ACCUFIND

Taking AI into production

A companion to the Digital Transformation Agency's guidance for AI proof of concept to scale, and the 5 decisions that stop a pilot becoming a service.

HAMISH CAMERON · ACCUFIND PTY LTD

VERSION 1.0 · SEPTEMBER 2026

A proof of concept succeeds, and then nothing happens

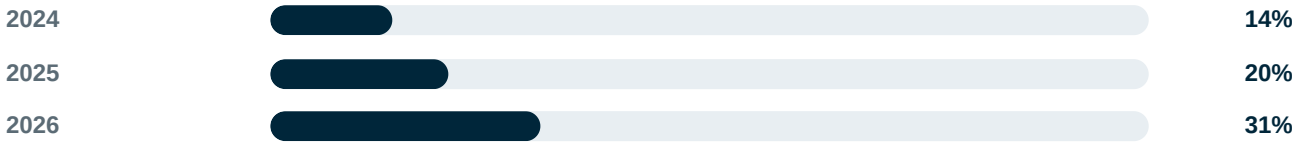
The Digital Transformation Agency publishes a scenario that will be familiar to anyone who has run one. A data science team builds a model to find corporations with questionable registration information. The model works. The team presents it to leadership, and the project stops.

The reason given is not technical. Leadership "did not fully grasp how the model worked and were hesitant to trust an AI to support decision-making".[1] The agency's own account names the causes as lack of trust, risk aversion and misaligned goals. Traditional processes felt safer.

The technology worked and the transition failed. That is the ordinary case.

I have run about a dozen of these in Australian government over the past 2 years, mostly in state and territory agencies. Some became services an agency now runs. Others answered the question they were asked and went no further. What separated them had little to do with the technology, which performed much the same either way. It came down to 5 decisions, and whether anyone had made them before the build started.

NEW ZEALAND GOVERNMENT AI USE CASES REACHING OPERATIONAL STATUS



Counts: 15 of 108 use cases in 2024, 55 of 272 in 2025, 167 of 545 in 2026. Source: Government Chief Digital Office (NZ), cross-agency survey for artificial intelligence use, 2024, 2025 and 2026 editions. digital.govt.nz

New Zealand runs the only published longitudinal count in the region, and it supports 2 readings. The share reaching operational status has more than doubled in 2 years, which is a real improvement and should be reported as one. The number nobody publishes is what became of the other 378 use cases in the 2026 survey. No jurisdiction in Australia or New Zealand reports discontinued AI use cases at all. The missing denominator is itself the finding.

Three things follow.

Agencies already own the processes. Gateway reviews, benefits management, evaluation policy, assurance frameworks and go-live gates all exist. AI work costs less than the thresholds that trigger them, so nobody applies them.

The Commonwealth has now written down the questions. The DTA's guidance for AI proof of concept to scale is a serious piece of work and it asks almost all of the right ones. It does not supply the instruments to answer them, and outside the Commonwealth it does not apply.

5 decisions decide whether it goes further. Most of what stalls a pilot comes down to 5 decisions that nobody made, in an order that matters. The rest of this paper sets them out, with what failure looks like and how to check.

AN INTEREST TO DECLARE

AccuFind sells grounded retrieval for legal and policy work in government. Decision 4 of this paper concludes that most agencies should buy rather than train, which is a conclusion that suits our commercial position. Read that section with the interest in mind. The evidence behind it is cited so you can check it yourself, and the exceptions where training is warranted are named rather than buried.

A note on the evidence, before any of it is used

Writing about AI adoption means working in a field where the most quoted numbers are the weakest. The figure you are most likely to be shown in a briefing is that 95 per cent of AI initiatives fail. Mistral AI publishes it on its own AI transformation page with no source or footnote attached.[2]

The figure traces to an unrefereed MIT NANDA report. It rests on 52 interviews, 153 survey responses and around 300 publicly disclosed initiatives. The report includes its own caveat that its numbers are "directionally accurate based on individual interviews rather than official company reporting".[3] Its own funnel is more interesting than its headline. 60 per cent of organisations evaluated custom tools, 20 per cent piloted, and 5 per cent reached production. Among organisations that actually piloted, that is roughly a 75 per cent failure rate. The 95 per cent is reached by counting the organisations that never piloted as failures.

The temptation at this point is to replace it with a better survey. Most of the alternatives have the same defect. Deloitte, McKinsey, BCG and IBM all run self-reported, non-random, unaudited surveys with a commercial interest in the answer. A paper that demolishes one and then leans on 6 has no standard at all.

So every number below is graded against a stated hierarchy, strongest first.

THE EVIDENCE HIERARCHY USED HERE

- Audit office and parliamentary committee findings
- Official statistics, including the Australian Bureau of Statistics and the OECD
- Controlled studies with a comparison group
- Peer-reviewed research
- Large surveys that publish their method and sample
- Analyst forecasts
- Vendor claims

Every framework author says to use the processes you already

have

There is a widespread belief that AI needs its own assurance apparatus because it is a different kind of technology. The people who write the frameworks do not believe this.

