
AI Consulting Playbook

From an AI opportunity to a governed decision and clean handoff

Michael Long, surfaces.systems

October 9, 2026 | Version 1.0.1

ORIENTATION

How to use this playbook

AI consulting starts with a request that is usually too broad: find the use cases, automate the work, choose a platform, or get the company ready. This playbook supplies the operating layer between that request and a defensible decision. It treats consulting as evidence work carried through delivery, adoption, and ownership transfer.

DECISION: The promise of this playbook

By the end, you will be able to turn a broad AI request into a bounded engagement, diagnose the real workflow, choose a useful pilot, coordinate delivery and risk, measure what changed, and leave the client able to continue without the consultant.

For: independent consultants, product and design leaders, transformation teams, technical advisors, and internal AI leads. The playbook assumes basic familiarity with AI products but does not require model training or production engineering expertise.

CAUTION: Useful guidance is not legal advice

The governance and regulation sections help a team identify questions, evidence, owners, and review gates. They do not determine legal obligations. Jurisdiction, sector, contracts, employment rules, privacy duties, and the exact system role still require qualified review.

Choose a path

PATH	TIME	WHAT TO DO	OUTCOME
Orientation	90 minutes	Read the concepts and worked examples; skip blank worksheets.	You can frame a better AI engagement and challenge a weak pilot.
Practice	6-8 hours	Complete one worksheet in every module.	You can draft a credible consulting packet.
Full field guide	12-15 hours plus capstone	Complete the exercises, worksheets, capstone, and decision memo.	You can make and defend a bounded recommendation and handoff.

The repeated learning loop

- 1 Rejoin the Harborline Service Operations case.
- 2 Learn one consulting move in plain language.
- 3 See the decision rule, model, or artifact that makes the move inspectable.
- 4 Apply it in a short exercise and compare your answer.
- 5 Complete a worksheet that carries into the next module.
- 6 End with the claim the artifact supports and the boundary it does not cross.

What you need

- This PDF, a pen or note-taking app, and access to the people who perform or receive the work.
- For a real engagement: the sponsor request, current workflow, service or product measures, policies, data inventory, vendor commitments, and responsible owners.
- Permission to observe the work and test assumptions. Interviews alone rarely expose queues, workarounds, exception handling, or unrecorded review burden.
- A place to keep a decision log, evidence ledger, risk register, and versioned deliverables.

FIELD NOTE: Start with the work, not the model

A technology-first discovery tends to collect possible features. A work-first discovery finds the decisions, handoffs, exceptions, and evidence that determine whether AI would help at all.

CAUTION: Learning-sized is not production-sized

The exercises use small fictional evidence packets so you can practice the method. A workshop score, demonstration, or short pilot cannot establish broad safety, durable value, legal compliance, or readiness to scale.

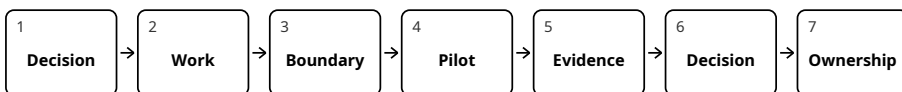
Sources were reviewed on October 1, 2026. Harborline, its data, the consulting sequence, rating labels, exercises, and templates are fictional examples or author synthesis. Regulatory dates are a snapshot, not a substitute for current legal review.

Contents

How to use this playbook	ii	10. Run the pilot as a controlled evidence loop	72
Choose a path	iii	11. Turn the evidence into the next decision	81
What you need	iv	12. Leave the client able to continue	89
Contents	v	Guided capstone: Harborline decision review	98
The consulting loop at a glance	vi	Guided capstone answer key	110
The consulting packet you will build	xi	Guided capstone self-assessment	116
1. Frame the engagement around a decision	1	Appendix A: consulting lifecycle	118
2. See the work as it actually happens	8	Appendix B: discovery interview and observation guide	119
3. Choose one opportunity worth testing	14	Appendix C: assist, recommend, decide, and act	121
4. Define the future workflow and AI boundary	22	Appendix D: measurement and economics cheat sheet	122
5. Make ownership, risk, and permissions explicit	29	Appendix E: governance and vendor question bank	123
6. Design a pilot that can answer a decision	38	Appendix F: plain-language glossary	124
7. Plan delivery around uncertainty	46	Appendix G: worksheet index	126
8. Build and verify the whole pilot system	55	Further reading and source notes	127
9. Prepare the people who will use and support it	63	Changelog	131

The consulting loop at a glance

A consulting engagement starts with a decision, not an AI feature. The team observes the work, defines the boundary, runs a controlled pilot, interprets the evidence, and chooses the next action. Ownership then shifts into the client operating system. New production evidence creates the next decision.



Operating evidence creates the next decision and restarts the loop

The recurring case: Harborline

Harborline is a fictional regional commercial-equipment maintenance company. Service coordinators receive requests through email, phone, and a customer portal. They identify the account and equipment, assess urgency, search manuals and contracts, and create a work order. Safety, technician certification, geography, parts, and service-level commitments complicate dispatch.

The COO asks the consulting team to 'automate dispatch with AI.' Field evidence points to a narrower first move: structure intake, retrieve approved evidence, and draft a work order for coordinator review.

Priority changes, technician assignment, customer commitments, and autonomous writes remain outside the first pilot.

Worked example: a bounded support pilot

The following customer support example shows a completed decision summary. The scenario and all numbers are fictional. The targets describe what the pilot will test, not results already achieved.

For **four experienced agents answering software setup questions**, the evidence supports testing **AI-generated reply drafts**. A review of 100 recent tickets found agents repeatedly searching the same help articles. Finding instructions, writing, checking, and sending a reply took a median of eight minutes; 96% of sent replies were factually correct. The target is six minutes while maintaining at least that accuracy.

The pilot runs for **two weeks and covers routine setup questions only**. It permits approved help articles and ticket text with customer identifiers removed. AI may draft replies and cite supporting articles. An agent must check and send every response. The system cannot send messages or change customer accounts.

The controls include **restricted access, logs, and an off switch**. The support lead owns the pilot and checks completed replies for factual errors. The engineering lead owns access permissions, identifier removal, and shutdown.

We **stop immediately** if customer identifiers reach the AI or an unreviewed reply is sent. At the end of two weeks, we **extend once, by one week**, if fewer than 150 eligible tickets are completed and neither incident has occurred. If volume remains too low, we close the pilot as inconclusive. With at least 150 tickets, we **revise** if median time exceeds six minutes or accuracy falls below 96%. We **expand to eight agents** only if both targets are met and no stop event has occurred. Timing includes checking and sending; accuracy measures completed replies.

The current evidence does not establish **that AI drafting will save two minutes, that replies can be sent without review, or that results will transfer to billing disputes or inexperienced agents.**

Build your decision summary

Answer the questions one at a time, using the worked example as a guide. Record what you know and mark missing evidence as unknown. Return to those gaps as you complete the modules; do not invent facts to fill the summary.

- 1

Which workflow and users are in scope? Name the task and the people who will take part.
- 2

What problem have you observed? Describe the difficulty in the current work.
- 3

What evidence supports that problem? Record observations, the sample, and baseline measures.
- 4

What bounded change will you test? Choose one change small enough to evaluate and reverse.
- 5

What outcome do you expect? Name the target and how you will measure it, including review effort.
- 6

What data may the system use? Identify permitted sources, required removal of identifiers, and excluded data.
- 7

What actions may the system take? State what AI may draft, recommend, send, or change, and what requires human approval.
- 8

Who may be exposed to the output? Define participants, recipients, duration, and activities outside the pilot.
- 9

What controls will enforce those limits? Specify access restrictions, review, logs, and how the pilot can be stopped.
- 10

Who owns the pilot and its controls? Assign each responsibility and the final decision to a named person or role.
- 11

When will you stop, revise, extend, or scale? Set measures, thresholds, minimum evidence, review timing, and any extension limit before starting.

12

What does the current evidence not establish? Name untested capabilities, populations, and outcomes.

Use your answers to complete the summary below. Keep unknowns visible and revise the summary as new evidence arrives.

DECISION: The sentence you will keep completing

For workflow and users ____, the evidence supports testing bounded change ____ because observed problem and expected outcome _____. The pilot permits data, actions, and exposure ____, with controls and owners _____. We will decide stop, revise, extend, or scale using measures and gates _____. The current evidence does not establish _____.

The consulting packet you will build

MODULES	PACKET
1-3	Engagement decision brief, current-state evidence, opportunity recommendation.
4-5	Target workflow, AI boundary, risk controls, permissions, governance map.
6-8	Pilot charter, integrated delivery plan, system manifest, readiness evidence.
9-10	Change-impact plan, training and support, weekly evidence board, incident loop.
11-12	Outcome scorecard, decision memo, operating runbook, ownership and handoff record.

