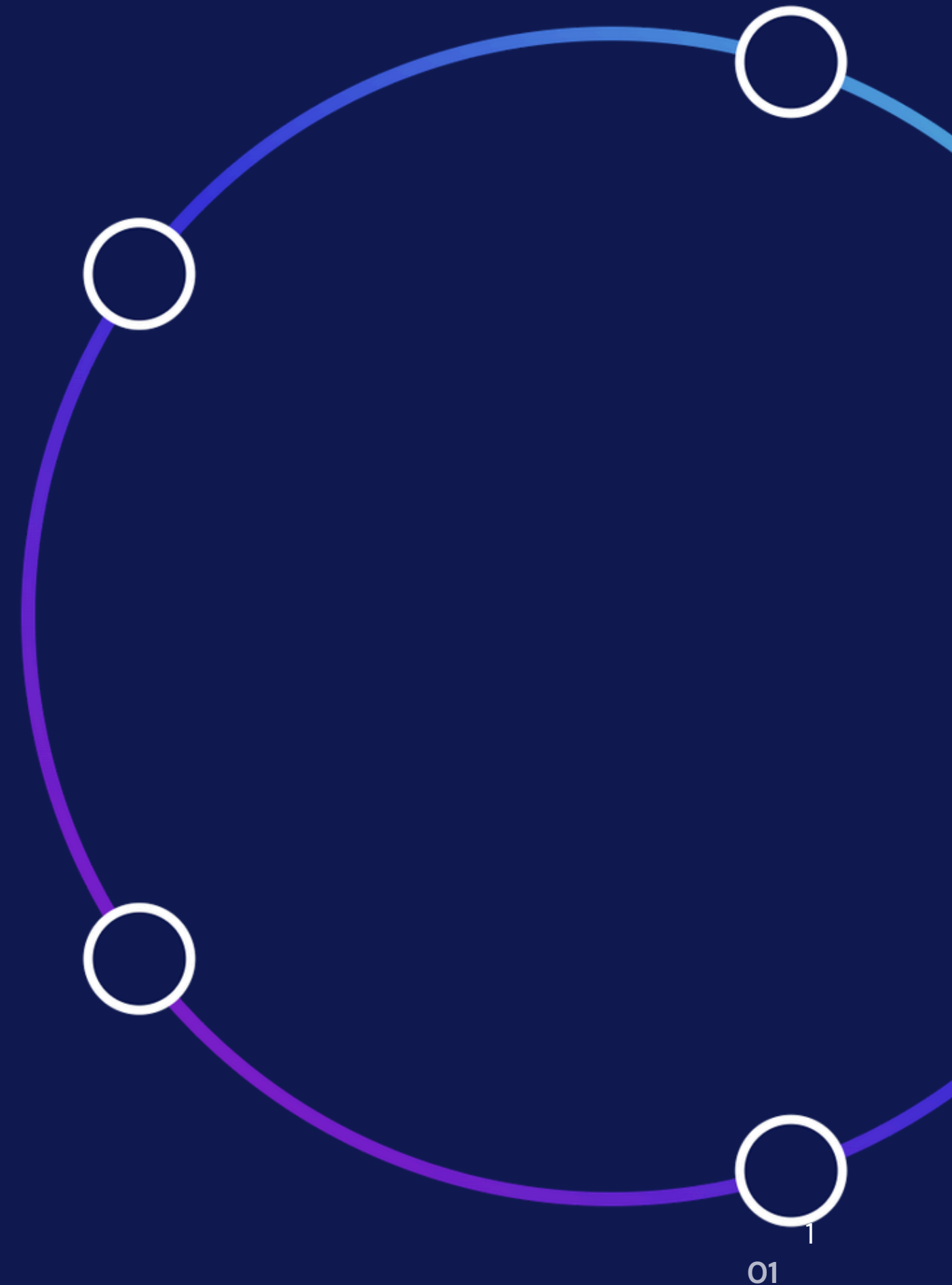


“Should we scale it?”

A practical guide for deciding whether an **AI PoC** is actually worth investing in.

AN AVIAN FRAMEWORK



A proof of concept should end in a **decision**, not a demo.



AI proofs of concept usually start with the best of intentions. Your team gets access to a new tool, or someone spots a task AI might be able to do, or you've been told to "start using AI".

Then the fun bit starts: you get to build something. And somewhere towards the end of that process, your exec asks whether it worked.