The NIST AI Risk Management Framework says AI risk management "should be integrated and incorporated into broader enterprise risk management strategies and processes".[4] Australia's National framework for the assurance of AI in government was agreed by all jurisdictions in June 2024. It says existing decision-making and accountability structures "should be adapted and updated".[5] The DTA said in December 2025 that its new impact assessment and procurement tools "complement and strengthen, not duplicate" existing frameworks.[6]

New South Wales has gone furthest and demonstrated that it works. Its AI Assessment Framework is mandatory under Circular DCS-2024-04. The framework is part of the existing Digital Assurance Framework, and an AI Assurance Review was added to the Gateway gates the state already ran.[7]

The same holds for evaluation, and this is the part most often missed. The Commonwealth has run an evaluation policy since 1 January 2022, administered by Finance. It established an Australian Centre for Evaluation in Treasury in 2023, to lift the volume and quality of impact evaluation across the service.[8] The United Kingdom has gone one step further and written the AI-specific guidance. HM Treasury's Magenta Book now includes a supplementary guide on evaluating AI interventions. It states the position plainly. "The key principles of robust impact evaluation are no different for AI interventions than for any other type of government programme."[9]

So the doctrine is settled and it has been for some time. The audit offices tell us what is happening in practice.

The Australian National Audit Office examined AI governance at the Australian Taxation Office and found the arrangements only partly effective. 74 per cent of AI models had no completed data ethics assessment. There was no centralised inventory of AI in use, and no evidence of structured and regular monitoring of models in production.[10] The Queensland Audit Office looked at the Department of Transport and Main Roads. It had not incorporated AI ethical risk management into its policies, or into its existing ICT governance, at all.[11] The Audit Office of New South Wales found 357 AI tools in use across 21 agencies. Only 38 per cent of those agencies held a formal AI policy.[12]

WHAT THREE AUDIT OFFICES FOUND

74%
of ATO AI models without a completed ethics assessment

357
AI tools in use across 21 NSW agencies

38%
of those NSW agencies with a formal AI policy

Agencies are not applying the doctrine they already have. Nobody needs to write another assurance framework. Agencies need to point the one they hold at this work.

There is a structural reason this keeps happening, and it is fixable. Commonwealth Gateway review thresholds sit at $30 million, or $30 million with $10 million in ICT.[13] A typical AI proof of concept sits nowhere near that. So AI work escapes the heaviest independent assurance without any person or committee ever deciding that it should be exempt. The exemption is an artefact of a threshold set for capital projects.

The Commonwealth has now written the questions down

In March 2026 the DTA published Guidance for AI proof of concept to scale.[14] It addresses this problem directly, and it is better than most things published on the subject anywhere.

It sets out 3 stages. A proof of concept is an "early, experimental build to test concepts, feasibility, or specific technical components". It runs on synthetic data in a sandbox, with high risk tolerance, where "failure is acceptable and expected". A pilot is a "limited-scale, real-world implementation to validate value, usability and readiness", on live or near-live data, with moderate risk tolerance. Production is a "fully deployed, enterprise-scale system integrated into business-as-usual operations". Risk tolerance is low. Full compliance applies: the Protective Security Policy Framework, the Information Security Manual, AI assurance, procurement and audit standards.[15]

It sets out 8 principles and 12 dimensions. It gives 6 scenarios drawn from real successes and failures, and 3 appendices covering readiness, evaluation and procurement.

The readiness checklist is the strongest part. It asks whether the proof of concept is linked to a clear business priority, and whether success criteria and measurable outcomes have been defined. It asks whether the Benefits Management Policy has been followed. It asks whether there is a principal business owner who can validate the outcomes. Before scaling it asks whether adequate funding has been sought and approved for the pathway to production. It also asks about provision for top-up funding. It even asks whether a decision register has been established.[16]

Those are the right questions. An agency that worked through all of them before starting would avoid most of what goes wrong.

A question list is not an instrument

The guidance describes itself as a complement to "existing Australian Government frameworks and standards". It offers "a flexible approach that agencies can adapt to their operational contexts". That is a deliberate and defensible choice. It also means the guidance stops at the point where the work starts.

Take the clearest example. Principle 5 requires that AI initiatives are tied to business priorities "with defined success metrics and baselines, systematic evaluation methods and pathways to value".[17] That is the right requirement, and it is the one least often met.

Now look for the method. Appendix 2 names 5 evaluation areas. It offers tactics such as testing on edge cases and noisy data, tracking latency, and running cost-benefit analysis.[18] Appendix 3's entire treatment of measurement is a single sentence: "Include performance indicators such as accuracy, fairness, explainability and user satisfaction."[19]

There is no method for establishing a baseline. There is no scoring method, no rubric, and no guidance on who scores what or how to tell whether 2 evaluators agree. An agency that reads Principle 5, agrees with it, and asks how to do it has nowhere in the document to go.

The same holds for the decision register the checklist asks for. The guidance supplies no register, and does not say which decisions belong in one.