Five rules for the full guide

- [1] Do not turn a sponsor request into a solution before observing the work.
- [1] Do not collapse business impact, system quality, risk, and adoption into one score.
- [1] Do not let a prompt serve as the control for permissions, money, safety, privacy, or irreversible actions.
- [1] Do not call a pilot successful unless it changes a named decision against predeclared evidence.
- [1] Do not call delivery complete until named owners accept the system, evidence, runbooks, and review cadence.

CLAIM: The operating principle

A useful engagement leaves behind clearer decisions, credible evidence, and an organization able to continue without the consultant.

Frame the engagement around a decision

Turn a broad AI request into a bounded question with an owner, evidence standard, and decision date.



Learning objectives. You will separate the sponsor's stated request from the decision the organization needs to make; name scope, non-goals, owners, assumptions, and evidence; and set an engagement boundary that can survive new ideas without quietly expanding.

The request is usually not the decision

'Find our AI use cases' names an activity. 'Automate dispatch' names a preferred answer. Neither tells the team what decision must be made, by whom, or from what evidence. A useful engagement brief translates the request into a question that can produce an action.

SPONSOR SAYS	DECISION THE ENGAGEMENT COULD SUPPORT	EVIDENCE NEEDED
Find our AI use cases	Which one or two workflow changes deserve a bounded pilot this quarter?	Observed work, baseline, alternatives, dependencies, value, risk.
Choose an AI platform	Which capability and operating model fit the named workflow and constraints?	Requirements, security and data posture, tests, cost, exit terms.
Automate dispatch	What is the smallest safe change that reduces intake delay without weakening safety or accountability?	Workflow evidence, exceptions, control points, pilot measures.
Create an AI strategy	Which decisions, capabilities, and governance routines should the organization fund next?	Portfolio evidence, maturity gaps, operating constraints, owners.

Write the decision before the workplan

- 1 Name the decision and the person accountable for making it.
- 2 Name the workflow, users, business boundary, and jurisdictions in scope.
- 3 State the choices that could follow: stop, repair the process, automate conventionally, pilot AI, buy, build, or defer.
- 4 Define the evidence the decision owner considers sufficient and the date the decision is due.
- 5 List non-goals and prohibited assumptions so discovery cannot silently convert them into commitments.
- 6 Record who may change scope, how the change will be evaluated, and what it does to time, cost, and evidence quality.

CLIENT CONVERSATION: A plain way to reframe the ask

'I hear the desired outcome. Before we turn it into a solution, I want to name the decision you need to make and the evidence you would trust. We may still arrive at that answer, but the engagement should be able to show when another path is stronger.'

A decision brief has six parts

PART	QUESTION IT ANSWERS	COMMON FAILURE
Decision	What action will this work support?	A deliverable replaces a decision.
Owner	Who can accept the recommendation and its risk?	The sponsor is influential but not accountable.
Boundary	Which workflow, population, locations, systems, and actions are included?	A promising demo becomes an enterprise claim.
Evidence	What observations, measures, tests, and approvals are needed?	Opinion and workshop enthusiasm are treated as proof.
Timing	When must the decision be made, and what can be learned by then?	The schedule assumes evidence that cannot be collected.
Non-goals	What will the engagement deliberately not establish?	Unasked strategy, architecture, or compliance work appears later.

Protect engagement integrity

- Disclose vendor relationships, referral fees, platform incentives, and reusable intellectual property that could shape a recommendation.
- Separate facts supplied by the client, observations made by the team, external sources, and consultant synthesis.
- Do not use client data in unapproved tools or promise that a provider contract, setting, or model behavior satisfies a legal duty.
- Record material disagreements and uncertainty. A clean narrative is not worth erasing a real decision risk.
- Make acceptance criteria visible before the final presentation, not while the recommendation is being negotiated.

Exercise: rewrite the Harborline request

The COO asks, 'Can you automate dispatch with AI before peak season?' Write a decision, owner, scope, and non-claim that would let discovery find a narrower answer.

ANSWER: One defensible frame

Decision: whether Harborline should run a bounded pilot that reduces service-intake time and rework before peak season. Owner: the COO with Service Operations accountable for workflow acceptance and IT/Risk accountable for system approval. Scope: two queues, coordinator-facing assistance, no autonomous assignments or customer commitments. Non-claim: the engagement will not establish that dispatch can be automated safely across all regions or service types.

Worksheet 1: engagement decision brief

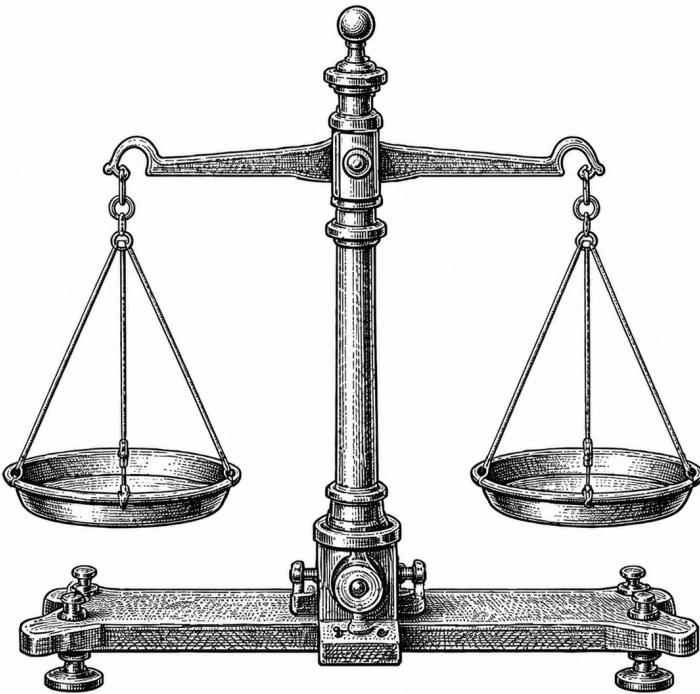
ENGAGEMENT DECISION BRIEF
Sponsor request in the sponsor's words
Decision the engagement must support
Decision owner, contributors, reviewers, and affected groups
Workflow, users, population, systems, locations, and jurisdictions in scope
Choices the evidence may support

ENGAGEMENT DECISION BRIEF
Evidence required and decision date
Non-goals, prohibited assumptions, and legal-review boundary
Scope-change authority and integrity disclosures
CLAIM: What this work supports A decision brief supports a bounded discovery and makes later scope changes visible. It does not prove the opportunity is valuable, feasible, safe, lawful, or ready for a pilot.

Sources: C1, C3-C5, C16, C18. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

See the work as it actually happens

Observe the actors, evidence, queues, handoffs, and exceptions that determine the real opportunity.



Learning objectives. You will map current work from evidence rather than policy alone, establish a usable baseline, distinguish normal flow from exception work, and separate an observed finding from an interpretation or recommendation.

The documented process is one source

A procedure explains how work is supposed to move. Interviews explain how people understand it. Logs and records show some of what the systems captured. Observation exposes switching, copying, waiting, review, workaround, and recovery that neither the procedure nor the logs may contain. Use the sources together.

EVIDENCE	BEST FOR	MAIN LIMITATION
Policy and procedure	Required sequence, roles, controls, official exceptions.	May lag practice or omit local workarounds.
Interview	Intent, judgment, pain, trust, incentives, unrecorded history.	Recall and social desirability can distort frequency.
Observation	Actual sequence, tool switching, hidden review, exception handling.	A small sample can overrepresent unusual days.
System logs	Volume, timestamps, transitions, errors, repeated behavior.	A log records what was instrumented, not the whole experience.
Work artifacts	Inputs, outputs, quality defects, provenance, rework evidence.	Retention and privacy constraints may limit access.
Outcome data	Service, financial, quality, safety, or user consequence.	Confounding and missing baselines weaken attribution.

Map the unit of work

Choose a unit that can be followed from entry to outcome: one service request, claim, support conversation, inspection, planning cycle, or invoice exception. For each unit, record the actor, input, decision, evidence, tool, handoff, wait, exception, output, and consequence.

FIELD	HARBORLINE OBSERVATION
Entry	Email, phone note, or portal request arrives with inconsistent detail.
Interpretation	Coordinator identifies account, equipment, symptom, urgency, and contract.
Evidence search	Manuals, prior work orders, customer contract, parts and certification data.
Decision	Create request, ask for missing information, escalate safety concern, or prepare dispatch.
Handoff	Coordinator to service supervisor, scheduler, technician, or customer.
Exception	Unknown serial number, conflicting urgency, stale manual, missing contract, safety phrase.
Outcome	Complete work order, rework, delay, escalation, wrong commitment, or prevented error.

Build a baseline that matches the decision

- Volume: eligible units, arrival pattern, seasonality, and important slices.
- Time: active work, elapsed time, waiting, review, and recovery time.
- Quality: complete-first-time rate, rework, escalation, defects, and downstream correction.
- Service and user outcome: time to resolution, missed commitment, satisfaction, safety, or burden.
- Economics: labor, platform, vendor, support, review, incident, and opportunity cost.
- Control and risk: access exceptions, policy overrides, privacy events, near misses, and audit gaps.

