MASTERCLASS SERIES · WEEK 2 OF 3 · 29 OCTOBER 2026

Designing Intentional AI Experiments

Measure what matters.

Mike Davis

Founder & CEO, Shackleton Labs

What has stayed with you since Week 1?

Scan with
your phone
to participate.

Insights, reflections or questions. Share aloud or in the chat.

01

An insight

What changed how you think about innovation or AI?

02

A reflection

What feels relevant to your organisation and the people you serve?

03

A question

What would you like to revisit or work through today?

10 minutes together. No preparation required; listening is welcome.

From an opportunity to a decision-ready experiment.

A practical session for boards and leaders in the for-purpose sector.

01

Choose a problem

Find a useful, mission-aligned starting point.

02

Design the test

Set the scope, people, time and safeguards.

03

Measure what matters

Agree evidence and stop / continue / scale decisions.

04

Work it through

Work through Harbour's six-month chatbot case.

ROI from innovation programs

IMPACT RETURN

5:1

Well-designed innovation programs in the social sector demonstrate a social return on investment of 5:1 — every $1 invested generates $5 in measurable social value.

GROWTH RETURN

+20% revenue

A well-structured innovation program can increase organisational revenue by up to 20%. Innovation-active organisations consistently outgrow their peers in financial sustainability.

GOVERNANCE RETURN

2× EBIT impact

Organisations with CEO-level oversight of AI governance are twice as likely to report meaningful EBIT impact from AI — governance is the single strongest predictor of return.

CULTURE RETURN

Up to 59% less turnover

Innovation cultures see up to 59% lower staff turnover. In Australia, replacing one employee costs 50–200% of their annual salary. Disengaged workplaces cost the global economy $438 billion every year.

[1] New Economics Foundation, SROI Guide, 2012 · [2] Project ROI Study, Firefly Giving, 2024 · [3] McKinsey, State of AI Trust, 2026 · [4] McKinsey, The Impact of Agility, 2021 · [5] Gallup, State of the Global Workplace, 2024 · [6] AHRI / ScaleSuite, 2025