WHAT THE GUIDANCE GIVES

The questions

- Three stages with risk, data and compliance settings
- Twelve dimensions to plan across
- A readiness checklist at 5 points in the lifecycle
- Six scenarios showing what went wrong

WHAT AN AGENCY STILL NEEDS

The answers

- A method for baselining a process before building
- A scoring method that two people apply consistently
- A register that says which decisions must be closed
- A funding route from pilot to appropriation

Twelve dimensions is a planning list rather than a sequence. Nobody closes 12 dimensions. The guidance does not say which of the 12 will actually stop a transition, which can be left until later, or what evidence closes one. That is the judgement an agency has to supply, and it is the judgement this paper tries to supply.

Outside the Commonwealth, the guidance does not apply

The DTA guidance is guidance. It binds nobody, including in the Commonwealth. Outside the Commonwealth it has no standing whatever, and that is where most Australian government service delivery happens.

The states and territories have gone in 4 different directions.

MANDATORY	**NSW** The AI Assessment Framework binds all agencies under Circular DCS-2024-04, across the full lifecycle. High and critical risk systems go to an AI Review Committee, which publishes no outcomes and no statistics.
MANDATORY	**WA** Every public sector entity must name an executive AI Accountable Officer. Each must also complete a self-assessment for any AI or automated decision-making system, under an AI Advisory Board.
GUIDANCE	**Vic** A generative AI guideline sets minimum expectations, enforced through codes of conduct and a written-justification mechanism rather than an assessment gate.
FUNDED	**SA** An Office for AI, the first in the country, with $28 million across the 2025-26 budget and 5 full-time staff. It funds agency proof of value projects by application. A Royal Commission into Artificial Intelligence was announced in August 2026, reporting by 1 July 2027.

In most Australian jurisdictions nothing obliges an agency to baseline the process before a proof of concept, or to evaluate the result afterwards.

Several jurisdictions now run a central pool that agencies apply to for a proof of value. The arrangement is a sensible way to begin. It gets work started in agencies that would never have found the money themselves. It also puts the early attempts somewhere they can be seen and compared.

The limit appears at the next stage. The pool funds the build, but not the production service. The central office does not hold the agency's operating budget, and the agency did not budget for something it had not yet proven.

This is a question of maturity rather than design. A jurisdiction standing up its first central AI capability has to start by funding builds, because there is nothing else to fund yet. The arrangements that move a proven build into an agency's recurrent budget come later, and have to be built deliberately. South Australia established the country's first government Office for AI in 2025 and is early in that sequence, as is every jurisdiction that has followed.

In my own engagements one question shows how far along an agency is. Ask at the first workshop who will own the service if it works, and which budget line it will sit in. The answer, or the absence of one, tells you what has to be arranged before the pilot rather than after it.

WHAT FAILURE LOOKS LIKE

- A steering committee with no member who controls the operating budget the service would sit in.
- A state agency that has read the DTA guidance, found it useful, and is not covered by it, with nothing in its own jurisdiction to take its place.
- An evaluation report that reaches the sponsor after the funding round has closed.

01 What decision changes, and who owns it

> *Name the decision or process the system will change, and the person accountable for that decision today. If no such person exists, there is nothing to hand the system to.*

The common framing is the use case. A use case describes what the system will do. It does not name anyone who is worse off if the system never exists. That is why a use case can be enthusiastically supported by a room in which nobody is responsible for anything.

Mistral AI's keynote framing is the iconic use case: strategically valuable, highly urgent, pragmatic and feasible within 6 months. Its own published version of the definition is tighter, requiring a prototype live within weeks and production within 3 months.[20] Those criteria are reasonable. They are also criteria for choosing work, not for making it stick.

The DTA checklist asks the better question. It asks whether the business has been consulted and whether there is "a principal Business Owner involved that can ensure proposed outcomes are validated".[21] Its own Scenario 5 shows what the absence costs. The lesson the DTA draws is blunt: "Technical teams alone cannot carry an AI initiative. Success requires business ownership and clear governance."[22]

The research literature reached the same place 20 years ago. Greenhalgh and colleagues, reviewing the diffusion of innovations in service organisations, separate adoption by individuals from assimilation by an organisation. Assimilation is non-linear and depends on absorptive capacity, slack resources, dedicated funding and leadership.[23] Those are precisely the things a proof of concept scope excludes.

WHAT FAILURE LOOKS LIKE

- A steering committee of enthusiasts, none of whom own a process the system touches.
- A use case written as a description of the technology, so that success means the technology worked.
- An executive sponsor who is sponsoring innovation rather than a decision they are accountable for.

HOW TO CHECK

- Name the decision, determination, assessment or task the system changes. One sentence, in the language the agency already uses for it.
- Name the person accountable for that decision today, by role. If the role does not exist, stop here.
- Confirm that person controls, or can reach, the operating budget the service would sit in.
- Record what that person will do differently once the system works. If nothing changes, the benefit is not real.
- Check the agency's AI use case register for the same problem already solved elsewhere, as the checklist requires.