CAUTION: Do not baseline only the easy measure

Handle time may fall while review time, error correction, employee burden, or downstream delay rises. Measure the whole workflow outcome and keep the unit of analysis visible.

Keep facts, interpretations, and proposals separate

TYPE	EXAMPLE	HOW TO RECORD IT
Observed fact	14 of 30 observed requests required a second system lookup after the work order was opened.	Source, sample, period, unit, and exact count.
Participant report	Coordinators say contract lookup is the least predictable step.	Speaker role, context, and whether the theme repeated.
Interpretation	Fragmented evidence may be a stronger constraint than dispatch logic.	Reasoning and alternative explanations.
Proposal	Test retrieval and drafting before automated assignment.	Expected outcome, risk, evidence plan, and owner.

Exercise: find the hidden work

A process map shows a two-minute 'create work order' step. Observation shows coordinators spend another seven minutes locating the contract, resolving serial-number differences, and rewriting customer language. What belongs in the baseline?

ANSWER: Follow the outcome, not the system event

Baseline active and elapsed time from request receipt to a reviewable work order. Record search, interpretation, corrections, waiting, and exception paths separately. The two-minute transaction is one component, not the unit of work.

Worksheet 2: current-state evidence

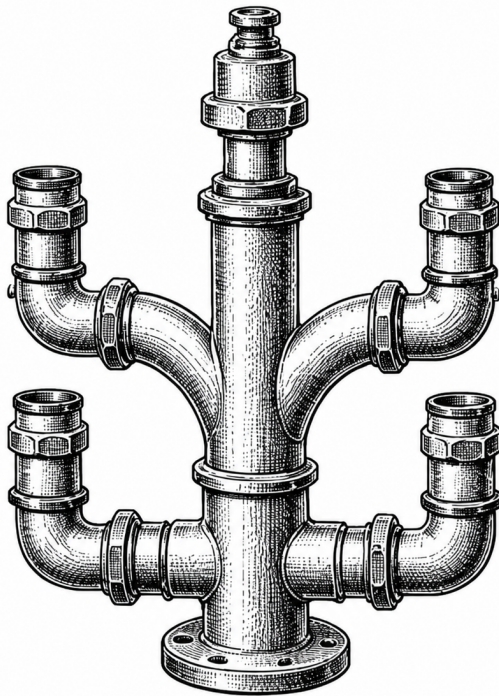
CURRENT-STATE EVIDENCE SHEET
Unit of work, start state, end state, and population
Actors, decisions, evidence, tools, and handoffs
Normal flow, exceptions, workarounds, and recovery
Baseline volume, time, quality, service, economics, and risk

CURRENT-STATE EVIDENCE SHEET
Evidence sources, sample, period, limitations, and missing data
Observed facts
Interpretations and competing explanations
Questions to resolve before recommending a change
CLAIM: What this work supports A current-state evidence sheet supports a defensible description of a named workflow and its baseline. It does not establish causality, full population prevalence, or that AI is the right intervention.

Sources: C16-C19. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Choose one opportunity worth testing

Compare AI with process repair, conventional automation, and no change before selecting a pilot.



Learning objectives. You will generate alternatives from workflow evidence, classify the proposed role of AI, compare value, feasibility, risk, and readiness, and recommend one bounded opportunity without pretending a workshop score is proof.

Start with alternatives, not use cases

A strong opportunity review asks what change would improve the workflow. AI is one possible mechanism. Process repair, better data, clearer policy, search, rules, integration, staffing, or no change may be stronger. Include those options before a favored solution gathers momentum.

OPTION	HARBORLINE EXAMPLE	WHEN IT MAY BE STRONGER
Process repair	Require serial number and safety indicator at intake.	The main failure is missing or inconsistent input.
Conventional automation	Validate account and contract fields; route by deterministic rules.	Rules are stable, observable, and exact.
Search and information design	Unify approved manuals and contract lookup.	People need attributable evidence more than generation.
AI assistance	Structure intake and draft a work order from approved evidence.	Language varies but a human can review the result.
AI recommendation	Suggest priority or technician candidates.	Judgment can be evaluated and a responsible human decides.
AI action	Assign technician and commit to customer timing.	Only after authorization, safety, recovery, and accountability are strong.
No change or defer	Keep current process while fixing data ownership.	Dependencies prevent useful learning or the risk is disproportionate.

Classify the role before scoring the idea

ROLE	SYSTEM BEHAVIOR	HUMAN RESPONSIBILITY	TYPICAL EXPOSURE
Assist	Draft, summarize, retrieve, transform, or flag.	Review and complete the work.	Low to moderate when actions stay reversible.
Recommend	Rank or propose a decision.	Understand evidence, decide, and record override.	Depends on consequence and review quality.
Decide	Select an outcome within policy.	Oversee exceptions and challenge outcomes.	High when rights, safety, money, or access are affected.
Act	Change an external system or communicate a commitment.	Authorize, monitor, reconcile, and recover.	Highest when actions are consequential or hard to reverse.

CAUTION: A human in the loop is not a control by itself

Name what the person can see, how much time they have, what authority they retain, how overrides work, and whether incentives or automation bias make the review meaningful.

Screen each option against evidence

DIMENSION	QUESTIONS	EVIDENCE
Outcome value	Which user or business outcome changes? How much does the problem matter?	Baseline, outcome data, affected volume, decision owner.
Workflow fit	Does the proposed behavior fit the real task, exception rate, and review path?	Observation, task analysis, error and recovery paths.
Technical feasibility	Can the complete system meet quality, latency, integration, and reliability needs?	Prototype, representative tests, architecture constraints.
Data and knowledge	Are permitted, current, attributable inputs available?	Inventory, access rules, provenance, quality samples.
Risk and rights	Who can be harmed or excluded? Which duties and controls apply?	Impact assessment, legal/risk review, affected-person input.
Adoption readiness	Will roles, workload, skills, trust, and support permit use?	Role map, capacity, incentives, prior change history.
Economics	What is the full cost to build, run, review, govern, and exit?	Baseline cost, vendor terms, staffing and sensitivity model.

Use High when observed evidence supports the rating, Medium when a material dependency remains, and Low when evidence shows a poor fit or a prerequisite is missing. Use Unknown when evidence is absent. Record the facts behind value and readiness separately; do not add the labels into a single score. These labels organize a discussion, not an investment calculation.

Show the reasoning behind one rating

Harborline's structured intake has High value because missing serial, contract, and safety fields repeatedly cause rework. Readiness is High only for validating required fields in the two observed queues, where Operations owns the field policy. Mandatory portal fields are the conventional alternative. Approved evidence retrieval has High value but Medium readiness because source owners, dates, and contract labels need repair. The recommendation is to test intake and retrieval together only after those dependencies close. These are fictional case judgments, not market benchmarks.

Rank Harborline's first opportunities

OPPORTUNITY	VALUE	READY	RECOMMENDATION
Structured intake	High	High	Include
Approved evidence retrieval	High	Medium	Include after access and freshness work
Work-order drafting	High	Medium	Include with coordinator approval
Priority recommendation	Medium	Low	Challenge separately
Autonomous assignment	Unknown	Low	Exclude from first pilot

Exercise: challenge the workshop winner

A workshop ranks autonomous technician assignment first because leaders expect the largest savings. Data ownership, safety exceptions, scheduling permissions, and recovery are unknown. What should the recommendation say?

ANSWER: Separate attraction from readiness

Record the idea as a longer-horizon hypothesis, not the first pilot. Recommend the narrower intake, retrieval, and drafting flow because it can test workflow value while keeping assignments and commitments under human control. List the evidence autonomous assignment would require before reconsideration.

Worksheet 3: opportunity recommendation

OPPORTUNITY RECOMMENDATION
Observed problem, affected people, and baseline evidence
Alternatives: process, rules, information, AI, defer, or no change
Proposed AI role: assist, recommend, decide, or act
Outcome value and workflow fit evidence

OPPORTUNITY RECOMMENDATION
Technical, data, risk, adoption, and economic readiness
Unknowns, dependencies, and disconfirming evidence
Recommended opportunity and excluded opportunities
Next decision and evidence needed

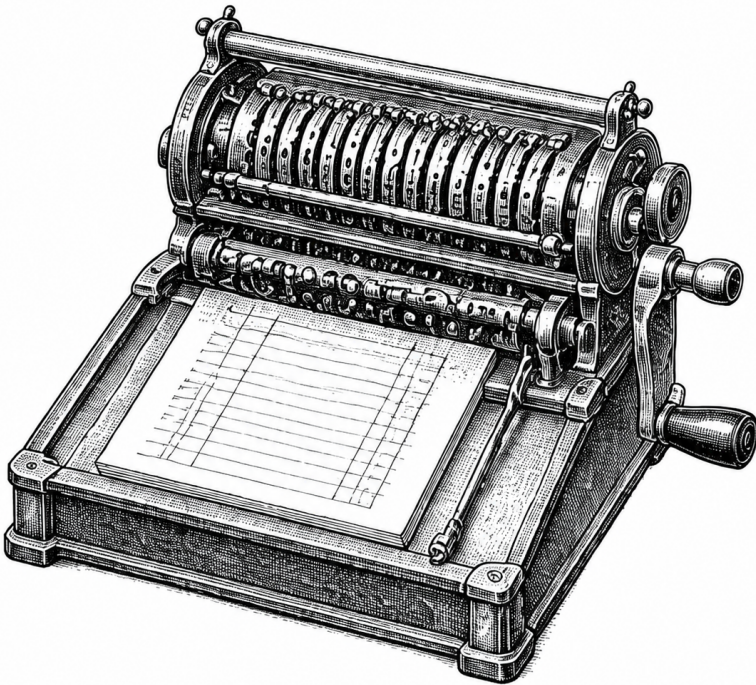
CLAIM: What this work supports

An opportunity recommendation supports the choice of one bounded problem to investigate or pilot. It does not establish production feasibility, return on investment, compliance, or organizational readiness to scale.