Instead, we need to start by designing the evaluation itself.

An AI PoC is an experiment, not the first phase of implementation. So it has to be designed like an experiment. Its purpose is to generate enough evidence to decide whether changing the way people work is actually worthwhile, and then to **scale, iterate or stop**.

"The AI can do it" is not the same as "this is worth it."

WHAT THIS FRAMEWORK IS FOR

An investment decision lens for AI-enabled workflow PoCs.

It is designed for proofs of concept where AI can change how people work. It should help you look at the business case, product and workflow, and set you up to answer one question.

Does this new way of working create enough value to justify the change?

WHAT IT IS NOT

It is not a replacement for:

AI assurance or impact assessment

security and privacy assessment

technical model validation

procurement or architecture review

regulatory compliance

ongoing monitoring once a system is deployed

It should sit alongside those disciplines.

For Australian Government teams, it complements the DTA's **Guidance for AI proof of concept to scale** and the **AI technical standard**.

Five questions before you scale AI



01	Is it the right PoC? Start with dissatisfaction, not AI.	PAGE <u>6</u>	⇒
02	What does it have to beat? No baseline, no benefit claim.	PAGE <u>8</u>	⇒
03	What would make this genuinely better? AI doesn't get a quality discount.	PAGE <u>10</u>	⇒
04	Does it hold up in real work? Test the work, not the demo.	PAGE <u>14</u>	⇒
05	Is the evidence good enough to change how we work? Scale, iterate or stop.	PAGE <u>20</u>	⇒
Seven guiding principles THE MINDSET BEHIND THE FRAMEWORK		PAGE <u>23</u>	⇒

PART ONE

The five questions

Five questions to guide your AI PoC, from opportunity to decision.

Each question comes with a model you can use in a workshop with your team.



01 QUESTION ONE

Is it the right PoC?

The existence of an AI capability is not a reason, by itself, to change a workflow. Start with what **isn't working** today: what is too **slow**, **expensive**, **inconsistent** or **poor quality**, how often it happens and who it affects.

Because AI PoCs aren't free.

And scaling them certainly isn't! Even when they offer savings, a new way of working itself carries costs in design, integration, governance, training, adoption and ongoing oversight. **A technically viable use case can still be a poor investment.**

AI capability is not business value.

THE PRINCIPLE

Start with
dissatisfaction,
not AI.

QUESTIONS TO ANSWER

What are we dissatisfied with?

How often does the work occur?