02 What the process does now

> *Measure the current process before anything is built: volume, time, cost and error rate. Without those numbers nothing measured afterwards means anything.*

This is the cheapest of the 5 decisions and the most often skipped, because the cost comes early and the benefit comes much later.

The consequence is not a vague loss of rigour. It is that the result reverses. METR ran a randomised controlled trial with 16 experienced open source developers across 246 real issues from their own repositories. The developers expected to be 24 per cent faster with AI. They were 19 per cent slower. Afterwards, having been slower, they still believed they had been 20 per cent faster.[24]

ONE RANDOMISED TRIAL, SIXTEEN DEVELOPERS, 246 REAL ISSUES

+24%	+20%	-19%
faster, expected before the trial	faster, believed afterwards	slower, actually measured

The study is small, the participants were maintainers working on codebases they knew intimately, and the models were from early 2025. METR says all of this itself, which is part of why the study is worth citing. The finding that matters survives the caveats. Self-reported productivity gain and measured productivity gain can point in opposite directions, and the people reporting cannot tell.

Australia has already run the natural experiment. The whole-of-government Microsoft 365 Copilot trial measured benefit through self-assessed perception, with self-nominated participants, executives over-represented, no baseline and no control group. The evaluation flagged its own headline figure as an upper bound.[25] The United Kingdom's Department for Work and Pensions ran the same product with a comparison group and regression controls. It measured about 19 minutes a day.[26]

THE SAME PRODUCT, MEASURED TWO WAYS: REPORTED DAILY TIME SAVING

Australia: self-nominated participants, no baseline, no control group, figure flagged by the evaluation as an upper bound (DTA, October 2024). United Kingdom: comparison group and regression controls (DWP, January 2026). The difference is the method, not the product.

The DTA's own recommendation after the trial was that agencies should analyse their workflows to identify use cases. That is the right advice and the sequence is telling. The process analysis arrived after the tool.

WHAT FAILURE LOOKS LIKE

- A satisfaction survey standing in for a measurement, with no number describing the process before the system existed.
- A time saving reported per user per day, multiplied by headcount, with no account of what the saved time was used for.
- A baseline collected after the pilot began, from people who have already used the system.

HOW TO CHECK

- Record volume, elapsed time, effort and error or rework rate for the current process, before the build starts.
- Take the measurement from records the agency already keeps where possible, rather than from asking people.
- Identify a comparison group, even an imperfect one. A team not using the system is worth more than a larger sample of users.
- Write down in advance what result would cause the agency to stop. An evaluation that cannot fail is not one.
- Say plainly what the saved time is expected to be used for, and check afterwards whether it was.

03 Whether the system is any good

> *Score the output against the source, and score whether it is worth acting on, and keep them apart. A demonstration is not an evaluation.*

A grounded system finds material and then writes from it. Either part can fail while the other works, and the answer alone does not tell you which. Scoring retrieval separately from generation is what makes a result actionable, because the 2 failures have different fixes. AccuFind has published the method it uses for this, including the variant for exploratory work where no gold standard can exist.[27]

Two findings explain why you cannot delegate this to a benchmark or to a general impression.

The first is that capability is uneven in ways nobody can predict from the outside. Dell'Acqua and colleagues ran a field experiment with 758 BCG consultants. On tasks inside the model's range, consultants completed 12.2 per cent more tasks and worked 25.1 per cent faster. On a task outside it, consultants using AI performed 19 per cent worse than the control group.[28] No earlier enterprise technology made its users worse at an adjacent task with no warning. It follows that a result on one task tells you very little about the next one, and that evaluation has to be per use case.

The second is that published benchmark scores carry less information than they appear to. Twenty nine expert reviewers examined 445 large language model benchmarks. Over a fifth never defined what they were measuring, and only about 1 in 6 reported any statistical uncertainty.[29] Using a model as the judge does not solve it either. Chance-corrected analysis of 21 model judges shows raw agreement figures overstate true discriminative ability by between 33 and 41 percentage points.[30]

3 POINTS, NOT 5

Give evaluators 5 options and the unsure ones choose the middle. Three force a call: wrong, partly right, right. The pile of threes that a 5-point scale produces is the uncertainty you most wanted to see, rendered invisible.

WHAT FAILURE LOOKS LIKE

- An accuracy figure quoted to one decimal place, from a test set written by the same team after seeing the output.
- A demonstration of ten questions chosen by the people who built the system.
- Two reviewers scoring the same output three and one, and a project that averages them rather than asking why.

HOW TO CHECK

- Score two axes separately: whether the output is supported by the source, and whether it is worth acting on.
- Put the checkable axis on anyone who can read. Reserve subject matter experts for the judgement axis, because they are the scarce resource.
- Have two people score a sample independently and report how often they agreed. Report it even when it is poor.
- Evaluate the use case in front of you. Do not carry a score across from another task or another agency.
- Record the questions the system got wrong, not only the proportion.