Sources: C1-C5, C16-C19, C25. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Define the future workflow and AI boundary

Specify what changes, what stays human, what data and actions are permitted, and how the work fails safely.



Learning objectives. You will design the future workflow around a complete service, write a behavior and outcome contract, set the AI boundary, define meaningful human review, and connect the proposed change to a testable value hypothesis.

Design the service, not the model call

The future workflow includes intake, identity, data access, context, model behavior, interface, review, approval, tool execution, logging, fallback, support, and recovery. A useful boundary names every point where responsibility or state changes.

STAGE	HARBORLINE FUTURE FLOW	CONTROL
Admit	Coordinator opens an eligible request from one of two pilot queues.	Queue, user, region, and request type are checked.
Structure	Assistant extracts account, equipment, symptom, urgency cues, and missing fields.	Source text remains visible; missing data is not invented.
Retrieve	System fetches approved manual and contract passages.	Identity, access, version, date, and citation travel with evidence.
Draft	Assistant proposes a work order and questions for the customer.	No priority, assignment, commitment, or external write.
Review	Coordinator compares draft with source request and cited evidence.	Editable fields, reasons, uncertainty, and unsupported claims are visible.
Commit	Coordinator corrects and submits through the existing system.	Existing authorization, audit, and confirmation remain authoritative.
Recover	System degrades to current workflow when evidence or service is unavailable.	No partial hidden state; issue is visible and traceable.

Write a behavior and outcome contract

DECISION: Harborline pilot contract

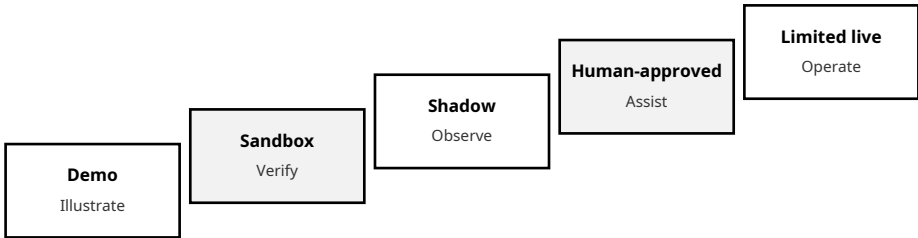
Given an eligible service request and permitted current sources, the assistant should produce a traceable intake summary, list missing information, retrieve approved evidence, and draft a work order for coordinator review. It must never invent source facts, change priority, select a technician, contact the customer, or write to the service system. When evidence is missing, conflicting, stale, or inaccessible, it must expose the problem and return the work to the coordinator.

The contract names behavior, outcome, exclusions, and fallback in language that product, engineering, operations, evaluation, risk, and users can inspect together. It should be specific enough to generate tests and workflow measures.

Make the boundary explicit

BOUNDARY	PERMITTED	EXCLUDED IN FIRST PILOT
Users	Named coordinators and supervisors in two queues.	Technicians, customers, contractors, and other regions.
Data	Approved request, account, contract, manual, and prior-work-order fields.	Unapproved email, personal folders, unrelated customers, hidden credentials.
Outputs	Structured intake, missing questions, citations, draft work order.	Final safety decision, customer promise, legal interpretation.
Actions	Save a draft in pilot workspace after coordinator confirmation.	Production write, priority change, assignment, message, purchase, refund.
Exposure	Sandbox, then shadow, then human-approved draft if gates pass.	Unobserved live automation or autonomous expansion.
Time	Six-week pilot with frozen weekly versions and change log.	Indefinite beta or silent model/provider changes.

Increase exposure only when the prior boundary is supported



Define meaningful human control

- Authority: the person can reject, edit, defer, escalate, and use the prior workflow.
- Information: source request, cited evidence, system uncertainty, and changed fields are visible.
- Time and workload: review is possible within real queue conditions and does not become a rubber stamp.
- Competence: the person understands the task, common AI failure modes, and when to escalate.
- Record: acceptance, edit, override, reason, and resulting state can be audited without punishing appropriate disagreement.

Connect behavior to value

LINK	HARBORLINE HYPOTHESIS	EVIDENCE
Capability	Structure varied requests and retrieve approved evidence.	System evaluation and trace review.
Task change	Coordinator spends less time interpreting and searching.	Task time and observation.
Workflow change	More work orders are reviewable on the first pass.	Complete-first-time and rework rate.
Service outcome	Eligible requests reach scheduling sooner without more safety or contract errors.	End-to-end time, quality, safety, and slices.
Business result	Capacity or service reliability improves at acceptable full cost.	Volume, staffing, service, cost, and sensitivity analysis.

CAUTION: A plausible chain is still a hypothesis

A model can produce cleaner drafts without improving cycle time, service, cost, or user burden. The pilot must measure each important link rather than infer the business result from output quality.

Exercise: draw the hard line

A coordinator asks the assistant to assign a technician because the recommended person is obvious. The pilot charter covers drafting only. What should happen?

ANSWER: Keep the charter authoritative

The assistant may summarize permitted evidence that helps the coordinator use the existing assignment process, but it must not choose or write the assignment. Record the request as evidence for a future opportunity review. Do not expand permissions through conversational convenience.

Worksheet 4: target workflow and boundary

TARGET WORKFLOW AND AI BOUNDARY
Target users, unit of work, start state, and end state
Future workflow: admit, structure, retrieve, generate, review, commit, recover
Behavior and outcome contract
Permitted users, data, outputs, actions, exposure, and time

TARGET WORKFLOW AND AI BOUNDARY
Excluded decisions, actions, populations, and claims
Human authority, information, time, competence, and audit record
Safe fallback and degraded workflow
Capability-to-business value hypothesis and measures

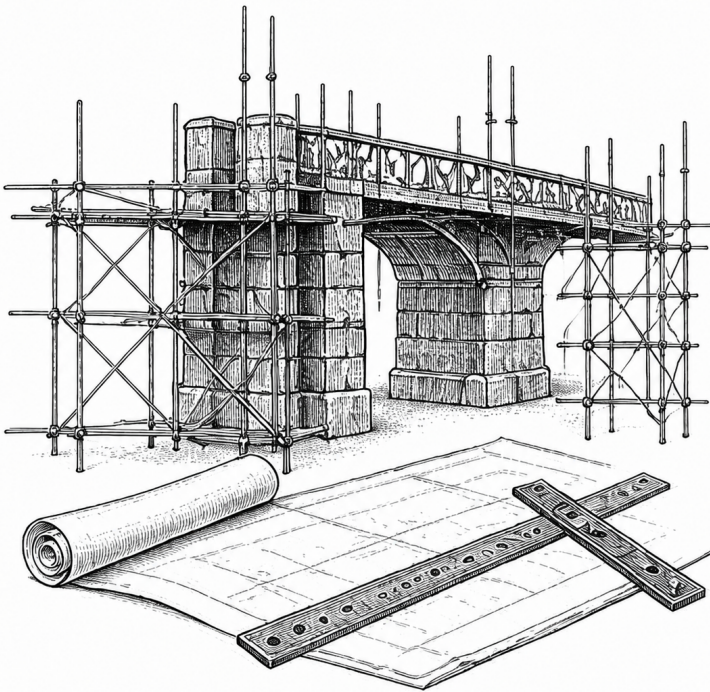
CLAIM: What this work supports

A target workflow and boundary support design, estimation, control selection, and pilot planning for a named use. They do not establish that the system can meet the contract or that the expected business outcome will occur.

Sources: C1-C5, C13-C18. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Make ownership, risk, and permissions explicit

*Translate affected people, obligations, and failure paths into
controls, gates, and accountable decisions.*



Learning objectives. You will map risk to observable controls, assign decision rights and escalation, define data and action permissions, perform vendor diligence, and use standards and law as overlays on the real system rather than as substitute checklists.

Governance is part of delivery

Governance decides who may propose, approve, operate, change, monitor, stop, and retire an AI system. A steering committee alone does not provide those controls. Name who can change the workflow, grant data access, accept vendor terms, approve a release, stop an incident, and represent affected people.

