and sampling choices affect generated text and diversity.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260917122236/https://arxiv.org/abs/1904.09751

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.

R21

Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020; author version arXiv:2005.11401v4, 2021.

Type: Paper

Live source: https://arxiv.org/abs/2005.11401

Claim supported: Chapter 3, note 4 and FIG-3.1: external retrieval combined with generation.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260915175627/https://arxiv.org/abs/2005.11401

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.

R14

Fabrizio Dell'Acqua et al. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science, published online 11 March 2026. DOI 10.1287/orsc.2025.21838.

Type: Paper

Live source: https://pubsonline.informs.org/doi/abs/10.1287/orsc.2025.21838

Claim supported: Chapter 3, note 5: uneven effects across tasks. Replaces the starter working-paper version for this claim.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: capture attempt failed; existing snapshot could not be established. Recapture needed before publication.

Status: Verify before publish. Recheck within 30 days of publication.

R15

Chuan Guo, Geoff Pleiss, Yu Sun and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. ICML 2017, PMLR 70, pp. 1321–1330.

Type: Paper

Live source: https://proceedings.mlr.press/v70/guo17a.html

Claim supported: Chapter 3, note 6 and FIG-3.2: numerical classifier confidence can be poorly calibrated; not a study of verbal certainty.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260920133227/https://proceedings.mlr.press/v70/guo17a.html

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.

R22

Shunyu Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023; arXiv:2210.03629v3.

Type: Paper

Live source: https://arxiv.org/abs/2210.03629

Claim supported: Chapter 3, note 7 and FIG-3.3: model-generated reasoning and actions interacting with tools or environments.

Verified on: 24 September 2026. Supporting claim checked against the live source; all remain subject to final publication verification.

Archive: http://web.archive.org/web/20260921035253/https://arxiv.org/abs/2210.03629

Archive status: existing snapshot, not a new capture. Recapture needed; snapshot contents not independently rechecked in this round.

Status: Verify before publish. Recheck within 30 days of publication.