04 Build, buy or tune

> *For almost every Australian government use case the answer is buy, or assemble from bought parts. Training or fine-tuning a model is warranted in named circumstances, and they are narrow.*

I disagree with the vendor framing here. Mistral's published method places model customisation at step 2 of 4, immediately after choosing the use case.[31] For a government agency that is the step to skip, and the evidence for skipping it is now reasonably settled.

For knowledge-intensive work, which covers most legal and policy tasks, retrieval beats fine-tuning for getting facts into a system. Ovadia and colleagues found retrieval consistently outperformed continued pre-training for injecting new factual knowledge, and that fine-tuning on new facts can actively degrade performance.[32] Nori and colleagues showed a frontier general model with structured prompting beating a purpose-built domain model with no fine-tuning at all.[33] Building a state-of-the-art legal domain model took 540 billion legal tokens and over 160,000 GPU hours. It finished 2 percentage points ahead of GPT-4 on the benchmark it was built for.[34]

The argument that fine-tuning is simply unaffordable is the wrong argument, and it should not be made. Those large figures describe pre-training. Light-touch fine-tuning is cheap. Work on legal annotation found that between 200 and 1,000 labelled examples were usually enough to beat frontier commercial models on narrow classification tasks.[35] An agency that wants to try it can afford to.

The real cost is the liability, and it recurs. A fine-tune is attached to a base model, and base models are withdrawn on the vendor's schedule rather than the agency's. OpenAI gives generally available models at least 6 months of notice, and runs a shutdown wave on 23 October 2026. From 6 January 2027 its existing customers can no longer create new self-serve fine-tuning jobs.[36] Every base model turn means redoing the work and maintaining an evaluation harness capable of proving the new version is no worse. Continual fine-tuning also degrades what the model could already do, and the effect gets worse as models get larger.[37]

Two Australian constraints close the question further. The Office of the Australian Information Commissioner treats fine-tuning on personal information already held as a secondary use. It says that use will rarely satisfy the reasonable expectations test.[38] And the workforce is not there. Seventy one per cent of APS agencies report critical digital skills shortages, with AI and machine learning among the hardest roles to fill.[39]

Meanwhile the Commonwealth has built the alternative. GovAI, run by Finance, brokers multiple models to agencies. GovAI Chat went from a 5,000-user alpha in April 2026 to a 20,000-user beta in July 2026. A market sounding the same month contemplated around 200,000 users, multi-vendor and onshore.[40]

WHEN TRAINING IS WARRANTED

Three cases, and they are about capability rather than accuracy. First, where the model genuinely cannot do the thing. Low-resource languages are the clean case. Building one took AI Sweden 560,000 GPU hours, and Tilde around 1.5 million on a national supercomputer. Second, narrow, high-volume, repetitive classification where labelled data already exists and the task will not change. Third, where a deployment constraint such as on-premises or edge operation rules out every hosted option. Singapore's HTX is the strongest public sector example of the first case. It is a dedicated science and technology agency, with its own engineering workforce and a national mandate. Very few Australian agencies have that shape.

There is a serious argument on the other side, and it deserves an answer. NVIDIA researchers argue that agentic systems perform a small number of specialised tasks repetitively. Fine-tuned small models are therefore better suited, and more economical than calling a frontier model every time.[41] The argument holds where an agency is already running a high-volume agentic system in production with an evaluation harness around it. An agency with no production AI and no evaluation capability is not that agency. The sequence matters. Optimise a system that is already running, not one that does not exist yet.

Retrieval has its own failure modes, and I should state them. Accuracy degrades when the relevant passage sits in the middle of a long context, a result known well enough to have its own name.[42] There is no universal winner between retrieval and long context. The choice depends on the model, the task and the material. That makes it an engineering decision to evaluate rather than a position to hold.

WHAT FAILURE LOOKS LIKE

- A sovereignty argument used to justify training, in circumstances where deployment architecture would have answered the concern.
- A fine-tune built by a vendor during a pilot, with no agreement about who redoes it when the base model is withdrawn.
- A business case that compares the cost of training against nothing, rather than against the bought alternative.

HOW TO CHECK

- Establish what a frontier model with good retrieval and structured prompting achieves on the task first. That is the baseline every other option has to beat.
- If fine-tuning is proposed, name the capability gap it closes, in one sentence, that retrieval cannot.
- Price the recurring obligation, not the build: who redoes the work on each base model turn, and who maintains the evaluation harness that proves it is no worse.
- Check whether the training data includes personal information the agency holds for another purpose.
- Check the whole-of-government arrangements before building anything. The agency may already have access.

05 What production costs, and who pays

> *Identify the operating budget the service will sit in. Name the person who owns the benefit from a business-as-usual role, and the process change the benefit depends on. Do this before the build, not after it.*

This is the decision that actually stops pilots, and it is rarely a technology decision at all.

The funding arrangements are the main cause. A proof of concept is usually paid for from an innovation line, a central pool or a discretionary budget. A production service needs ongoing money, which needs a costed benefits case, which needs the baseline from Decision 2 and the evaluation from Decision 3. An agency that skipped those has no way to build the case. The work stops for reasons that look like risk aversion and are actually a funding problem.

