AI project worksheets

Ten worksheets for team sessions

BEFORE YOU START

These are meant to be printed, written on and argued over in a room. They work alone, but they work better with someone in the room who is allowed to disagree with you.

Do them in order if you are starting something new. Pick one out of the middle if you already know which bit is stuck. Nobody needs to do all ten in one sitting, and a session that gets through three honestly is worth more than one that gets through ten politely.

Two rules make them work. Write the answer down, because an answer everyone agrees with out loud is usually four different answers. And write the real answer, not the one that keeps the meeting short. A worksheet full of sensible-sounding nothing is worse than a blank one, because it looks like a decision.

They pair with the Excel toolkit, which does the sums, and the readiness check, which scores a use case in five minutes. Same stages, same language, so you can move between them.

01	**Start with the problem**	Name what you are trying to change, before anyone says the word AI.
02	**Where are you losing value?**	Find where it leaks before you decide where to spend.
03	**Is this use case ready?**	Five things have to be true before a pilot is worth running.
04	**Measures that prove it worked**	Agree what success looks like now, because afterwards everyone remembers it differently.
05	**Who could slow this down?**	Map the people whose questions will decide your timeline.
06	**Questions to ask a supplier**	Ten questions to ask before anyone signs anything.
07	**How much oversight does it need?**	Not every AI tool needs a committee. A few need considerably more than they are getting.
08	**Designing human review**	A human in the loop is only a control if the human can actually do the job.
09	**Your risk appetite**	The hardest page here, and the one that lets everything else move.
10	**Continue, pause or stop**	The decision most pilots never formally make.

Start with the problem

Name what you are trying to change, before anyone says the word AI.

HOW TO RUN IT Twenty minutes. Whoever owns the process, and at least one person who does the work every day.

> Most AI projects do not fail because the technology does not work. They fail because nobody properly answered this page.

The problem, in one sentence. No technology in the sentence.

Who has this problem, and how often?

How do you know it is a problem? What have you seen, counted or been told?

What does it cost you now — time, money, complaints, risk, people leaving?

What does good look like in twelve months? Be specific enough that someone could check.

What happens if you do nothing at all?

Is this worth solving this year? ☐ Yes ☐ No ☐ Not yet

If not yet, write what has to change first.

Where are you losing value?

Find where it leaks before you decide where to spend.

HOW TO RUN IT Thirty minutes. Bring someone who sees the volumes, and someone who handles the complaints.

> Take one process and walk it end to end. You are looking for the steps that are slow, repeated, or go wrong often enough that somebody has built a workaround.

STEP IN THE PROCESS	HOW MANY A MONTH	TIME OR COST EACH	A MONTH	WHAT IT LOOKS LIKE WHEN IT GOES WRONG

Which three of those hurt most? Say why — the biggest number is not always the worst problem.

Which of them would still hurt in two years if nothing changed?

Has anyone already tried to fix one? What happened?

NOTES, AND ANYTHING ELSE THAT CAME UP

The one step you would fix first

Name it, and say what makes it first rather than second.

Is this use case ready?

Five things have to be true before a pilot is worth running.

HOW TO RUN IT Twenty minutes per use case. Score it honestly. A low score is useful; a flattering one costs you a quarter.

> Score each one from 1 to 5, and write the evidence beside it. If you cannot write evidence, the score is a 1.

The problem is specific, and one named person owns it. 1 2 3 4 5

EVIDENCE

The data exists, you can get at it, and it is good enough. 1 2 3 4 5

EVIDENCE

The people who would use it have been asked, and want it. 1 2 3 4 5

EVIDENCE

You know which rules apply and who signs it off. 1 2 3 4 5

EVIDENCE

There is money and time for the boring bits — integration, training, support, the second year. 1 2 3 4 5

EVIDENCE

OUT OF 25

Under 12	Not yet. Fix the ones scoring 1 and 2 before you spend anything.
12 to 18	Close. Pick the lowest score and make it the first job.
Over 18	Ready to test. Go to sheet 4 and agree what would prove it.