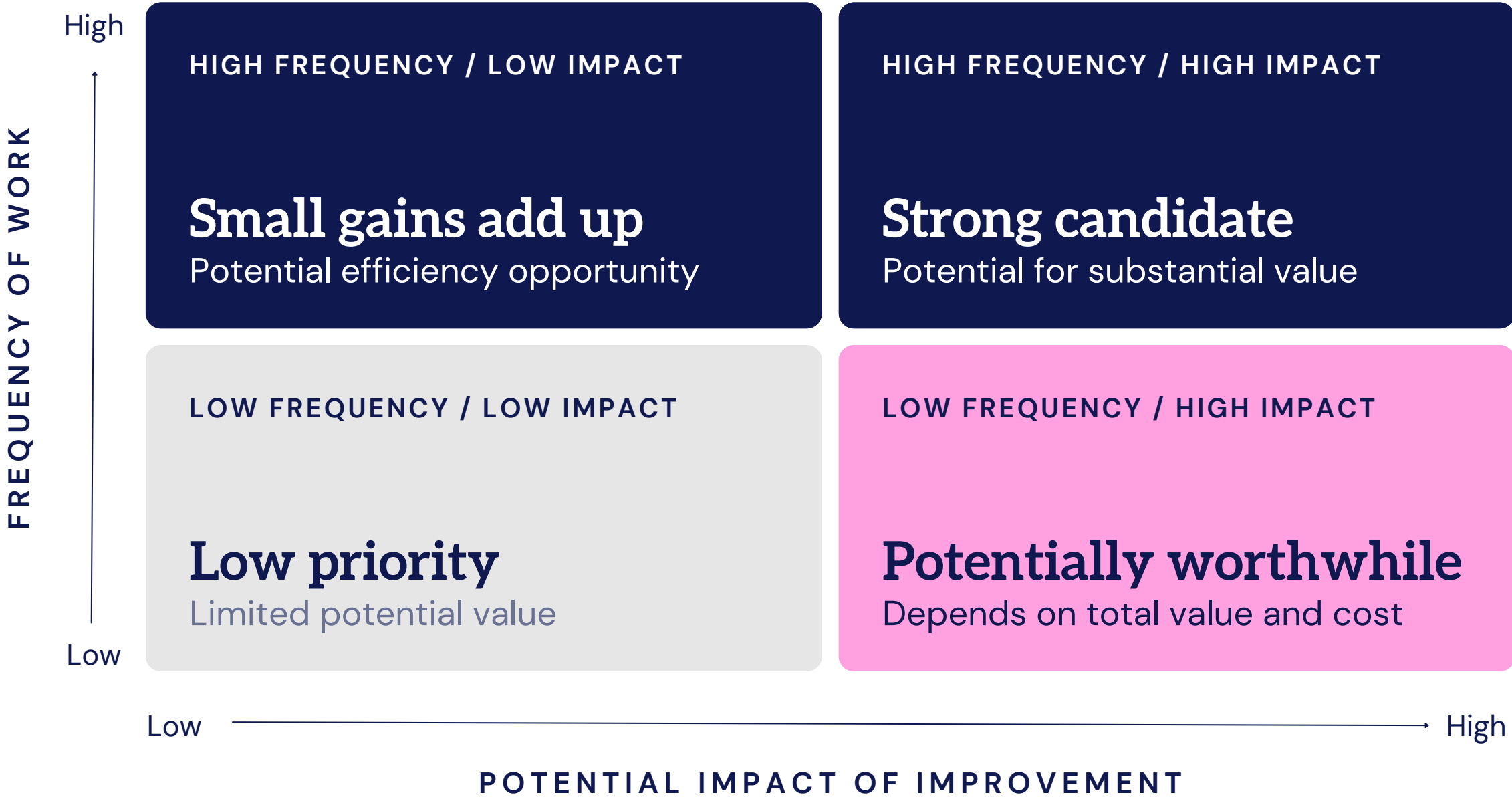
What would improving it be worth?

How difficult would it be to change?

What happens when the work goes wrong?

Is this worth testing?

Start with the opportunity: how often does the work happen, and how much value could improving it create?



TWO MORE CONSIDERATIONS

1. COST OF CHANGE

Consider implementation, integration, training and ongoing operating costs.
Does the potential benefit justify the investment?

2. CONSEQUENCE OF FAILURE

Consider the consequences of failure, and what testing, controls and assurance would be required.
High consequences raise the evidence bar, not necessarily rule out the use case.

No baseline? No benefit.

~~“People think it feels faster.”~~

“We have evidence that this way of working is materially better.”

02 QUESTION TWO

What does it have to beat?

If you don't understand how the work performs today, **you can't credibly show that AI improved it.**

Your baseline may not be simply how long the existing process takes. You can also look at what the work achieves, then how good it is, what it takes to produce and where it goes wrong.

Of course, sometimes that data doesn't exist.

That's not a reason to skip the baseline; without it, you will struggle to make an evidence-based investment decision.

Measuring and understanding the current state may need to be the first phase of the PoC.

Before you test AI, measure what happens today.

You need a baseline to compare the PoC against. Use existing operational data where it exists. Where it doesn't, collect it first: timing studies, SME scoring, surveys, user research or workflow observation.

Start with the reason the work exists, then measure how well the current workflow delivers it. Add measures such as consistency, accessibility, throughput, risk or compliance where they matter.

If you can't measure the current state, you can't credibly claim the new one is better.

OUTCOME	QUALITY	EFFORT	COST	FAILURE
What happens for the user or organisation today?	How good is the output today?	How much human work does it take?	What resources does it consume?	Where and how does it go wrong?
MEASURE	MEASURE	MEASURE	MEASURE	MEASURE
Task completion	SME score / qualitative assessment	Minutes or hours per task	Software and licensing	Error rate
User success	Completeness	Rework time	Labour cost per task	Rework rate
Business outcome	Accuracy and error rate	People involved	External spend	Escalation rate
Satisfaction	Consistency	Number of handoffs	Infrastructure	Severity of failures
Complaints and escalations				Detectability
MEASURE WITH	MEASURE WITH	MEASURE WITH	MEASURE WITH	MEASURE WITH
Analytics, service data, surveys, user testing	Agreed rubric, Likert scoring, expert review, sample audit	Time study, workflow observation, staff logs and interviews	Finance data, plus time × loaded labour cost	Incident data, sample review, SME assessment, workflow logs

Use numbers where numbers make sense and structured judgement where they don't.