NIST AI RMF FUNCTION	CONSULTING MOVE	EVIDENCE ARTIFACT
Govern	Set policies, roles, risk tolerance, review rights, and accountability across the lifecycle.	Decision-rights map, policy, inventory, review cadence.
Map	Describe context, affected people, intended use, impacts, dependencies, and risk.	Workflow, system boundary, impact and risk register.
Measure	Select and run methods for quality, safety, rights, security, reliability, and impact.	Evaluation plan, test evidence, user and workflow measures.
Manage	Prioritize, treat, accept, transfer, monitor, and respond to risk.	Controls, gates, owners, rollout, incident and retirement plan.

NIST AI RMF 1.0 is a voluntary risk-management framework. Use its Govern, Map, Measure, and Manage functions to organize evidence, not to certify a deployment. ISO/IEC 42001 provides an organization-level management-system bridge for policy, objectives, risk, performance review, audit, and continual improvement. Neither framework determines the law that applies to a specific engagement.

Write risk as a testable chain

ELEMENT	HARBORLINE EXAMPLE
Asset or affected interest	Worker and customer safety, correct contract service, personal and commercial data.
Hazard or failure	A stale manual passage leads to an incorrect urgency cue in a draft.
Exposure	Coordinator sees the draft during a live queue and may accept it under time pressure.
Consequence	Wrong service path, delayed safety escalation, bad commitment, or downstream rework.
Preventive controls	Approved-source inventory, freshness rules, citation, safety phrase routing, excluded priority field.
Detective controls	Exact source/version trace, sampled expert review, safety slice, override and correction logging.
Recovery	Stop pilot, revert to current workflow, notify owner, preserve evidence, correct affected records.
Owner and gate	Service Safety owns consequence; Product owns fix; Risk approves restart after fresh evidence.

CAUTION: Do not use likelihood to erase severity

A rare High-severity failure may require a hard architectural control or a narrower boundary even when the aggregate score is strong. Record frequency, severity, detectability, reversibility, affected groups, and uncertainty separately.

Map permissions from source to effect

LAYER	QUESTIONS	CONTROL
Identity	Which person, role, tenant, and device is acting?	Authenticated identity and role mapping.
Source data	Which fields and records may this task read?	Least-privilege access, field limits, purpose and retention.
Model context	What is sent to which provider and under which terms?	Data classification, minimization, approved route, logging limits.
Output	What may be shown, stored, copied, or used as evidence?	Labeling, provenance, sensitive-output handling, retention.
Action	Which resource and operation may be proposed or executed?	Server-side authorization, explicit approval, idempotency, audit.
Change	Who may alter prompt, model, retrieval, tool, policy, or population?	Versioned change control and revalidation.

Use law as a jurisdictional overlay

As of October 1, 2026, the European Commission reports that the AI Act generally applies from August 2, 2026, with earlier application of prohibited-practice and AI-literacy obligations on February 2, 2025, and general-purpose AI obligations on August 2, 2025. Its current timeline places Annex III high-risk rules on December 2, 2027 and product-integrated Annex I rules on August 2, 2028. Those dates do not classify this pilot. Counsel must check the current legal text, entity roles, use, exceptions, and sector rules before exposure. [C8]

QUESTION	CONSULTING EVIDENCE	REVIEW OWNER
Where and for whom will the system be used?	Locations, users, affected people, entity roles, sector, population.	Legal, privacy, risk, business owner.
What decisions or rights can it affect?	Workflow, outputs, actions, consequence, human authority, appeal.	Legal, compliance, domain and affected-person representation.

QUESTION	CONSULTING EVIDENCE	REVIEW OWNER
What data enters, trains, or leaves the system?	Purpose, lawful basis, source, minimization, processors, retention, rights.	Privacy, security, data owner, procurement.
What classification or duty may apply?	Use-case description, provider/deployer role, transparency, literacy, impact evidence.	Qualified legal and regulatory review.
How will change be detected?	Version inventory, vendor notice, monitoring, periodic reclassification.	Product owner, risk, vendor manager.

The OECD's 2026 responsible-AI due-diligence guidance connects adverse impacts to enterprise policies, assessment, prevention, tracking, communication, and remediation. Use supplier questions to identify who will act on a finding, not simply collect assurance documents. This playbook's question bank adapts that approach to an engagement. [C27]

Ask vendors for evidence, not assurance

- Exact service, model, region, subprocessors, data flows, retention, training use, deletion, and access controls.
- Security and development practices, provenance, vulnerability handling, incident notice, audit evidence, and supply-chain dependencies.
- Evaluation methods, limitations, representative and challenge results, monitoring, known failures, and change-notification terms.
- Service levels, rate and cost behavior, support, continuity, export, portability, deletion, and exit assistance.
- Responsibility for documentation, downstream information, regulatory cooperation, intellectual property claims, and indemnity or liability terms.
- Rights to test, audit, suspend, restrict versions, reconcile state, and terminate when evidence or obligations change.

FIELD NOTE: A certification is one input

A certification or standard alignment can improve diligence, but it does not establish that the configured system, workflow, data, population, and controls are fit for this use. Ask for evidence at the level of the actual service and deployment.

Exercise: assign the decision rights

Harborline's pilot retrieves contract and safety-manual content through a vendor service. Who should be able to approve pilot entry, stop the pilot, accept residual risk, and authorize a new model version?

ANSWER: Name separate accountable decisions

Service Operations accepts workflow fit; IT and Security verify architecture and access; Privacy and Legal review data and obligations; Safety owns safety risk; the Product owner controls versions and evidence; the executive risk owner accepts residual exposure. Any named stop owner may halt the pilot. Restart requires the owners affected by the incident and fresh evidence, not only the vendor's reassurance.

RISK, PERMISSIONS, AND GOVERNANCE MAP

High, Medium, and Low risk chains: failure, exposure, consequence, uncertainty

Preventive, detective, corrective, and recovery controls

RISK, PERMISSIONS, AND GOVERNANCE MAP
Identity, data, context, output, action, and change permissions
Decision rights: propose, approve, operate, change, stop, restart, retire
Vendor evidence, contract gaps, and exit requirements
Legal, privacy, security, safety, labor, and sector reviews needed

RISK, PERMISSIONS, AND GOVERNANCE MAP

Gates, residual risk, owner, review cadence, and escalation

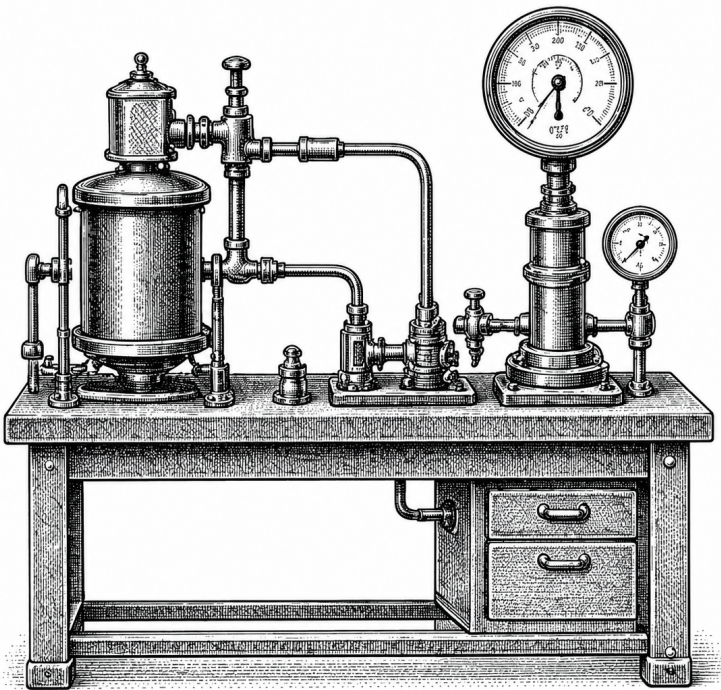
CLAIM: What this work supports

A governance map supports accountable review, control implementation, supplier diligence, and a bounded residual-risk decision. It does not establish compliance or transfer accountability away from the organization deploying the system.

Sources: C1-C15, C18, C20-C22, C25. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Design a pilot that can answer a decision

Choose a representative boundary, comparison, measures, exposure limit, and predeclared next-action rule.



Learning objectives. You will distinguish a demonstration from a pilot, write a testable hypothesis, build separate evidence tracks, define a comparison and exposure ladder, and predeclare entry, exit, stop, and next-decision rules.

A pilot exists to reduce named uncertainty

ACTIVITY	PRIMARY JOB	WHAT IT CANNOT ESTABLISH
Concept or storyboard	Make a workflow and experience discussable.	System feasibility or measured value.
Demonstration	Show a capability on selected examples.	Representative quality, safety, reliability, adoption, or economics.
Technical prototype	Test an architecture, integration, or hard constraint.	End-to-end workflow impact.
Usability study	Observe whether people understand and can use a design.	Production performance or business impact by itself.
Pilot	Test a bounded system and operating model under decision-relevant conditions.	Broad scale, rare-risk absence, or durable value outside its boundary.

