Responsible AI assessment

A decision path: first whether AI belongs in the process, then how to use it responsibly

AI is a tool, not a goal. This assessment walks the decision in order: step one asks whether AI is the right tool at all, step two lists the conditions for using it responsibly. Stopping after step one is a valid outcome, and often the best one.

Step 1: is AI the right tool

- [ ] Is the input genuinely messy: free text, images, judgement calls that do not reduce to rules?
- [ ] Have you checked that a rule, a lookup table or a well-designed form does not solve it?
- [ ] Is a wrong answer cheap, or caught by a person before it does damage?
- [ ] Does the task need to happen at a scale or speed no person can sustain?

Any box you cannot tick is a stop sign. The simpler solution is faster, cheaper, testable and explainable, and nobody has to keep watching it. Build that first; if it already solves the problem, you have your answer.

Step 2: conditions for responsible use

AI passed step one. These are the conditions before it touches real people or real data:

- [ ] You know exactly which data the model sees, and you send the minimum the task needs
- [ ] For personal or sensitive data there is a legal basis, and what the model provider stores is treated as a data-processing question with a real answer
- [ ] A person reviews the output before it reaches the people affected, wherever a mistake would hurt
- [ ] When the system is uncertain it stops, rather than producing something plausible and wrong
- [ ] Errors have an owner: who catches them, how they are corrected, who informs the person affected
- [ ] The people affected know AI is involved, in plain language
- [ ] Quality is measured: a baseline before, an evaluation set of hard cases after, and a re-test whenever the model changes
- [ ] The ongoing cost is budgeted: monitoring, evaluations, spend caps

Every unticked box is work to plan before launch, not a note for later.

Red flags: where AI should not decide

Tick any that apply to your case. A single tick means AI may prepare the decision, but must not take it:

- [ ] The decision has legal, medical or financial consequences for a person
- [ ] Nobody in your organisation could explain why the system refused someone
- [ ] A mistake stays invisible until it has already done damage
- [ ] People believe a human is doing this work, and would feel deceived to learn otherwise
- [ ] You would hesitate to explain this use of AI on your own homepage

This is how we decide at Punfyre, including deciding no. If your case fails step one, that is a result: the boring solution wins on every axis.