03 QUESTION THREE

What would make this genuinely better?

Define success **before** you see what the AI can do. The acceptable quality threshold is set independently of the technology.

THE PRINCIPLE

AI doesn't get a quality discount.

A. WHAT IS NON-NEGOTIABLE?

What are your non-negotiable thresholds: accuracy, safety, compliance, privacy, accessibility, trust?

If the new workflow was more efficient, but wrong 50% of the time, would that be tenable? What about 25% of the time? 10%?

You **must be clear** on these thresholds before starting your PoC.

B. WHAT IS 'VALUE'?

What does 'value' mean for your organisation? Sometimes the primary driver is increasing profit. For other organisations, it is increasing accuracy or uptake.

Not **everything** has to improve (and it likely won't). Which measures matter most, and how should they improve?

This, balanced with your non-negotiable thresholds, will help you identify whether the overall proposition is 'worth it.'

WHAT WOULD MAKE US SCALE, ITERATE OR STOP?

Before testing, agree what data you would need to see to scale the AI use case, what would show you that it's time to iterate, and what would tell you to stop.

Define it now so the goalposts can't move later.

If your business case only works
because you've lowered the definition
of “good enough”

...you probably don't have a strong AI use case.

Clear the bar. Then prove the value.

An AI-enabled workflow has to do two things: stay acceptable on the non-negotiables, and create enough value somewhere to justify changing the way you work.

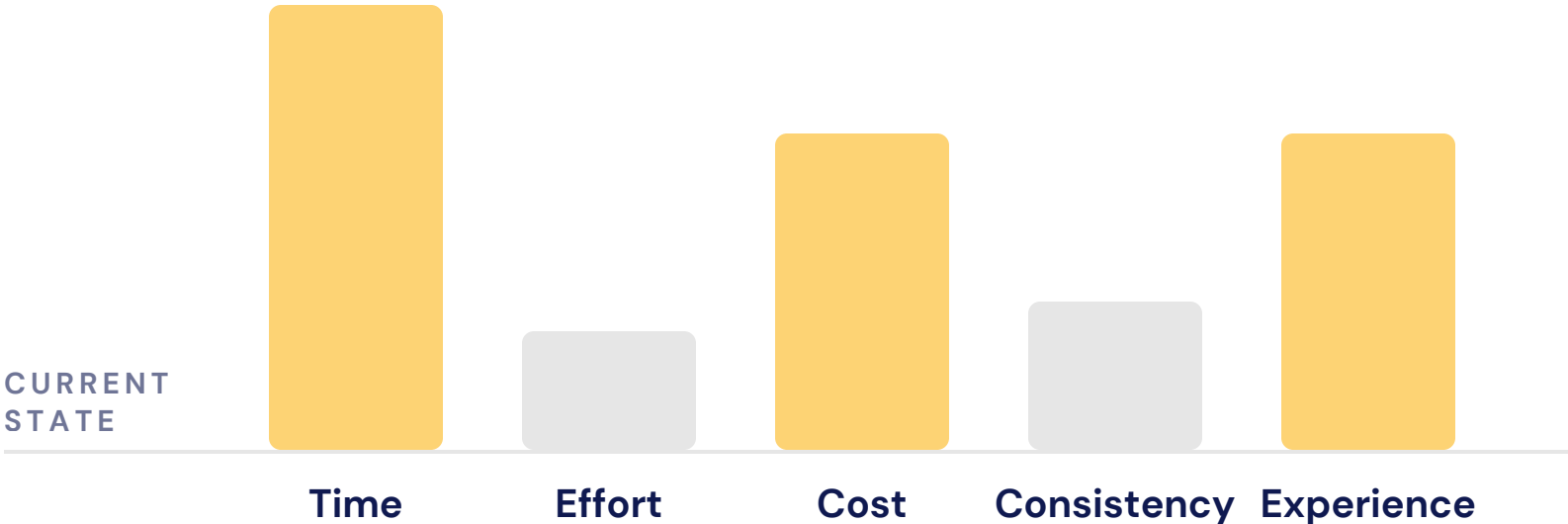
1. MUST CLEAR THE THRESHOLD

These are the conditions the new workflow cannot trade away for speed or cost.