The DTA checklist asks about this twice, which suggests the DTA knows. Before scaling it asks whether adequate funding has been sought and approved for the pathway to production. It also asks whether there is provision for top-up funding.[43] Asking the question in a checklist does not create an appropriation.

The benefit also has to belong to someone who is not on the project. Australian Government Architecture guidance already says benefits require business process re-engineering, training and redeployment. It also says the benefit owner must sit in a business-as-usual role rather than a project role.[44] That is settled Commonwealth guidance and it long predates AI.

The economics behind it is older still. Brynjolfsson, Rock and Syverson argue that general purpose technologies deliver measured productivity only after a lagged wave of complementary investment. That investment goes into process, skills and organisational structure. They conclude that implementation lag, rather than model capability, is the main cause of the current productivity paradox.[45] Hammer made the operational version of the point in 1990, warning against layering new technology onto an unchanged process.[46]

Two further costs belong in the production case and are usually missing. The first is workforce. Where the work itself changes, consultation obligations under enterprise agreements apply, and they apply before the change, not after it. The second is continuous evaluation. Model behaviour changes while the system is running, with no version change the customer can see. GPT-4 accuracy on one task fell from 84 per cent to 51 per cent between the 2 2023 releases.[47] Evaluation is therefore a standing operating cost.

The DTA's Scenario 6 shows how this fails. A complaint triage model reached 85 per cent accuracy and could not be deployed. The agency's change management process required manual approvals for code changes, and the model needed frequent retraining. The lesson drawn is that AI development and IT modernisation "should be treated as a unified program". That means shared governance, shared resourcing and early agreement on deployment pathways.[48]

WHAT FAILURE LOOKS LIKE

- A successful pilot with no line in any forward estimate, presented to a committee that cannot create one.
- A benefit owned by the project manager, which means it is owned by nobody once the project closes.
- A production business case whose benefits depend on a process change that no one has agreed to make.
- An operating cost that omits evaluation, monitoring and the work of moving to the next base model.

HOW TO CHECK

- Name the operating budget line the service will sit in, and the person who controls it.
- Name the benefit owner by role, and confirm the role is business-as-usual rather than project.
- Write down the process change the benefit depends on, and get the person who owns that process to agree to it in writing.
- Cost the recurring items: evaluation, monitoring, support, and migration to successor models.
- Check the consultation obligations that apply if the work changes, and start them on the right side of the decision.
- Confirm the production pathway before the proof of concept starts. The answer to who pays is cheapest to get wrong at the beginning.

Four things are genuinely different

An argument that AI adoption follows the same playbook as every prior technology wave is easy to overstate. Four differences are real, well evidenced, and worth conceding plainly.

Capability is uneven, and the limits are not visible from outside. The BCG field experiment found consultants using AI on a task outside the model's range performing 19 per cent worse than the control group. Earlier enterprise systems failed predictably. This one fails at a boundary nobody can see from the outside.

Self-reported benefit can be wrong in the opposite direction. The METR trial found participants confidently reporting a gain while they were measurably slower. No prior technology evaluation had to defend against that.

The system you depend on changes without notice. A deployed model's behaviour can change between releases with no version change visible to the customer, and vendors withdraw models on their own schedule. No enterprise resource planning vendor ever altered the behaviour of a live system overnight.

Prompt injection has no reliable fix. The United Kingdom's National Cyber Security Centre position is that "at present, there are no failsafe security measures that will remove this risk". The advice it gives is architectural rather than technical.[49] It remains first in the OWASP top 10 for large language model applications.[50] There is no analogue in the cloud, data warehouse or robotic process automation waves.

Each of those makes evaluation harder, and makes continuous evaluation necessary. None of them removes the need for a baseline, a benefit owner, a funding route or a gate. The technology has new properties. Taking it into production has not changed.

The prior waves are instructive on exactly this point. After roughly 5 years of enterprise robotic process automation, Deloitte's surveys found only 13 per cent of organisations had scaled past 50 automations. They named process fragmentation as the top barrier to scale for 4 consecutive years.[51] That is a survey finding and sits low in the hierarchy set out earlier, so treat it as corroboration rather than proof. The peer-reviewed work on enterprise systems says the same thing more carefully. Implementation failure is an organisational learning problem rather than a configuration problem.[52]

How the 5 decisions map to the DTA guidance

This paper is a companion to the DTA guidance rather than an alternative to it. The guidance supplies the stages, the dimensions and the questions. The 5 decisions are a reading of which questions stop a transition, when they have to be answered, and what evidence closes them.

DECISION	CLOSE IT BY	DTA DIMENSIONS COVERED
1. What decision changes, and who owns it	Before the PoC starts	Business alignment, governance, non-AI alternatives
2. What the process does now	Before the PoC starts	Business alignment, experimentation, people
3. Whether the system is any good	End of PoC, repeated at pilot	Experimentation, data, AI technique selection
4. Build, buy or tune	End of PoC	AI technique selection, architecture, technology, non-AI alternatives
5. What production costs, and who pays	Before the pilot starts	Delivery, scalability, sustainment, people, governance

The timing is where this and ordinary practice diverge.