Your lowest score. What would move it up by two, and who would do it?

NOTES, AND ANYTHING ELSE THAT CAME UP

Is this the one to start with? ☐ **Start it** ☐ **Fix the gaps first** ☐ **Park it**

If you have scored several use cases, the highest score wins, not the loudest sponsor.

Measures that prove it worked

Agree what success looks like now, because afterwards everyone remembers it differently.

HOW TO RUN IT Thirty minutes. Do not leave the room without a baseline. A target with no baseline cannot be met or missed.

WHAT YOU WILL MEASURE	WHERE THE NUMBER COMES FROM	TODAY	TARGET	BY WHEN	WHO READS IT

> Now the one most teams skip. A guardrail is the measure that tells you it is going wrong while the headline number still looks fine — handling time down, complaints up.

Your guardrail measure, and the level at which it stops being acceptable.

What would make you stop? Write it down while you still have nothing invested.

NOTES, AND ANYTHING ELSE THAT CAME UP

Agreed by

Name and date. Measures agreed by nobody in particular get renegotiated the week the results arrive.

Who could slow this down?

Map the people whose questions will decide your timeline.

HOW TO RUN IT Thirty minutes, with someone who knows how the organisation actually works rather than how the chart says it does.

> Nobody says no. Everybody asks another question. Months pass. Work out now who is going to ask what, and go and answer it before it is asked in a steering group.

NAME OR ROLE	WHAT THEY NEED TO BE SURE OF	WHAT WOULD SATISFY THEM	WHO TALKS TO THEM	BY WHEN

Who can actually say no? Who can only say not yet? Those are two different problems and they need two different conversations.

Who has to change how they work for this to pay off? Have they been asked, or just told?

NOTES, AND ANYTHING ELSE THAT CAME UP

The three conversations to have this month

Names, not departments.

Questions to ask a supplier

Ten questions to ask before anyone signs anything.

HOW TO RUN IT Take this into the demo. Write the answers in the room, not from memory afterwards.

> You are listening for two things: whether they answer the question you asked, and whether they have been asked it before.

1 Where does our data go, who else can see it, and which country is it in? ☐ OK ☐ NOT YET

2 Is our data used to train anything? Can we turn that off in writing? ☐ OK ☐ NOT YET

3 What happens to our data if we leave, and how long does that take? ☐ OK ☐ NOT YET

4 What does it do when it is not sure? Does it say so, or guess? ☐ OK ☐ NOT YET

5 How would we know it had got something wrong? ☐ OK ☐ NOT YET

6 Who is accountable when it does — you, us, or the model provider? ☐ OK ☐ NOT YET

7 What does this cost at ten times the volume? ☐ OK ☐ NOT YET

8 What happens when you change the model underneath us? ☐ OK ☐ NOT YET

9 Who else our size is using this? Can we speak to them without you there? ☐ OK ☐ NOT YET

10 What is the exit, and what does it cost? ☐ OK ☐ NOT YET

Verdict ☐ Shortlist ☐ Park ☐ No

Write the answer that decided it.

How much oversight does it need?

Not every AI tool needs a committee. A few need considerably more than they are getting.

HOW TO RUN IT Ten minutes per tool. Do it again when the tool changes what it does.

Six questions. Tick the ones that are true.

- [ ] It makes or shapes a decision about a person — a customer, an employee, an applicant.
- [ ] A mistake could cost real money, a customer, or the organisation's name.
- [ ] It touches personal data, or something commercially sensitive.
- [ ] It acts on its own, rather than a person approving each output.
- [ ] It is in front of customers.
- [ ] You would struggle to explain how it works to the person it affects.