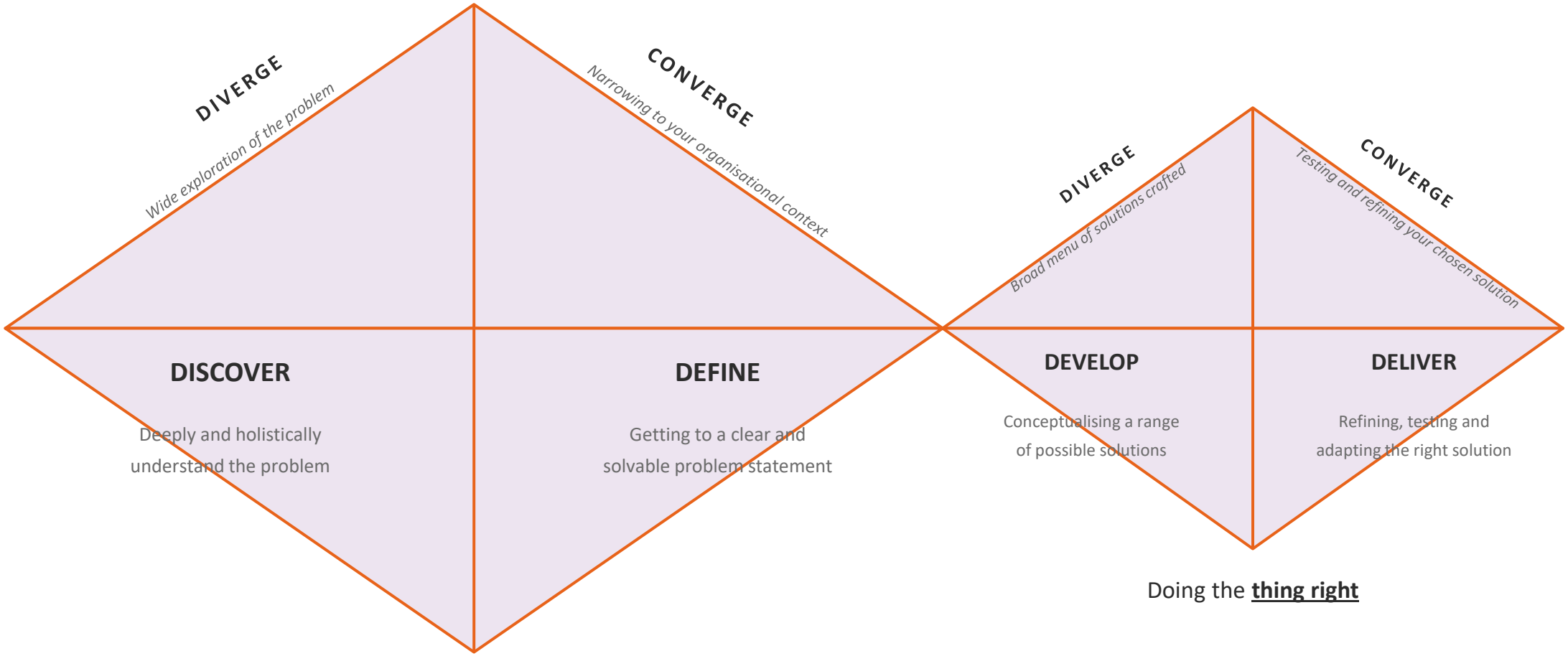
CHALLENGE SPACE

Spending most of our time to identify and clarify the problem to be solved

SOLVE SPACE

Here is where we stress test, learn, adapt and refine our solution to the problem

70% — Challenge Space | **30% — Solve Space**



Doing the **right thing**

Doing the **thing right**

Framing problem statements well is mission critical

AI PROJECT FAILURE

80%+

More than 80% of AI projects fail, and RAND's research found the number one cause isn't the technology. It's misunderstanding the problem being solved.

PROBLEM DIAGNOSIS

85%

of C-suite executives say their organisations are bad at problem diagnosis, and 87% say this carries significant costs. Leaders jump into solution mode before they understand the problem.

AI PILOT RETURNS

95%

of enterprise generative AI pilots delivered no measurable return. The cause wasn't model quality. It was tools that never fitted the real work they were meant to change.

[1] RAND Corporation, The Root Causes of Failure for Artificial Intelligence Projects, 2024 · [2] T. Wedell-Wedellsborg, Are You Solving the Right Problems?, Harvard Business Review, 2017 · [3] MIT NANDA, The GenAI Divide: State of AI in Business, 2025

PART ONE

Choose the right problem.

Start with a meaningful need, not a tool looking for something to do.

Harbour Community Hub: start with the challenge.

Following the Double Diamond: discover the context, then define the problem before choosing a solution.

DISCOVER · CONTEXT

A small community hub

Harbour connects local people with community programs, appointments and support services. A small team balances routine enquiries with face-to-face support and follow-up.

DISCOVER · FRICTION

Repeat admin crowds out people

About 60 clients attend each week with simple queries: booking steps, opening hours and contact details. Repeating these answers leaves less staff capacity for clients needing more complex help.

DEFINE · WHY THIS?

A manageable starting point

Routine questions can use approved public facts. Test an optional website chatbot for six months, while keeping phone and walk-in access for anyone who needs human help.

Defined challenge: reduce time on repeat enquiries and free staff for complex human help.

What makes a good first challenge for AI?

Choose a task that releases capacity for higher-value work, without handing over human judgement.

01

Repeated pattern

Similar inputs and steps recur across many cases, rather than a one-off task.

02

Meaningful workload

The time and resource burden is big enough to justify setup, review and running costs.

03

Lower-value task work

Routine admin consumes time better spent on client support, judgement or human connection.

04

Clear information

Approved information is available; the task and a good output can be defined clearly.