Write a hypothesis with an exposure boundary

DECISION: Harborline pilot hypothesis

For eligible requests in two service queues, a coordinator-facing assistant that structures intake, retrieves approved sources, and drafts a work order will reduce median active preparation time and rework without increasing safety, contract, privacy, or service errors. The pilot begins in sandbox and shadow mode, then permits human-approved drafts only if system and governance entry gates pass.

The hypothesis names the workflow, population, intervention, comparison, expected outcomes, protected outcomes, exposure, and action that could follow. Avoid 'prove AI works.' A pilot can support a narrower next step or a stop decision.

Keep four evidence tracks separate

TRACK	QUESTION	EXAMPLE MEASURES
Business and service impact	Did the workflow outcome change?	Cycle time, complete-first-time, rework, service-level attainment, full cost.
System behavior	Did the complete AI system meet its contract?	Field accuracy, citation validity, unsupported claims, reliability, latency, cost.
Risk and governance	Did controls and ownership operate as intended?	Permission violations, High-severity failures, control coverage, incident response.
Human and adoption	Could people use, review, challenge, and support the system?	Eligible use, edits, overrides, review time, trust calibration, workload, help.

CAUTION: Do not average unlike evidence

A faster workflow does not cancel a privacy failure. High adoption does not prove output quality. A strong offline score does not establish business value. Report each track and hard blocker separately before making a combined decision.

Choose a comparison that fits the claim

- Before-and-after: practical, but vulnerable to seasonality, staffing, policy, and learning changes.
- Concurrent comparison: stronger when eligible units or teams can be allocated fairly and spillover is controlled.
- Crossover: useful when teams can use both conditions in a balanced sequence without lasting carryover.
- Matched historical baseline: useful when a live comparison is not possible, but matching and missing-data assumptions must be explicit.
- Qualitative observation: necessary for workflow fit, comprehension, burden, and unexpected consequences; it complements rather than replaces measures.

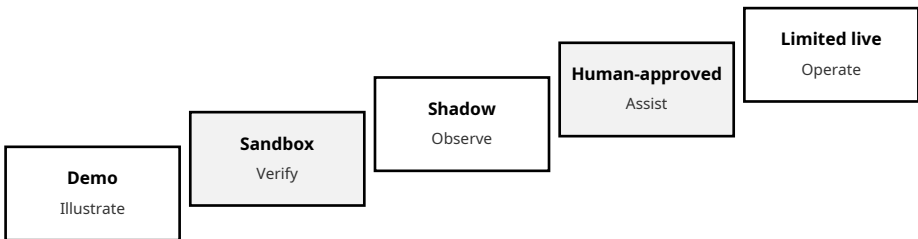
Predeclare the decision rules

RULE	HARBORLINE EXAMPLE
Entry	Approved source inventory, access tests, frozen boundary, trained reviewers, incident path, baseline locked.
Quality floor	Required fields and citations meet named thresholds overall and in safety and contract slices.
Hard blockers	Any cross-customer access, invented safety instruction, unauthorized action, or hidden production write stops expansion.
Workflow outcome	Median active preparation time improves without worse rework, service, or review burden.
Adoption	Eligible use and review behavior show the system fits the work; non-use reasons are understood.
Economics	Observed value remains plausible under full recurring cost and sensitivity ranges.
Next action	Stop, revise, extend, or expand one exposure level; no automatic leap to autonomy.

Limit exposure deliberately

Constrain users, population, locations, data, actions, volume, duration, versions, and hours. Define the kill switch, degraded workflow, reconciliation steps, communication path, and authority to stop. Exposure should grow one supported boundary at a time.

Increase exposure only when the prior boundary is supported



Exercise: reject a weak pilot

A vendor proposes a two-week trial with hand-selected requests, no baseline, and a satisfaction survey. Leaders will decide whether to buy an enterprise license. What is missing?

ANSWER: The evidence does not match the decision

The design can show whether selected users like a demonstration. It cannot support an enterprise purchase. Define eligible population, current baseline, comparison, representative and challenge cases, system and risk gates, usage and workflow measures, full cost, exposure, version control, and a predeclared purchase or stop rule. If those cannot fit, narrow the decision to whether a larger pilot is justified.

Worksheet 6: pilot charter

PILOT CHARTER
Decision, uncertainty, and testable hypothesis
Population, eligibility, slices, duration, volume, users, and locations
Intervention, comparison, baseline, and assignment method
Business and service measures

PILOT CHARTER
System behavior measures and evaluation sets
Risk, governance, human, adoption, and economic measures
Entry gates, floors, hard blockers, exit and stop rules
Exposure, kill switch, fallback, reconciliation, and communication

PILOT CHARTER

Predeclared stop, revise, extend, or scale decision

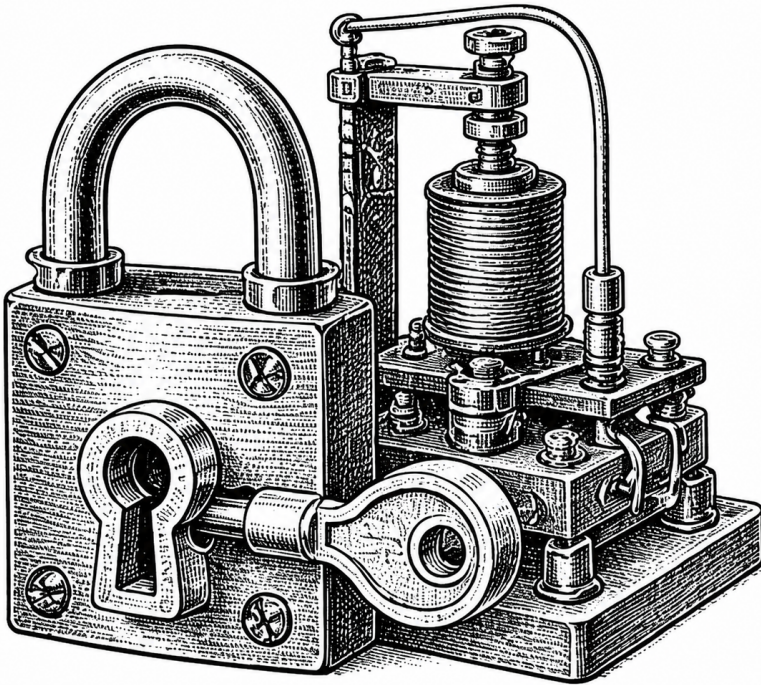
CLAIM: What this work supports

A pilot charter supports a controlled learning decision for a named population and exposure. It does not make the sample representative, the measures valid, or the system ready; those claims depend on execution and evidence quality.

Sources: C1-C3, C16-C18, C23, C25. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Plan delivery around uncertainty

Coordinate product, data, engineering, evaluation, governance, and adoption through visible decisions and dependencies.



Learning objectives. You will choose build, buy, or partner based on the named constraint; organize workstreams around evidence and decision gates; maintain assumptions and dependencies; and define done as accepted behavior, controls, and evidence rather than completed tasks.

The plan should expose what is not yet known

A conventional project plan can hide AI uncertainty inside a sequence of build tasks. A useful delivery plan pairs every major uncertainty with an experiment, owner, evidence, date, and consequence. Some decisions must be made early because they change data access, architecture, vendor terms, pilot exposure, or the claims the work can support.

WORKSTREAM	CORE QUESTION	DECISION EVIDENCE
Product and workflow	What user outcome and behavior contract are we delivering?	Target workflow, prototypes, usability, acceptance.
Data and knowledge	Which permitted current sources make the behavior possible?	Inventory, quality sample, access, provenance, freshness.
Engineering	Can the complete system meet reliability, latency, integration, and recovery needs?	Architecture, prototype, tests, operations plan.
Evaluation	How will quality, safety, slices, and regressions be measured?	Cases, graders, gates, versioned results.
Governance and security	Which controls, reviews, records, and decisions are required?	Risk treatment, threat model, approvals, incident path.
Adoption and operations	Can people use, review, support, and own the workflow?	Role design, training, support, workload, operating owner.
Commercial	Do terms, full cost, continuity, and exit support the use?	Diligence, contract, cost model, portability test.

Choose build, buy, or partner from the constraint

PATH	BEST WHEN	WATCH FOR
Buy	The workflow is close to a supported product, speed matters, and vendor controls and terms fit.	Hidden limitations, generic workflow fit, data route, lock-in, weak test rights, changing service.
Build	The workflow, data, integration, control, or differentiation requires custom behavior.	Underestimated operations, evaluation, security, support, model and tool lifecycle.
Partner	The client needs specialized delivery or temporary capability while retaining ownership.	Blurred accountability, dependency, IP ambiguity, poor knowledge transfer.
Hybrid	A vendor model or platform can support a custom governed workflow.	Responsibility gaps across provider, integrator, client, and downstream systems.
Defer	A prerequisite such as source ownership, baseline, or operating capacity is missing.	Pressure to disguise foundational work as an AI pilot.