Decisions 1 and 2 close before the proof of concept starts. That runs against the instinct that a proof of concept is cheap and exploratory, so the paperwork can wait. The baseline in particular cannot be recovered later, because once people have used the system there is no clean measurement of the process without it.

Decision 5 closes before the pilot, not before production. A pilot puts a system in front of real users on real data. Doing that without knowing who will pay for the production service builds an expectation the agency cannot meet. The cost of that is higher than anything the pilot would have returned.

The register is the artefact. The DTA checklist asks whether a decision register has been established. 5 decisions, each with a named owner, the evidence that closes it, and a date, is a register an agency can actually keep. AccuFind publishes a companion workbook alongside this paper.

References

1. Digital Transformation Agency, *Guidance for AI proof of concept to scale: Scenario 5, Gap in business buy-in* (Canberra: Australian Government, 2026). digital.gov.au ↩
2. Mistral AI, *AI transformation*, accessed 11 September 2026. mistral.ai/ai-transformation/ ↩
3. A. Challapally, C. Pease, R. Raskar and P. Chari, *The GenAI Divide: State of AI in Business 2025* (MIT NANDA, July 2025). Not peer reviewed. Report PDF ↩
4. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (Gaithersburg: NIST, 26 January 2023). NIST AI 100-1 ↩
5. Department of Finance, *National framework for the assurance of artificial intelligence in government*, agreed at the Data and Digital Ministers Meeting, 21 June 2024. finance.gov.au ↩
6. Digital Transformation Agency, *AI policy overhauled with new impact assessment tool and procurement guidance*, media release, 2 December 2025. dta.gov.au ↩
7. NSW Government, *Circular DCS-2024-04: Use of artificial intelligence by NSW Government agencies*, effective 1 July 2024, since superseded by DCS-2026-02. arp.nsw.gov.au ↩
8. Australian Centre for Evaluation, Department of the Treasury. The Commonwealth Evaluation Policy is administered by the Department of Finance under RMG 130 and took effect 1 January 2022. evaluation.treasury.gov.au ↩
9. HM Treasury, *Guidance on the impact evaluation of AI interventions*, Magenta Book supplementary guide, first published 17 December 2024, updated 15 May 2026. gov.uk ↩
10. Australian National Audit Office, *Governance of Artificial Intelligence at the Australian Taxation Office*, Auditor-General Report No. 26 of 2024-25 (Canberra: ANAO, February 2025). Model figures reflect the position as at August 2024. anao.gov.au ↩
11. Queensland Audit Office, *Managing ethical risks of artificial intelligence* (Brisbane: QAO, 24 September 2025). qao.qld.gov.au ↩
12. Audit Office of New South Wales, *Internal controls and governance 2025: procurement and technology* (Sydney: Audit Office of NSW, 29 October 2025). audit.nsw.gov.au ↩
13. Department of Finance, *Gateway reviews process*, current as at 2026. finance.gov.au ↩
14. Digital Transformation Agency, *Guidance for AI proof of concept to scale* (Canberra: Australian Government, 2026), released 11 March 2026. The guidance pages carry no on-page date; the date is from the DTA media release New guidance to support AI project success. digital.gov.au ↩
15. Digital Transformation Agency, *Guidance for AI proof of concept to scale: AI transition stages and dimensions*. digital.gov.au ↩
16. Digital Transformation Agency, *Guidance for AI proof of concept to scale: Appendix 1, readiness checklist*. digital.gov.au ↩
17. Digital Transformation Agency, *Guidance for AI proof of concept to scale: context and principles*. digital.gov.au ↩
18. Digital Transformation Agency, *Guidance for AI proof of concept to scale: Appendix 2, AI evaluation*. digital.gov.au ↩
19. Digital Transformation Agency, *Guidance for AI proof of concept to scale: Appendix 3*. digital.gov.au ↩
20. C. Petit and S. Beldo, 'The crucial first step for designing a successful enterprise AI system', *MIT Technology Review*, 2 February 2026. Sponsored content produced by Mistral AI, not written by MIT Technology Review editorial staff. technologyreview.com ↩
21. DTA, *Appendix 1, readiness checklist*, as above. digital.gov.au ↩
22. DTA, *Scenario 5, Gap in business buy-in*, as above. digital.gov.au ↩
23. T. Greenhalgh, G. Robert, F. Macfarlane, P. Bate and O. Kyriakidou, 'Diffusion of innovations in service organizations: systematic review and recommendations', *The Milbank Quarterly* 82, no. 4 (2004): 581-629. Milbank Quarterly ↩
24. METR, *Measuring the impact of early-2025 AI on experienced open-source developer productivity*, 10 July 2025. Randomised controlled trial, n=16. metr.org ↩