05

Safe, reviewable scope

Start small. Protect sensitive data, keep human checks and make errors easy to catch.

06

Measurable benefit

Compare time, quality and total cost with the current process; agree success before testing.

Repetition alone does not justify AI. Compare a better process, template or rules-based automation.

How to prioritise: The Innovation Rubric

Score each idea 1–5. Add a fifth criterion that matters to your organisation.

IMPACT

Does it meaningfully improve outcomes for those we serve?

→ Addresses a significant unmet need
→ Benefits are measurable and observable
→ Potential for sector-wide influence

Score: ___ / 5

FEASIBILITY

Can we actually do this with what we have?

→ Realistic resource requirements
→ Skills and capability exist or can be built
→ Timeline is achievable

Score: ___ / 5

ALIGNMENT

Does it fit our mission, strategy and culture?

→ Connected to strategic priorities
→ Consistent with values and brand
→ Has leadership commitment

Score: ___ / 5

SCALABILITY

Could this grow, replicate or attract funding?

→ Evidence base can be built
→ Replicable in other contexts
→ Fundable or revenue-generating

Score: ___ / 5

YOUR CRITERION

?

Score: ___ / 5

What separates pilots that work from pilots that die

Four factors show up in almost every successful AI and innovation pilot. The board's job is to test whether they are actually present before saying yes.

CLEAR PROBLEM

Anchored to a real problem

The pilot starts from a specific, measurable user or operational problem — not from a technology looking for a use.

BOARD QUESTION

"What specific problem does this solve — and who told us it was a problem?"

BOUNDED SCOPE

Small, time-boxed, decision-ready

The pilot is deliberately small, time-boxed, and ends at a clear go / iterate / stop decision point at a set date.

BOARD QUESTION

"What exactly will we decide at the end of this pilot — and how long is the runway?"

READY FOUNDATIONS

Data, policy and capability in place

Data quality, AI policy, governance and staff capability are in place — or explicitly scoped as a Phase 0 before build.

BOARD QUESTION

"Do we have the data, policy and skills to run this responsibly — or is that Phase 0?"

CLEAR OWNERSHIP

One owner, one sponsor, one metric

One accountable delivery owner, one executive sponsor, one budget line, and one agreed measure of success from day one.

BOARD QUESTION

"Who owns delivery, who is the exec sponsor, and how will we know it worked?"

If a proposal can't answer all four questions with confidence, it isn't ready for board approval — it's ready for more thinking.

PART TWO

Design a bounded experiment.

One question, a small test and a clear decision point.

Can Harbour shift routine enquiries online, responsibly?

Fictional case: an optional AI chatbot on the Hub website, with a clear route to human help.

Over six months, can Harbour’s website chatbot answer routine questions accurately, reduce simple-query attendance, and free more staff time for clients with complex needs?

Accuracy: ≥90% of routine answers correct against human-approved answers.

Clients attending: 30% reduction in clients with simple queries.

Staff focus: 30% more staff time spent on clients with more complex needs.

A six-month website pilot with clear boundaries.

Fictional plan: approved FAQs, live website use and human handoff. Baseline and testing come before launch.

01

Problem + people

Simple enquiries compete with complex human support.

02

Hypothesis

Accuracy, simple-query attendance and staff time on complex needs.

03

Scope + comparison

Website FAQs; human support stays available.

04

People + safeguards

Service lead owns it; staff audit and handle handoffs.

05

Evidence + limits

Six months; budget and staffing agreed pre-launch.

06

Decision

Monthly checks; reviews at months 1, 3 and 6.

Answer routine questions. Make human help easy.

Illustrative public-facing exchange. The bot identifies itself as AI and uses approved FAQs.

WEBSITE ENQUIRY

A routine booking question

"Can I book a Tuesday appointment? I may need an interpreter."

APPROVED FAQ

Two facts, no assumptions

Tuesday bookings are made by phone. Interpreter requests must be confirmed by staff. No eligibility rules are supplied.