2. MUST MEASURABLY IMPROVE

Not everything needs to improve. But the overall proposition needs enough uplift somewhere to make the change worthwhile.



Define the evidence you need before you test.

Set clear pass/fail criteria for every non-negotiable. Agree how you'll measure improvements in value, and what would count as enough to justify the change.

Turn the criteria into decision rules.

Define the minimum threshold

Measure	Assessment method	Pass condition
Accuracy	SME review, agreed rubric	4/5 or better, no critical errors
Compliance	Check against mandatory requirements	100% met
Safety	Review of critical failure modes	Zero unacceptable failures
Quality	Independent reviewer scoring	Agreed minimum score or above

These are **gates**. If a non-negotiable fails, the PoC has not met the minimum standard.

Use numbers where numbers make sense. Use structured expert judgement where they don't.

For every criterion, define what you will measure, how you will assess it, and what result will count as success.

Define the value target

Measure	Baseline	Target
Time	6 hours per outcome	2 hours or less
Human effort	Current handling time	50% reduction or more
Cost	Current cost per task	Agreed target reduction
Consistency	Current variation	Measurable improvement

These are the value measures. The PoC does not need to improve all of them, but the improvement needs to be material enough to justify change.

A Likert scale, expert rubric, pass/fail requirement or user test can all be valid measures. What matters is that the method and decision rule are defined before the results arrive.
Example values are illustrative, not recommended thresholds.

04 QUESTION FOUR

Does it hold up in real work?

A handful of well-chosen prompts can make almost anything look impressive.

That is not evidence it will **create value** in normal operations.

Test two things separately:

- **Representative work:** How does it perform on the tasks people actually do?
- **Critical failures:** What happens in the rare cases where getting it wrong really matters?

Report these separately. Don't let strong average performance conceal critical failures.

THE PRINCIPLE

Test the **work**, not the demo.

1 · REPRESENTATIVE PERFORMANCE

What happens during normal operations?

Real work, in the proportions it occurs. **Tells you about value.**

2 · STRESS AND CRITICAL FAILURE

What happens in rare but consequential situations?

The high-risk cases, even if rare. **Tells you about risk and controls.**

Start with the outcome.

Evaluate from the top down.

A technically impressive model can still create a poor workflow, a worse output or no meaningful business value.

Model performance is evidence. It is not the outcome.

EVALUATE FROM THE TOP DOWN



Where did the work go?

Many AI-enabled workflows don't just reduce human effort, but also push it into other (often unmeasured) parts of the process or organisation. This is a key impact to watch for in evaluating AI PoCs. If AI pushes work downstream, or onto other teams, that also needs to be measured.



Measure the difference in total human effort between workflows, not model generation time.

Compare total human effort before and after. Include work removed, retained and created by AI, while ensuring the new workflow still meets the quality threshold.

WHY OVERSIGHT IS PART OF THE TEST

Human in the loop is not the same as human oversight.

Most AI-enabled workflows keep a person in the loop: someone reviews the output before it goes anywhere. That step is usually treated as the control that ensures the workflow is reliable and safe.

But... it is only a control if the reviewer can actually catch failures that matter. Once a system is trusted, checking gets lighter, attention drifts and the approval becomes a formality.

Automation research has known this for decades: **the more reliable the system, the harder it is for the person to stay engaged.**

So the PoC has to **test** the oversight rather than **assume** it. The next two pages examine what supports meaningful review, and the failure modes that reviewers need to be able to detect.

An approval step is not a control just because a person clicks approve.

THE RISK

Trusted systems are checked least.

QUESTIONS TO ANSWER

Who reviews the output, and what do they actually see?

Do they have the evidence and context to judge it?

What happens when they disagree with it?