FIELD NOTE: A product comparison is not a logo grid

Use representative workflow tasks, failure cases, data and security requirements, operating cost, change terms, and exit tests. A feature list rarely exposes the decision-relevant differences.

Plan by gates, not phases alone

GATE	QUESTION	MINIMUM PACKET
Discovery lock	Is the problem and decision worth continued work?	Decision brief, current-state evidence, opportunity recommendation.
Pilot design	Is the boundary testable and governed?	Target workflow, controls, charter, owners, data and vendor path.
Build entry	Are requirements, sources, architecture, cases, and acceptance ready?	Behavior contract, source inventory, plan, risk and evaluation design.
Pilot entry	Has the complete candidate passed the bounded readiness review?	System manifest, test results, training, support, stop and recovery rehearsal.
Exposure change	Does evidence support the next population or action?	Results by track and slice, incidents, residual risk, owner acceptance.
Handoff	Can the client operate and change the system responsibly?	Accepted runbooks, monitoring, ownership, backlog, vendor and decision history.

Keep evidence and risk records connected

Give each observation, claim, test, risk, decision, and change a stable ID. Link those IDs across the packet. The evidence record says what was seen; the risk record says what could fail and how it is controlled; the delivery ledgers say what the team decided and changed.

RECORD	WORKED HARBORLINE ROW
Evidence E-14	Observed fact: 14/30 requests needed a second lookup. Discovery week; two queues; observation notes O-14; analyst owner. Small convenience sample, not population prevalence. Supports decision D-03 to test retrieval.
Evidence E-41	Test result: HSP-0.8 contract citations 89/100 supported the field. Weeks 4-6; review set CT-08; domain reviewers; Product owner. Multiple passages per draft; not independent requests. Below the 95% floor; D-11 holds expansion.
Risk R-07	Stale manual may distort a safety cue. Data owns source versions and review dates; stale sources block dependent fields. Verify with stale-source challenges and an update drill. Safety accepts only the tested two-queue exposure; restart requires fresh evidence.

Maintain three ledgers

- 1 Assumption ledger: the belief, evidence, owner, test, due date, and consequence if false.
- 2 Decision log: the question, options, evidence, decision maker, date, rationale, dissent, and revisit trigger.
- 3 Change record: the requested change, scope and evidence impact, owner, approval, version, and validation needed.

Use a risk-and-dependency log for threats to delivery, but do not hide product, evidence, or governance decisions inside a generic status field. A material unresolved choice needs a decision owner and date.

Define done at the system boundary

- [1] Named behavior and outcome contract works on representative and challenge evidence.
- [1] Permissions, exclusions, controls, degraded behavior, recovery, and state reconciliation are verified.
- [1] Required owners have accepted the evidence for the exact pilot boundary.
- [1] Training, support, feedback, incident response, and monitoring are ready for real queue conditions.
- [1] Versions, known limitations, residual risks, cost assumptions, and non-claims are documented.
- [1] The next decision and the evidence it will use are scheduled before exposure begins.

Exercise: respond to a delivery shortcut

The team can meet the pilot date only by skipping source-access integration and pasting contract text into prompts manually. The pilot aims to measure real workflow time. Should it proceed?

ANSWER: The shortcut changes the claim

Manual paste may support a technical or usability study, but it cannot measure the intended end-to-end workflow or permission model. Either change the decision and label the smaller study, or move the date. Do not keep the pilot label while removing the dependency that determines its value and risk.

Worksheet 7: integrated delivery plan

INTEGRATED DELIVERY PLAN
Decision gates, dates, owners, and minimum packets
Product and workflow workstream
Data and knowledge workstream
Engineering and operations workstream

INTEGRATED DELIVERY PLAN
Evaluation, governance, privacy, and security workstreams
Adoption, training, support, and commercial workstreams
Build, buy, partner, hybrid, or defer decision
Assumptions, dependencies, decisions, risks, and change control

INTEGRATED DELIVERY PLAN

Definition of done for build entry, pilot entry, and handoff

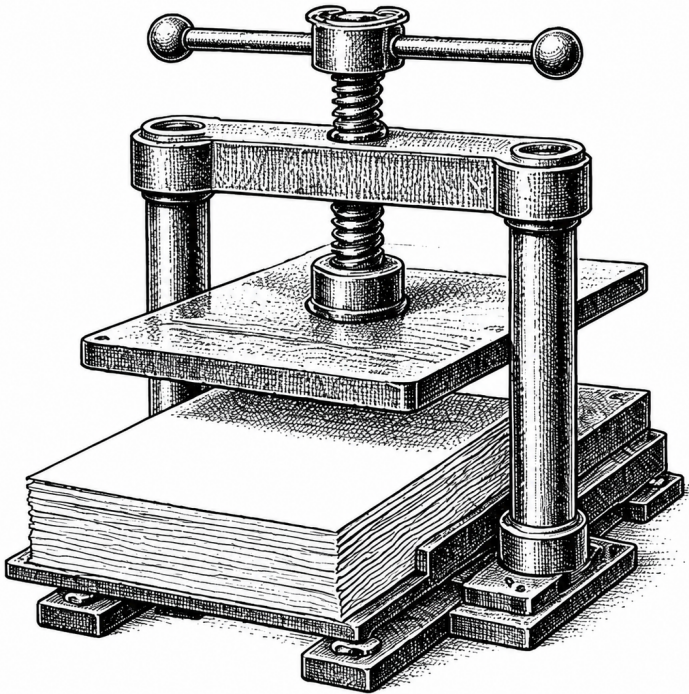
CLAIM: What this work supports

An integrated plan supports coordination and makes decision risk visible across workstreams. It does not eliminate uncertainty or authorize a gate; the named evidence and owners still have to accept the boundary.

Sources: C3-C5, C13-C16, C20-C22, C27. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Build and verify the whole pilot system

Version the complete workflow and require evidence for behavior, controls, recovery, and operations before exposure.



Learning objectives. You will define the complete system under test, build a traceable manifest, connect requirements to verification, distinguish technical checks from owner acceptance, and run a readiness review that can hold or narrow the pilot.

The model is one component

The pilot system includes identity, policy, interface, instructions, context, retrieval, model and provider, tools, validators, permissions, telemetry, human review, support, fallback, and external state. Changing any response-affecting component can weaken prior evidence.

MANIFEST AREA	RECORD
Workflow	Pilot population, eligibility, task, UI, human role, fallback, and operating hours.
Application	Code commit, configuration, feature flags, schemas, prompts, routing, tools, time and cost budgets.
Models and vendors	Provider, service, model snapshot or alias, region, settings, contract and data terms.
Knowledge and data	Source IDs, owners, versions, access rules, index build, freshness, retention, and lineage.
Evaluation	Case IDs, set purposes, rubrics, graders, judge versions, thresholds, run settings, and results.
Controls	Authentication, authorization, validation, approval, logging, rate limits, kill switch, rollback.
People and process	Reviewers, training version, escalation, support, decision owners, and run date.

Build a verification matrix

REQUIREMENT	METHOD	EVIDENCE	OWNER
No cross-customer source access	Exact authorization and adversarial tests	Request, policy decision, source and denial trace	Security and data
Draft cites approved current evidence	Representative and stale-source challenge cases	Citation validity, source version, unsupported-claim review	Product and domain
No autonomous production write	Architecture inspection and runtime state check	Tool list, permissions, audit and resulting state	Engineering and risk
Coordinator can review meaningfully	Usability, queue simulation, workload observation	Comprehension, edit, override, time and error evidence	Operations and design
Failure returns to safe workflow	Timeout, provider, retrieval and partial-state rehearsal	Visible status, no hidden commit, recovery and reconciliation	Operations
Full cost remains bounded	Load and workflow measurement	Tokens, requests, tools, review, support and sensitivity	Product and finance

Use different methods for different claims

- Exact tests for schemas, fields, calculations, permissions, tool calls, external state, and hard invariants.
- Representative evaluations for expected workflow cases and meaningful slices.
- Challenge and red-team cases for boundary, safety, privacy, injection, misuse, outage, and recovery behavior.
- Human evaluation for context-dependent quality and domain judgment, with instructions, independence, disagreement, and adjudication.
- Usability and workflow studies for comprehension, review behavior, burden, and error recovery.
- Load, resilience, observability, and cost tests for real operating conditions.

CAUTION: Validate the grader as well as the system

A polished rubric or model judge is still a measurement component. Use an independent human reference process, inspect error by important class and slice, test bias and injection sensitivity, and limit what the grader is permitted to decide.