CHATBOT + HANDOFF

A bounded answer

"Tuesday bookings are made by phone. Please contact our team to confirm your interpreter request."

Offer human contact clearly. Hand off unclear, personal or complex requests; never invent eligibility.

Six months live, with checkpoints along the way.

The six-month live trial starts after baseline collection, offline testing and launch approval.

Before launch	Four-week baseline; approved FAQs; accuracy, safety and access checks.
Months 1–2	Launch an FAQ chatbot; keep phone, walk-in and staff handoff.
Months 3–5	Audit answers monthly; review visits, handoffs, costs and incidents.
Month 6	Compare outcomes with baseline and decide: stop, refine or expand.

Formal reviews at months 1, 3 and 6. Unsafe behaviour can trigger a pause at any time.

Give the experiment an owner, not just a tool.

Fictional roles. Use equivalent responsibilities within your organisation's existing delegations.

DELIVERY OWNER

Service lead

Owns the chatbot, approved FAQs, monitoring and incident response.

EXECUTIVE SPONSOR

CEO

Approves the six-month scope, staffing, budget and review checkpoints.

STAFF + REVIEWER

Staff + checker

Staff audit answers against human references and support complex or escalated enquiries.

DECISION-MAKER

CEO + board oversight

CEO decides within delegation. The board oversees risk, outcomes and any larger expansion.

PART THREE

Measure what matters.

Choose evidence that supports a decision, not a dashboard full of activity.

Accuracy, attendance and staff focus.

Measure the ROI of these outcomes and include it in your early-stage business case.

Three six-month targets, plus a separate check on clients attending for complex human help.

ANSWER ACCURACY

Target: at least 90%

Audited routine answers judged correct against human-approved reference answers.

CLIENTS ATTENDING

Target: 30% fewer

Clients attending with simple queries each week, compared with baseline.

STAFF FOCUS

Target: 30% more

Weekly staff hours spent on clients with complex needs, compared with baseline.

COMPLEX ATTENDANCE

More clients helped

Clients attending for complex human support. Track separately from staff time.

Keep human access open; monitor safety, costs, waiting times and whether people’s needs are met.

Set the baseline before the chatbot goes live.

Define enquiry types and measures consistently. All Harbour figures remain fictional teaching assumptions.

Attendance + time	Four weeks of client attendance and staff hours on complex needs.
Human benchmark	Staff approve reference answers; use these to assess routine chatbot replies.
Other influences	Log changes in opening hours, outreach, service capacity and local demand.
Evidence owner	Service lead tracks attendance, staff hours and costs; reviewers audit replies.

Compare weekly rates and the final four-week average, not unequal totals across time periods.

Set a baseline for each service target.

Illustrative baseline only. These numbers are invented, not results from an actual organisation.

SIMPLE-QUERY CLIENTS

60 clients / week

Clients attending with simple queries. Target: 42 per week or fewer, a 30% reduction.

COMPLEX HUMAN HELP

20 clients / week

Clients attending for complex help. Track separately: attendance is not staff time.

STAFF TIME

40 hours / week

Team time on complex client needs. Target: 52 hours per week or more, a 30% increase.

Compare baseline and final four-week averages. More clients attending does not prove more staff time.

Set stop / continue / scale rules before the test.

Agree the criteria before launch. The 90% accuracy target does not excuse unsafe advice.

STOP

Unsafe or not worthwhile

Pause for unsafe advice or a data incident. Restrict or pause affected functions if accuracy falls below 90%.

CONTINUE

Promising, but unclear

Safe and ≥90% accurate, but either 30% target is unmet or unclear: refine within agreed limits and review again.

SCALE CAREFULLY

Evidence meets the gates

At month 6: ≥90% correct; ≥30% fewer simple-query clients; ≥30% more staff time on complex needs; safe human access.

Safety overrides averages. Keep human support available during any pause; expansion needs approval.

