AI DATA READINESS · WORKING WORKSHEET

Can You Say Yes?

Five sequential gates for deciding whether an AI team can have a regulated dataset — and for locating precisely what is stopping you.

How to use this. Do not score it. Take **one real dataset request currently sitting in your queue** and walk the gates in order. These are sequential dependencies, not independent dimensions: you cannot protect what you have not found, or defend a protection choice you cannot measure. **The first FAIL is your answer.** Stop there. Everything below it is blocked, and investment below it is wasted until it clears.

REQUEST UNDER ASSESSMENT

DATASET / SOURCE SYSTEM	REQUESTING TEAM	DATE / ASSESSED BY

AI SYSTEM AND INTENDED PURPOSE	EU AI ACT POSTURE
	☐ High-risk (Annex III) ☐ Embedded (Annex I) ☐ GPAI / Art. 50 only ☐ Out of scope ☐ Undetermined

1 LOCATE — Can you enumerate, at element level, every sensitive data type in this source — including the copies nobody chartered?

Passes only if: you can name the specific elements — which identifiers, which Art. 9 special categories, which regulated fields, in which columns and which document types — from an existing catalogue, without commissioning a new scan. A system-level sensitivity label is not a pass.

IF FAIL — discovery is your constraint. Every protection decision below this line is a guess, and no downstream investment will hold.

☐ PASS ☐ FAIL

EVIDENCE — CATALOGUE REF. / LAST REFRESH

2 PROTECT — Can you produce a protected copy that stays useful — without a bespoke engineering project?

Passes only if all of: format and type survive (a masked ID still validates); referential integrity holds across *every* system the pipeline joins against; natural-person consistency holds in unstructured sources; distributions survive well enough for bias testing to mean something — and it is invoked as a service, not hand-built for this request.

IF FAIL — you have a capability, not a service. Capabilities scale to three models. Services scale to fifty.

☐ PASS ☐ FAIL

EVIDENCE — POLICY / METHOD, SYSTEMS JOINED

3 VALIDATE — If the AI team reports that the protected copy degraded model performance, could you tell whether that is true?

Passes only if: you can separate *anonymisation destroyed the signal* from *this model was going to be mediocre anyway* — via a comparable evaluation, a held baseline, or an agreed utility metric owned jointly with the AI team.

IF FAIL — every disagreement resolves in favour of raw data, and it resolves that way permanently.

☐ PASS ☐ FAIL

EVIDENCE — METRIC / BASELINE, JOINT OWNER

4 EVIDENCE — Could you produce the provenance record today — not in a fortnight?

Passes only if: source, classification, protection applied, approver and timestamps are a *query* against a system of record — not a reconstruction. This is the Art. 11 / Annex IV question, and the one the deferral to 2 Dec 2027 was meant to buy time to answer.

IF FAIL — a week of assembly per model does not survive contact with a portfolio. The reprieve is being spent, not used.

☐ PASS ☐ FAIL

EVIDENCE — WHERE THE RECORD LIVES

5 MEASURE — What is the observed elapsed time from request to protected dataset in hand?

Not the target in the policy — the observed median across the last ten requests. This gate is not a capability; it is the arithmetic of Gates 1–4, which is why it is the only number worth reporting upward.

Median request → protected dataset, last 10 requests		Longest of those ten	
Time for a member of staff to paste the same data into a consumer AI tool		Ratio — sanctioned vs. shadow path	

Where You Stopped — and What It Means

Find the first gate you failed. That row is your constraint, your owner, and the reason your roadmap is the shape it is.

GATE	YOUR REAL CONSTRAINT	WHAT IS BLOCKED UNTIL IT CLEARS	WHO HAS TO OWN IT
1	**Discovery**	Everything. Protection policy, lawful-basis arguments, Annex IV documentation and every downstream control are being applied to an inventory you cannot vouch for. Classification quality sets the ceiling on protection quality.	Security or data governance. Requires continuous, element-level discovery across structured and unstructured stores — including artefacts the AI pipeline itself creates.
2	**Protection at scale**	Portfolio growth. You can serve the AI programme you have, not the one being planned. Expect per-request engineering, silent join failures, and an AI team that learns to route around you.	Data / test-data-management, jointly with platform engineering. The deliverable is an API-invoked service driven by classification metadata.
3	**Utility measurement**	Your negotiating position. Without a shared utility metric you will lose every argument about protected data on assertion rather than evidence — and lose it permanently.	Jointly owned with data science. This is the cheapest gate to close and the most commonly skipped.
4	**Evidentiary trail**	Regulatory defensibility and audit cost. Art. 11 and Annex IV assume a record that exists as a by-product of the pipeline, not one assembled retrospectively per model.	DPO with platform engineering. Provenance must be written as a side effect of protection, not as a separate compliance task.
5	**Cycle time**	Nothing — and that is the point. If Gates 1–4 pass and cycle time is still measured in weeks, the constraint is the handoff between three org charts, not any single capability.	Whoever is willing to own the seam. Protection becomes a pipeline stage, not a review gate the pipeline waits on.

CARRY THIS SENTENCE INTO THE PLANNING SESSION

For _____, we can deliver a protected, still-usable copy in _____. The binding constraint is _____ (Gate _____). Until it clears, _____ planned AI use cases that depend on regulated data are blocked — and the practical alternative our people are already using is _____. Closing it needs _____ and lands in _____.

Three failure modes this worksheet is designed to catch

- **Scoring instead of stopping.** A 3-out-of-5 reads as respectable progress. It is not — if Gate 1 was the failure, the other four passes are decorative.
- **Assessing the policy, not the practice.** Gate 5 asks for observed medians precisely because target SLAs and lived cycle times diverge, and only one of them drives shadow AI.
- **Treating the 2027 deferral as relief.** Annex III high-risk obligations moved to 2 Dec 2027 and Annex I to 2 Aug 2028, but Art. 50 transparency, GPAI penalty powers and market surveillance authority activated on 2 Aug 2026.

The regulatory point worth restating

- Art. 10(5) permits special-category processing for bias work only where it **cannot be effectively achieved using synthetic or anonymised data**. Anonymisation is the regulatory default, not the concession.
- Where identifiable sensitive data *is* used, safeguards must include **state-of-the-art privacy-preserving measures, including pseudonymisation**.
- The structure is a permission conditioned on technique. Organisations that have industrialised anonymisation get to use more data, for more purposes, with shorter review.

NOTES, DECISIONS AND OWNERS