25. Digital Transformation Agency, *Evaluation of the whole-of-government Microsoft 365 Copilot trial* (Canberra: Australian Government, October 2024). Trial ran January to June 2024. digital.gov.au ↩

26. UK Department for Work and Pensions, *An evaluation of DWP's Microsoft Copilot 365 trial*, 29 January 2026. gov.uk ↩

27. AccuFind, *Evaluating grounded AI systems*, version 1.0, August 2026. /resources/evaluating-grounded-ai-systems/ ↩

28. F. Dell'Acqua, E. McFowland III, E. Mollick et al., *Navigating the jagged technological frontier*, Harvard Business School Working Paper 24-013, 15 September 2023. Field experiment, n=758. SSRN ↩

29. A. Bean, R. Kearns, A. Romanou and others, *Measuring what Matters: Construct Validity in Large Language Model Benchmarks*, arXiv:2511.04703, 3 November 2025. A systematic review of 445 benchmarks from leading NLP and ML conferences. arXiv ↩

30. J. D. Norman, M. U. Rivera and D. A. Hughes (UC Berkeley School of Information), *Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias*, arXiv:2606.19544, 17 June 2026. arXiv ↩

31. Mistral AI, *AI transformation*, as above. mistral.ai/ai-transformation/ ↩

32. O. Ovadia, M. Brief, M. Mishaeli and O. Elisha, 'Fine-tuning or retrieval? Comparing knowledge injection in LLMs', *Proceedings of EMNLP 2024*. ACL Anthology ↩

33. H. Nori, Y. T. Lee, S. Zhang, E. Horvitz et al., *Can generalist foundation models outcompete special-purpose tuning? Case study in medicine*, arXiv:2311.16452, 28 November 2023. arXiv ↩

34. Equall.ai and others, *SaulLM-54B and SaulLM-141B: scaling up domain adaptation for the legal domain*, NeurIPS 2024 Datasets and Benchmarks. NeurIPS proceedings ↩

35. R. Dominguez-Olmedo, V. Nanda, R. Abebe et al., *Lawma: the power of specialization for legal annotation*, arXiv:2407.16615, first version 23 July 2024, revised 23 April 2025. arXiv ↩

36. OpenAI, *Deprecations*, developer documentation, accessed 11 September 2026. developers.openai.com ↩

37. Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou and Y. Zhang, *An empirical study of catastrophic forgetting in large language models during continual fine-tuning*, arXiv:2308.08747, revised 5 January 2025. arXiv ↩

38. Office of the Australian Information Commissioner, *Guidance on privacy and the use of commercially available AI products*, 21 October 2024, updated 17 January 2025. oaic.gov.au ↩

39. Digital Transformation Agency and Australian Public Service Commission, *APS Digital Workforce Insights Report 2025*, November 2025. Report PDF ↩

40. Department of Finance, *GovAI Chat*, accessed 11 September 2026. Market sounding reported July 2026. govai.gov.au ↩

41. P. Belcak and others (NVIDIA), *Small language models are the future of agentic AI*, arXiv:2506.02153, 2 June 2025, revised 15 September 2025. arXiv ↩

42. N. F. Liu, K. Lin, J. Hewitt et al., 'Lost in the middle: how language models use long contexts', *Transactions of the Association for Computational Linguistics* (2024). arXiv ↩

43. DTA, *Appendix 1, readiness checklist*, as above. digital.gov.au ↩

44. Australian Government Architecture, *Benefits management guides and tools*, Department of Finance and Digital Transformation Agency. architecture.digital.gov.au ↩

45. E. Brynjolfsson, D. Rock and C. Syverson, *Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics*, NBER Working Paper 24001, November 2017. NBER ↩

46. M. Hammer, 'Reengineering work: don't automate, obliterate', *Harvard Business Review*, July-August 1990. hbr.org ↩

47. L. Chen, M. Zaharia and J. Zou, *How is ChatGPT's behavior changing over time?*, arXiv:2307.09009, July 2023, revised 31 October 2023. arXiv ↩

48. Digital Transformation Agency, *Guidance for AI proof of concept to scale: Scenario 6, new AI model and legacy systems*. digital.gov.au ↩

49. UK National Cyber Security Centre, *Thinking about the security of AI systems*, 30 August 2023. ncsc.gov.uk ↩

50. OWASP, *Top 10 for large language model applications*, 2025 edition. genai.owasp.org ↩

51. Deloitte, *Automation with intelligence*, intelligent automation survey, 25 November 2020, and the 2022 edition. Self-reported survey data. deloitte.com ↩

52. D. Robey, J. W. Ross and M.-C. Boudreau, 'Learning to implement enterprise systems: an exploratory study of the dialectics of change', *Journal of Management Information Systems* 19, no. 1 (2002): 17-46. The publisher blocks automated access; the DOI resolves in a browser. DOI ↩

Version 1.0, 11 September 2026. Hamish Cameron, AccuFind Pty Ltd. Published as a companion to the DTA's Guidance for AI proof of concept to scale. The 5 decisions and the companion register are free to reuse.