



AI EXPERIMENT CANVAS

Define a worthwhile AI opportunity, design a credible test, manage material risks, and make an evidence-based decision.

Instructions and a worked example here: cleverbirdtech.org/canvas

Experiment: _____
Owner: _____
Team: _____

Date: _____
Version: _____
Review date: _____

PHASE 1: FRAME

01 OPPORTUNITY

Define the problem before proposing the technology.

Problem or opportunity: _____

Who experiences it? _____

Evidence and current baseline: _____

Desired outcome: _____

Out of scope: _____

02 AI FIT

Confirm that AI is appropriate and better than a simpler alternative.

AI capability (check all that apply)

- | | | | |
|--------------------------------------|------------------------------------|------------------------------------|-----------------------------------|
| ☐ Generate | ☐ Retrieve | ☐ Summarize | ☐ Classify |
| ☐ Predict | ☐ Recommend | ☐ Automate | ☐ Act |
| ☐ Other _____ | | | |

Why AI? _____

Simpler non-AI alternative: _____

Why AI is preferable: _____

Human-led decisions that cannot be delegated: _____

03 PEOPLE AND WORKFLOW

Map who uses, governs, supports, and may be affected.

Primary users: _____ Affected parties: _____ Owner: _____ Human reviewers: _____ Tech / data support: _____ Escalation owner: _____ Required reviews (privacy, legal, security, domain): _____ _____	Current workflow (as it happens today) • _____ • _____ • _____ • _____ Human oversight points ☐ Before AI use ☐ During AI use ☐ Before output is acted upon ☐ When system fails or confidence is low	Proposed workflow (what changes with AI) • _____ • _____ • _____ • _____
--	--	--

PHASE 2: DESIGN

04 HYPOTHESIS

State what you believe will improve and what evidence will challenge it.

We believe using _____ for _____
will improve _____ from _____ to _____
because _____

Supported when: _____

Unsupported when: _____

Most uncertain assumption: _____

05 EXPERIMENT DESIGN

Define the smallest credible test.

What will be tested: _____ Participants or test cases: _____ Baseline or comparison: _____ Duration and environment: _____ Minimum credible evidence: _____	Tool / model: _____ Prompt / configuration / workflow version: _____ Resources needed: _____ Max acceptable cost: _____ Owner: _____
--	---

06 DATA AND INPUTS

Confirm the information used is suitable, permitted, and protected.

Data and knowledge sources: _____ Owner and permission: _____ Sensitivity: _____ Quality / representation concerns: _____ Prohibited inputs: _____ Known limitations: _____	INPUT CHECKLIST ☐ Authorized ☐ Relevant ☐ Accurate enough ☐ Representative enough ☐ Noticed appropriately ☐ Traceable to source ☐ Free from prohibited or unnecessary data
--	---

PHASE 3: EVALUATE

07 MEASURES AND EVIDENCE

Measure value, performance, behavior, and feasibility separately.

☐ BUSINESS / USER OUTCOME Metric: _____ Baseline: _____ Target: _____ Minimum threshold: _____ Evidence source: _____	☐ AI PERFORMANCE AND EVALUATION Quality measure: _____ Fairness / error measure: _____ Reliability measure: _____ Baseline / benchmark: _____ Minimum threshold: _____	☐ ADOPTION AND APPROPRIATE RELIANCE Usage measure: _____ Correctness / overrides: _____ User confidence / trust: _____ Overreliance signal: _____ Minimum threshold: _____	☐ EFFICIENCY AND FEASIBILITY Time impact: _____ Cost impact: _____ Effort required: _____ Operational complexity: _____ Scalability concern: _____
---	--	--	--

Evidence collection method (how results will be recorded, reviewed, and compared): _____

08 RISKS AND GUARDRAILS

Define prevention, detection, response, and stop conditions.

RISK	PREVENTION	DETECTION	RESPONSE	STOP TRIGGER

Non-negotiable guardrails (must be met at all times):

- _____
- _____
- _____

Immediate stop conditions (pause or stop if):

- _____
- _____
- _____

PHASE 4: DECIDE

09 LEARNING AND DECISION

Make an explicit decision based on the full body of evidence.

EVIDENCE SUMMARY When: _____ Performance: _____ Adoption: _____ Workflow: _____ Data: _____ Risk: _____ Feasibility: _____ Remaining uncertainty: _____	DECISION (select one) ☐ Scale - evidence supports broader use ☐ Iterate - revise and test again ☐ Narrow - proceed with a smaller use case ☐ Pause - wait for missing evidence or readiness ☐ Stop - value insufficient, risk unacceptable, or AI not the right solution
--	---

DECISION RATIONALE

NEXT ACTION
Action: _____
Owner: _____ **Due date:** _____

► **DECISION RULE**
Decisions require more than data alone. Consider values, evidence, adoption, workflows, data suitability, risk, feasibility, and human oversight.

☐ **Be responsible**
Act with care and protect people and data.

☐ **Be experimental**
Test small, learn fast, and iterate.

☐ **Be evidence-based**
Use the best available evidence, do no harm.



Cleverbird Labs | AI Experiment Canvas | v1.0
© 2025 Cleverbird Labs. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



HOW TO USE THE AI EXPERIMENT CANVAS

The canvas helps you decide whether an AI idea deserves more time, money and trust before you commit to it. It works for a solo owner testing one idea and for an enterprise team building the case to scale. Plan on 30 to 60 minutes for a first pass.

BEFORE YOU START

One person owns the canvas. Put their name in the header. Review it with the people affected before the test begins, and again when it's time to decide. If you can't answer a field, don't skip it. Write "unknown" and add it to your remaining uncertainty in section 09. A blank is a finding.

WORK THROUGH THE FOUR PHASES

Frame (01–03). Define the problem before the technology. Check whether a simpler, non-AI fix would do the job. Map who uses the output, who reviews it, and who feels the result.

Design (04–06). Write a hypothesis you could be wrong about. Choose the smallest test that would still convince a skeptic. Confirm the data you need is permitted, protected and fit for purpose.

Evaluate (07–08). Measure four things separately: user outcome, AI performance, adoption and feasibility. A tool can perform well and still go unused. Set your thresholds and stop triggers before the test starts.

Decide (09). Compare results to the thresholds you set. Choose one decision and write down why.

THE FIVE DECISIONS

Scale: the evidence supports broader use.

Iterate: promising, so change one thing and test again.

Narrow: it works for part of the use case, so proceed only there.

Pause: evidence is missing or the organization isn't ready.

Stop: not worth it, too risky, or AI isn't the answer.

Stop is a good outcome. You learned something before it got expensive.

THREE RULES WORTH KEEPING

Set thresholds before you see results. Keep a person accountable for every decision the AI informs. Write down what surprised you, because that's usually the most useful thing you'll learn.

A WORKED EXAMPLE (FICTIONAL)

Harbor & Pine, an 8-person online home goods shop, tested AI-drafted replies to routine customer emails for three weeks. A person approved every reply before it was sent. Reply time fell from 26 hours to 9, which beat the minimum but missed the 8-hour target. Shipping and order drafts were strong. Both errors came from returns questions. **Decision: Narrow.** Keep AI drafts for shipping and order emails, fix the returns FAQ, then retest.

See the full worked example at cleverbirdtech.org/canvas

NEED MORE ROOM?

The Expanded Workbook uses the same framework with space for longer answers. Download it at cleverbirdtech.org/canvas

Want help running your first experiment? Visit cleverbirdtech.com