DISCOVERY INSTRUMENT

The questions, and how each is answered

The facilitated session we run to walk **one citizen-facing automated decision** through its seven surfaces (S1–S7). Each question is tagged by how its answer is established — **confirmed with the client** in the room, or **verified in a technology audit** — so the readiness picture is evidence-backed, not self-attested.

StepInsight × Protiviti | 7 surfaces · 29 questions | Confirm + technology audit | Draft — shows the shape of the instrument

WHAT THIS SESSION DOES

We walk one already-chosen decision end to end. At each surface we ask how it actually works today, reflect the answer back, and tie it to the obligations that bite there. The output is a single picture: a heatmap of where the decision is compliant, exposed, or unknown — with the three highest-priority gaps and a straight answer on readiness for the **10 December 2026** ADM transparency deadline.

Because the obligation mesh is the rubric, every stage is scored against the law itself — not opinion. That is what makes the result defensible, and it is a live preview of the continuous-assurance "eval loop" without having to build it.

BEFORE THIS SESSION — THE DECISION IS ALREADY CHOSEN

Choosing **which** decision, confirming the entity type (non-corporate Commonwealth / corporate / state), and securing access happen in a separate partner-scoping step with Rich beforehand. Those scoping questions are not run here — they gate *whether* we run this. This instrument is the diagnostic itself.

Who needs to be in the room

ROLE	WHY WE NEED THEM	SURFACES
Decision / process owner	Holds the end-to-end picture	**S1–S7**
Data / data-matching owner	The Robodebt zone lives here	**S1, S2**
Model / scoring owner	Explainability + fairness answers	**S3, S7**
Legally-authorised decision-maker	The accountability anchor	**S4, S5**
Caseworker / operational lead	What actually happens vs. the policy	**S4, S5, S6**

SETTING UP FOR A CLEAN, HONEST SESSION

Pre-session checklist (30 seconds, the facilitator's job)

- ✓ Attendee full names confirmed (not email handles), with roles and agency-name spelling
- ✓ The chosen decision named in one line; its known source systems listed
- ✓ The model / scoring type pre-loaded — even rules-based automation can be caught by APP 1.7–1.9
- ✓ Obligation acronyms ready to expand once (APP, ART, ADM, OAIC) so they anchor in the transcript
- ✓ Audio sorted: remote attendees on their own devices; no boardroom single-mic
- ✓ Mindset: reflect back the hard words; name people as they speak; this is mapping, not auditing

The opening move (warm, not a script)

> *"Before we dive in — so I don't misquote anyone later, a quick go-round: your name and what you do here. We're going to walk one decision end to end. At each step I'll ask how it actually works today, so we can lay it against the obligations that bite by the 10th of December. We're not auditing anyone today — we're mapping the surface so the gaps are obvious and fixable. Anything that's 'we're not sure' is exactly what we want to hear."*

THE WALK — SEVEN SURFACES, THE QUESTIONS TO ASK

CONFIRM Established with the client in the room. *AUDIT* Verified in a follow-on technology audit.

> **RISK & COMPLIANCE NOTE — HOW ANSWERS BECOME EVIDENCE**
>
> Most technical-control questions below **cannot be scored "compliant" on the client's say-so** — self-attestation is precisely the failure mode the Robodebt Royal Commission identified, a control "treated as correct by default." In the room we **confirm** governance, intent and ownership; everything tagged **Audit** is flagged for verification in a follow-on technology audit (system, logs, model artefacts, sample notices). A stage scores green only once the audit evidence backs the answer. The tags double as the scope for that audit.

S1 Intake & data capture

Where the citizen data comes from

G6 · G7 · G8

QUESTIONS TO ASK THE CLIENT

1. Where does the citizen data feeding this decision come from — how many systems and third-party sources? CONFIRM AUDIT
2. Is the lawful basis and consent recorded **at field level**, or assumed? AUDIT
3. Is the data trimmed to only what the decision actually needs? AUDIT
4. What security baseline applies — PSPF, Essential Eight, certified hosting — and can the current assessment evidence be produced? AUDIT

REFLECT BACK	Each source system by name; the security-baseline acronym expanded once; any count of systems repeated.
AUDIT VERIFIES	Field-level provenance in the data schema · fields ingested vs. used · current PSPF / Essential Eight / IRAP assessment evidence.

S2 Data matching — the Robodebt zone

Where averaging and proxy bias hide

G4 · G5 · G9

⚑ HEAVY-LOAD SURFACE — PROTECT THIS IF TIME RUNS SHORT

QUESTIONS TO ASK THE CLIENT

1. Is data matched, averaged or integrated across systems to build the case picture — anything resembling **income averaging**? CONFIRM AUDIT
2. Is the matching logic documented and tested, or treated as correct by default? AUDIT
3. Are low-confidence matches flagged for a human, or do they flow straight through? AUDIT
4. Have the matching assumptions been screened for indirect discrimination through **proxy variables**? AUDIT

REFLECT BACK	The phrase "income averaging" if it surfaces — it's the Robodebt failure point. Reflect any threshold/confidence number; name the proxy variables mentioned (postcode, age band).
AUDIT VERIFIES	The actual matching logic + test evidence · the human-referral threshold in the system · any bias / proxy-variable screening artefacts.