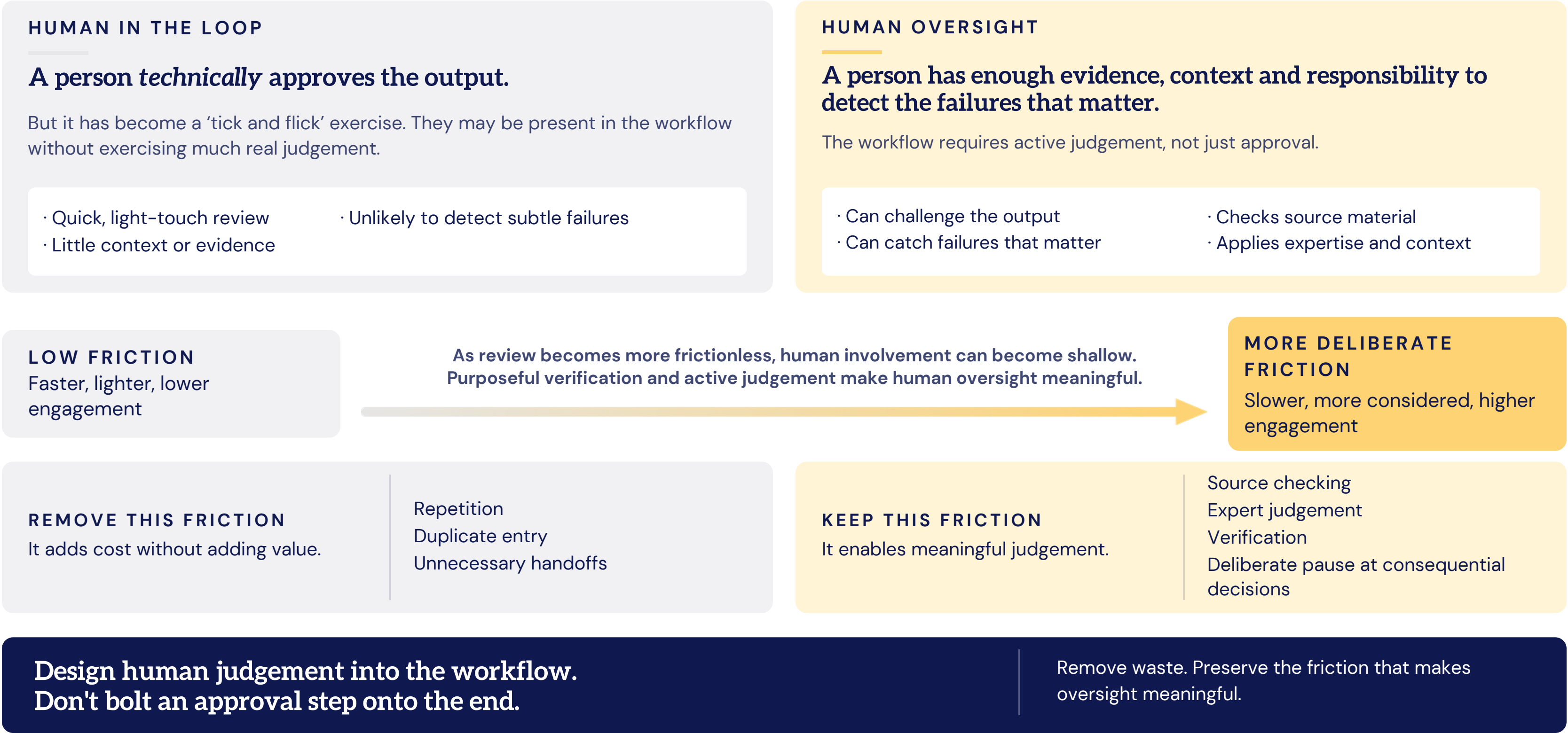
Did they catch the errors we seeded?

Does the checking hold up at real volume, over time?

Draws on Bainbridge's *Ironies of Automation* and research on meaningful human oversight.

Don't optimise away **useful effort**.

Some human effort is waste. But sometimes, a bit of friction is what keeps people engaged enough to notice when something is wrong.



Not all errors are alike

An overall error rate doesn't tell the whole story. Not all errors or failure modes will have the same impact on your customers or your organisation.

Failure needs to be defined in context. A common minor mistake and a rare catastrophic one are not equivalent.

A **rare** failure that is serious, hard to detect and difficult to recover from may matter more than a **frequent** but obvious minor error.

For each meaningful failure mode, ask:

FREQUENCY

How often could it happen?

Is it occasional, common or systemic?

SEVERITY

How serious would the consequence be?

Is the impact trivial, material or high-stakes?

DETECTABILITY

How likely is someone to notice?

Would a reviewer reliably catch it?

RECOVERABILITY

How easily can it be corrected?

Can the workflow recover before harm is done?

Use this lens to decide which failures are tolerable, which require stronger controls, and which make the workflow unacceptable.