PART FOUR

Guardrails that enable learning.

Make the safe path clear enough for people to use it.

A public-facing chatbot needs visible safeguards.

Fictional trial boundaries. Test before launch, monitor live use and make human contact easy.

DATA

Use approved inputs

Approved FAQs; avoid collecting sensitive details.

TOOLS

Check the settings

Access, retention and vendor settings approved.

PEOPLE

Offer human handoff

Keep phone and walk-in support; audit bot replies.

USE

Define exclusions

No eligibility, legal, clinical or crisis advice.

INCIDENTS

Make pausing easy

Staff can pause the bot; alert the CEO.

OVERSIGHT

Review at the gate

Monthly audits; reviews at months 1, 3 and 6.

Approve a six-month, supervised website pilot.

Fictional proposal: limited FAQ scope, human handoff and monthly monitoring.

Purpose	Reduce time on repeat enquiries; free capacity for complex human support.
Bounded ask	Six months live; agree budget, staffing, security and accessibility before launch.
Readiness + risk	Approved FAQs, ≥90% accuracy gate, tested handoffs and pause controls.
Decision evidence	Accuracy, simple-query clients and complex-support staff hours; month-6 review.

Keep human support available. Agree the full six-month cost, rather than reusing a prototype budget.

Good headline results. One safety failure.

Use these results to illustrate ROI. If strong, make the case to scale the pilot when safety gates pass.

Invented results, not a real study. Attendance and staff time compare baseline and final four-week averages.

ANSWER ACCURACY

92% correct

92/100 correct vs human-approved answers.

SIMPLE-QUERY CLIENTS

60 → 40 / week

33% fewer clients; 30% target met.

COMPLEX CLIENTS

20 → 28 / week

40% more clients; not a staff-time measure.

STAFF FOCUS

40 → 52 hrs / week

30% more time on complex needs; target met.

SAFETY

One unsafe answer

Month 6 audit found an invented eligibility rule.

YOUR DECISION

Stop, continue or scale?

Does 92% accuracy override the safety failure?

Pause the chatbot. Keep human support open.

All three numeric targets are met. Unsafe eligibility advice still requires a pause.

PAUSE + REVIEW

Understand the failure

Log the failure. Review affected chats and correct misinformation with users where possible.

REDESIGN

Use approved facts

Constrain answers to approved FAQs. Hand off uncertain queries; add eligibility edge cases to testing.

FRESH APPROVAL

Retest before restarting

Recheck accuracy, safety and access. Restart only with approval; do not expand while critical failures remain.

Different organisations. A similar starting point.

Possible website chatbot FAQ scopes, not claims about attending organisations' practices.

01

Community + housing

Booking steps and service locations.

02

Legal + advocacy

Appointment admin, never legal advice.

03

Arts + culture

Event details, never restricted material.

04

Members + networks

Event and membership enquiries.

05

Access + inclusion

Approved access info; staff confirm needs.

06

Operations

Routine contact and process enquiries.

Preparation for Week 3 (optional)

No homework required. A few things to consider before our next session.

01

A use case

Where could a chatbot free staff for complex human support?

02

Your readiness

What approved information, tools and safeguards would you need?

03

People + adoption

Who would need to be involved, supported or consulted?

04

Evidence + decisions

What result would make you pause, even if the tool saved time?

Week 3: From AI Scoping to Implementation

12 November 2026 · Use cases for impact.

From scope to use	Turn a defined opportunity into a working use case.
Adoption	Build with the people who will use and oversee it.
Return on investment	Measure costs, capacity and outcomes credibly.
Beyond the pilot	Make useful changes stick, with ongoing oversight.

SHACKLETON LABS

Thank you.

Mike Davis

Founder & CEO, Shackleton Labs

Helping for-purpose organisations build practical and scalable innovation and AI capability for growth and impact.

shackletonlabs.com

mike@shackletonlabs.com

0402 208 883

Let’s connect
human to human

linkedin.com/in/mikedav84