OUT OF 6

None or one	Light. A named owner, a note of what it does, and a look at it twice a year.
Two or three	Standard. Add a risk entry, a review at each release, and someone who is not the builder checking the output.
Four or more	Heavy. Human review by design (sheet 8), a written appetite (sheet 9), a named senior owner, and a date to switch it off if it is not working.

Which box does this tool land in, and is that what it is getting today?

NOTES, AND ANYTHING ELSE THAT CAME UP

Oversight level ☐ **Light** ☐ **Standard** ☐ **Heavy**

Owner, and the date of the next look.

Designing human review

A human in the loop is only a control if the human can actually do the job.

HOW TO RUN IT Thirty minutes, with the person who will be doing the reviewing in the room. Not their manager.

What exactly is the reviewer checking? "The output" is not an answer — name the thing that could be wrong.

What are they allowed to change, and what do they have to escalate?

How many will they see an hour, and how long does one take to judge properly? Multiply those. Is that a real job?

What does the reviewer see that helps them decide — the source, the confidence, the original request?

How would you know the review was working, rather than someone clicking approve at the end of a shift?

Where does a rejected output go, and does anything learn from it?

Reviewer named

Volume a day, minutes an item, and who checks the checkers.

Your risk appetite

The hardest page here, and the one that lets everything else move.

HOW TO RUN IT Forty minutes, once, for the organisation — not per project. Then revisit it yearly.

> You do not have one appetite for AI. You might be averse to fairness risk in anything touching recruitment, and quite open to technical risk in a tool that summarises internal documents. Writing that down is what lets a team get on with the low-risk work without asking permission every time.

	AVERSE	MINIMAL	CAUTIOUS	OPEN	EAGER	WHY, IN ONE LINE — AND WHO OWNS IT
Financial	☐	☐	☐	☐	☐	
Legal and regulatory compliance	☐	☐	☐	☐	☐	
Transparency and explainability	☐	☐	☐	☐	☐	
Fairness	☐	☐	☐	☐	☐	
Accountability and governance	☐	☐	☐	☐	☐	
Contestability and redress	☐	☐	☐	☐	☐	
Technical robustness	☐	☐	☐	☐	☐	
Security	☐	☐	☐	☐	☐	
People and the environment	☐	☐	☐	☐	☐	

> Multiplying likelihood by impact is fine for sorting a long list, but two estimates multiplied together look like a fact. A rare, catastrophic risk scores the same as a likely, trivial one and slips into the middle of the register.

Anything with a severe impact, whatever it scored. Someone senior reads this list.

If this gets something wrong about a customer or a colleague, how would they know, and how would they challenge it? If there is no good answer, that is your first job.

The nine categories, the five appetite levels and the four risk treatments (avoid, limit, transfer, accept) are from the Department for Science, Innovation and Technology's AI Risk Management Toolkit, published 8 September 2026.

Continue, pause or stop

The decision most pilots never formally make.

HOW TO RUN IT Forty minutes at the end of a pilot, with the measures from sheet 4 on the table.

What did you say you would see by now? Did you see it? Use the numbers from sheet 4, not the impression.

What did it actually cost, against what you said it would? Include the time your own people spent.

What surprised you? The surprises are usually where the real answer is.

Who is using it? Who quietly stopped, and did anyone ask them why?

What would have to be true to keep going — and is it?

> Pausing is not failure. Carrying on because stopping would be embarrassing is.

Decision ☐ **Continue** ☐ **Pause** ☐ **Stop**

What you are doing about it, who owns that, and the date you look again.

What next

The Excel toolkit does the arithmetic these sheets leave to you: the use case scorecard, the cost of doing nothing, a business case stress test, a risk register on the nine categories, a pilot tracker, the supplier questions and the governance level. Every figure is worked out for you.

The readiness check is fifteen questions and about five minutes, and gives you a score with the gaps called out. It is the quick version of sheet 3.

Both are free, at deliverbetter.co.uk/free-tools.

> If the answers raised more questions than they settled, ping me a message and tell me what is getting stuck.