Adapted from failure-mode and risk analysis practice (FMEA), applied here to AI-enabled workflows.

THE PRINCIPLE

The PoC
should end
in a **decision**,
not a demo.

05

QUESTION FIVE

Is the evidence good enough to change how we work?

The final question is not “did the AI work?”. It is whether there is enough evidence of value to justify changing how the organisation works.

You agreed the decision rules before testing. Now apply them to what the PoC produced.

WEIGH THE EVIDENCE

Demonstrated business or user value

Quality thresholds held

Total workflow effort

Failure risk and controls

AGAINST WHAT ADOPTION TAKES

Implementation and integration

Governance requirements

Adoption and capability

Ongoing operating cost

Choose deliberately

SCALE

The evidence **supports** broader adoption.

There is material value and the benefits justify the cost, risk and effort of change. Scaling may mean moving into a controlled pilot, rather than immediately deploying into production.

ITERATE

There is evidence of value, but the approach **needs work.**

Change the workflow, scope, controls, use case, tooling or implementation approach and test again.

STOP

The evidence **does not** justify further investment.

The value may be too small, the cost too high, the quality insufficient or the risk unacceptable.

A stop decision is a successful outcome of a good experiment.

Avoiding a poor investment creates value too.

The principles

The five questions guide the process. But these seven positions are what make the evaluation effective.

Seven principles for evaluating AI-enabled work

01

Start with dissatisfaction, not AI.

AI capability is not business value.

02

**No baseline?
No benefit.**

If you don't understand the current state, you cannot credibly claim improvement.

03

AI doesn't get a quality discount.

The acceptable threshold is set independently of the technology.

04

Test the work, not the demo.

Typical work dominates value testing. Consequential failures get their own stress test.

05

Evaluate the work, not just the model.

Outcome, then output, workflow and technology.

06

Count the work AI moves, not what it removes.

Measure total workflow effort, not generation time.

07

Don't optimise away useful friction.

Oversight only works when people still exercise meaningful judgement.

**And...
be willing to stop.**

A PoC that tells you to stop **has** created value.

IN PRACTICE

~6 hours
→ ~30 min

to a workable first draft

Includes AI generation plus content designer review and editing, not just first-draft output.

READ MORE AT [AVIAN.AU](https://avian.au)

[Case study: building and evaluating AI-assisted content workflows →](#)

We use this approach in practice

In a recent AI-assisted content PoC with an Australian Government regulator, we started with the work, not the technology.

We established the baseline, defined quality criteria with the content team and measured the complete human + AI workflow.

QUALITY FINDINGS

4.3 / 5 Free of substantive errors or hallucinations

3.0 / 5 Readability at the target reading level

EVIDENCE SHOWED:

- 1 The AI produced accurate, comprehensive first drafts quickly.
- 2 Outputs still needed expert editing for readability, narrative structure and regulatory referencing.
- 3 The value was not just faster drafting. It was understanding where AI helped, and where human expertise remained essential.

The PoC evidence showed that AI-assisted content could materially reduce effort for structured content, provided the workflow retained expert review and quality controls.



Evidence base and related guidance

Built from our delivery experience and established product, evaluation, risk and human–automation practice. Government guidance covers the wider lifecycle. This framework adds the work–level lens inside it.

[Australian Government, Guidance for AI proof of concept to scale →](#)

[Australian Government, Technical standard for government's use of AI →](#)

[Commonwealth Evaluation Policy →](#)

[NIST, AI Risk Management Framework →](#)

[Bainbridge \(1983\), Ironies of automation →](#)

[Zhu, Lu, Ming, Lee and Wang, Designing meaningful human oversight in AI →](#)

USE THE MODELS

Every model in this publication works as a workshop canvas.

Use–case screen	07
Baseline card	09
Success criteria	12
Representative and stress testing	14
Evaluation hierarchy	15
Work removed, work moved	16
Failure lens	19
Scale, iterate, stop	21

Want to run them with your team? [avian.au →](#)

We help organisations gather **evidence** before **investment**.

From evaluating AI proofs of concept to designing better digital services, we bring research, product thinking and practical delivery together to help teams make better decisions.

**Have an AI PoC underway?
Let's talk. →**

info@avian.au
avian.au

The Avian AI PoC Evaluation Framework

© Avian Services Pty Ltd 2026.

Please attribute the framework when you reuse it.

AVIAN

info@avian.au

avian.au