S3 AI scoring / triage

Whether a flag can be explained

G2 · G9 · G10

QUESTIONS TO ASK THE CLIENT

1. Can you reconstruct, in plain language, **why** a given person was flagged or scored? AUDIT
2. Does each output carry a reason code a caseworker — and a tribunal — could read? AUDIT
3. Is the model a black box, or is explainability built in? AUDIT
4. Has fairness been tested across the affected population against documented thresholds? AUDIT

REFLECT BACK	The model / tool name; whether outputs carry a reason code; the fairness-testing cadence if named.
AUDIT VERIFIES	Explainability tested on real cases · reason-code content in actual outputs · model architecture · documented fairness thresholds + results.

S4 The human decision — the Seam

Decide, or rubber-stamp?

G2 · G4 · G10

⚑ HEAVY-LOAD SURFACE — THE ACCOUNTABILITY ANCHOR

QUESTIONS TO ASK THE CLIENT

1. In practice, does the human meaningfully decide, or effectively **rubber-stamp** the AI output? CONFIRM AUDIT
2. Are there defined review gates before a decision is finalised? CONFIRM
3. When a human overrides the AI, is the override and its reason logged? AUDIT
4. Under what documented legal authority and delegation is the final decision made? CONFIRM
5. Is that authorised decision-maker the **same person** who actually reviews the AI output day to day? CONFIRM

REFLECT BACK	The named decision-maker's role; the distinction between deciding and rubber-stamping; whether overrides are logged.
AUDIT VERIFIES	The **override / acceptance rate** and time-on-task in the logs — the hard evidence of whether oversight is meaningful or nominal — and that override reasons are actually captured.

S5 Notification, reasons & review

The 10 December 2026 deadline

G1 · G4 · G10

⚑ HEAVY-LOAD SURFACE — THE DEADLINE LANDS HERE

QUESTIONS TO ASK THE CLIENT

1. Today, is AI involvement disclosed to the citizen — in the privacy policy **and** at the point of decision? AUDIT
2. Are the reasons given adequate enough to support a merits review at the **ART**? CONFIRM AUDIT
3. Is the route to internal review and the ART clear to the citizen? AUDIT
4. Does the client know whether this decision is caught by **APP 1.7–1.9**, and is there a plan to be compliant by 10 December 2026? CONFIRM

REFLECT BACK	APP 1.7–1.9 expanded once ("the new ADM transparency rule"); ART expanded once ("the Administrative Review Tribunal"); the 10 Dec 2026 date repeated; disclosure live-today vs. planned.
AUDIT VERIFIES	The actual privacy collection notice + decision letters (the documents, not the description) · whether reasons in a real sample would survive ART scrutiny · review-route wording.

S6 Logging & provenance

Can you reconstruct a past decision?

G4 · G7 · G8

QUESTIONS TO ASK THE CLIENT

1. Could you reproduce **exactly** what data and model version produced a specific decision made months ago? AUDIT
2. Is the model version pinned and logged against each decision? AUDIT
3. Are the complete decision records kept in an Archives-compliant way, or scattered across operational logs? AUDIT

REFLECT BACK	The phrase "model version pinned"; where records actually live (case-management system and the model logs — two places).
AUDIT VERIFIES	An actual reconstruction attempt on one past decision · model-version pinning in the logs · whether the full record set meets Archives Act retention.

S7 Monitoring & assurance

Protiviti's lane — with an AI lens

G2 · G3 · G9 · G10 · G11

QUESTIONS TO ASK THE CLIENT

1. How is the model currently monitored — continuously, or checked when someone remembers? CONFIRM AUDIT
2. Is drift and fairness re-tested between audits, or only point-in-time? AUDIT
3. Does a rule or policy change — like the Dec-2026 date — trigger a re-assessment today? CONFIRM
4. Is the assurance sample-based and manual, or moving toward continuous and evidence-rich? CONFIRM AUDIT
5. If the model or platform is vendor-supplied, do the contracts carry AI accountability terms (e.g. DTA AI Model Clauses) and audit / access rights? CONFIRM

REFLECT BACK	The monitoring cadence; the point-in-time vs. continuous distinction — where Protiviti's value and the continuous-assurance product live (APRA, April 2026: point-in-time AI assurance is "no longer fit for purpose").
AUDIT VERIFIES	The monitoring configuration + drift/fairness re-test artefacts · sample-vs-continuous assurance evidence · the supplier contract terms (Q5 — a document review).

CLOSING THE SESSION — SOFT CHECK

1. **2–3 minute recap.** "Here's what I heard: the decision is [X]; the heaviest exposure looks like S2 / S4 / S5; the systems are [A, B, C]; the thing that surprised me was [Y]."
2. **The check.** "Any red flags from that? Anything that doesn't sound like how it actually works for you? Anything important we didn't get to?"
3. **Sit with the silence.** The first five seconds of quiet is where the real corrections come out.
4. **Name the audit scope.** "A handful of these we can only confirm by looking at the system itself — the logs, the model, the actual notices. I'll send a short list of what we'd need access to." (That list = every AUDIT question above.)
5. **What happens next.** "From this we build the readiness heatmap — every stage scored against the obligations, the three priority gaps named, and a straight answer on readiness by 10 December. That's what goes in front of Lauren and Rita."

StepInsight · CONFIRM = established with the client in the room. AUDIT = verified in a follow-on technology audit before a stage can score compliant. The 29 questions map to the brief's sections C–I, the S4 accountability probe (B-Q9), and a procurement question at S7 (G11 — DTA AI Model Clauses). Use-case selection and access (brief A, B, J) sit in the separate partner-scoping step.

Status flags current as at June 2026. Orientation only, not legal advice — confirm currency before any external use. Draft artefact showing the shape of the instrument; not a deployed engagement.