For more detail, Michael Long's AI Evaluation Field Guide (Version 1.0.0) covers test sets, human rating, judges, uncertainty, and launch decisions (handbooks.mikelong.ai/ai-evaluation/). His Production AI Engineering (Version 1.4.0) covers the complete runtime, retrieval, tools, monitoring, and failure design (handbooks.mikelong.ai/production-ai-engineering/). Start here by naming the evaluated unit, separating development from decision cases, writing pass and stop rules, having domain reviewers check representative and failure cases, and preserving counts, disagreement, versions, and limitations. A technical owner must verify permissions and resulting state independently of the draft.

Rehearse failure before the pilot

FAILURE	EXPECTED BEHAVIOR	EVIDENCE TO PRESERVE
Source unavailable	Show unavailable state; do not fabricate; return to current lookup path.	Request, source status, fallback, user action.
Provider timeout	End bounded attempt; retain draft state safely; permit retry or manual path.	Timing, attempt count, user state, retry result.
Stale or conflicting evidence	Expose sources and conflict; block unsupported field; escalate.	Versions, conflict, block reason, owner.
Malicious retrieved instruction	Treat source as data; prevent permission or destination change.	Input lineage, policy decision, tool and output trace.
Partial external write	Stop further action; reconcile authoritative state; notify owner.	Idempotency, request, result, resulting state, reconciliation.
High-severity finding	Stop affected exposure; preserve evidence; assess affected units; require fresh gate.	Finding, versions, population, containment and restart decision.

Run an evidence-based readiness review

- [1] The system manifest identifies the exact candidate and pilot boundary.
- [1] All hard requirements have a method, result, limitation, owner, and evidence link.
- [1] High-severity findings are closed, controlled by architecture, or keep the exposure blocked.
- [1] Representative, challenge, slice, usability, security, privacy, reliability, recovery, and cost evidence match the decision.
- [1] Reviewers, support, monitoring, stop, incident, reconciliation, and communication paths have been rehearsed.
- [1] Product, Operations, Engineering, Data, Security, Privacy/Risk, and executive owners accept only the boundary their evidence covers.

Exercise: hold the right boundary

The candidate drafts strong work orders, but the runtime still exposes a production write tool that the prompt says not to use. No test observed a write. Is the pilot ready?

ANSWER: No prompt-only security boundary

Remove or hard-disable the write capability for the pilot identity, verify the exact runtime tool and authorization state, rerun affected tests, and update the manifest. Zero observed writes does not establish that an available consequential action is controlled.

Worksheet 8: readiness evidence packet

READINESS EVIDENCE PACKET
Complete system manifest and candidate identity
Requirement-to-method-to-evidence matrix
Representative, challenge, slice, usability, and human-review results

READINESS EVIDENCE PACKET
Security, privacy, permission, vendor, and supply-chain results
Reliability, latency, cost, fallback, recovery, and reconciliation results
Open findings, severity, owner, mitigation, and affected boundary
Training, support, monitoring, incident, kill-switch, and rollback rehearsal

READINESS EVIDENCE PACKET

Owner acceptance, permitted exposure, non-claims, and fresh evidence required

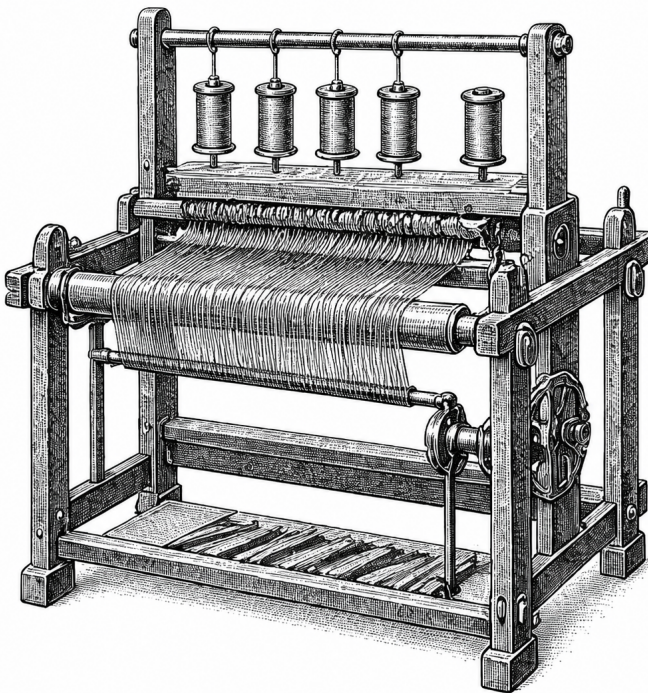
CLAIM: What this work supports

A readiness packet supports entry into the exact pilot boundary accepted by named owners. It does not establish production success, legal compliance, or human approval for broader populations, actions, or later versions.

Sources: C1-C3, C12-C18, C21, C23. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Prepare the people who will use and support it

Design the role, review, training, support, and feedback work with the people who will carry it.



Learning objectives. You will map role and task changes, diagnose rational reasons for non-use, design meaningful review and escalation, build role-specific literacy, and measure adoption as observable workflow behavior rather than enthusiasm.

Adoption starts in the target workflow

People adopt a system when it helps them complete accountable work under real conditions. They may reject it because it adds review, hides evidence, threatens role clarity, creates new risk, or performs poorly on the cases that define their expertise. Treat those reasons as product and operating evidence before labeling them resistance.

CHANGE	QUESTION	HARBORLINE EXAMPLE
Task	What is added, removed, accelerated, or made harder?	Less manual drafting; new citation and missing-field review.
Decision	Who now proposes, reviews, decides, and records rationale?	Assistant proposes; coordinator retains decision and edit authority.
Knowledge	Which expertise becomes visible, encoded, or newly required?	Coordinators need source and AI-failure literacy; exceptions remain domain work.
Workload	Where do time, queue pressure, and cognitive burden move?	Preparation may fall while review and feedback initially rise.
Responsibility	Who is accountable when the system is wrong or unavailable?	Operations owns the work; product and technical owners own system correction.
Identity and incentives	What status, autonomy, metric, or job concern changes?	Experts may see a drafting tool as devaluing judgment or creating surveillance.
Support	Who helps, how fast, and with what evidence?	Named supervisor path, issue capture, status visibility, manual fallback.

Involve affected people at decision points

- 1 Discovery: observe the work and validate which problems, exceptions, and outcomes matter.
- 2 Boundary: review what the system may do, what stays human, and who could be affected by failure.
- 3 Design: test source visibility, correction, override, escalation, and recovery in realistic conditions.
- 4 Pilot entry: confirm training, capacity, support, feedback use, and the right to use the fallback path.
- 5 Evidence review: include frontline interpretation of usage, burden, workarounds, and unintended effects.
- 6 Scale and handoff: agree role design, staffing, operating ownership, continuing literacy, and challenge routes.

CAUTION: State which decisions people can influence

Be clear about which decisions affected people can shape, which constraints are fixed, how input will be used, and who remains accountable. Asking for feedback after the material decisions are locked damages trust without improving the system.

Build role-specific AI literacy

Literacy should match the person's role, knowledge, context, and the system's risks. It includes what the system does, where it fails, how to verify evidence, which data and actions are permitted, when to challenge or stop, and how issues are reported. Completion of a generic course is not proof that a person can operate this workflow safely.

ROLE	NEEDS TO KNOW AND PRACTICE
End user	Purpose, boundary, source review, uncertainty, correction, override, privacy, fallback, issue reporting.
Supervisor	Queue effects, exception policy, support, review quality, incident escalation, coaching without punishing overrides.
Product owner	Behavior contract, versions, evidence, change control, adoption, monitoring, release and retirement.
Technical operator	Architecture, permissions, telemetry, service changes, failure, recovery, reconciliation, cost.
Risk and control owner	Affected people, obligations, evaluation limits, incidents, residual risk, review triggers.
Executive decision owner	Value, exposure, evidence quality, hard blockers, total cost, accountability, scale conditions.

Check access to the workflow

Test source review, edits, override, help, and fallback with people who use keyboards, assistive technology, or alternative ways of perceiving the interface. WCAG 2.2 provides testable web-accessibility criteria; it does not establish that a particular AI workflow is usable or that an employment duty has been met. Include accessibility findings in pilot entry and handoff evidence. [C28]

Design training around work

- Use representative and failure cases from the actual boundary, with sensitive details governed appropriately.
- Practice accepting a good result, correcting a plausible error, rejecting a dangerous result, escalating uncertainty, and using fallback.
- Show how feedback is interpreted and which reports trigger product, data, policy, or support action.
- Assess demonstrated behavior, not attendance alone. Retrain when role, model, interface, data, policy, or risk changes materially.
- Provide job aids at the decision point: source checks, prohibited actions, escalation reasons, and stop conditions.

Measure adoption without blaming people

MEASURE	WHAT IT CAN SHOW	WHAT TO INVESTIGATE
Eligible use	Whether the system is used when available and applicable.	Availability, speed, fit, training, trust, workload, incentives.
Completion through workflow	Whether users reach a valid reviewable outcome.	Abandonment, fallback, missing data, confusing states.
Edit and override	Where proposals differ from accepted work.	Quality, ambiguity, policy, user strategy, over- or under-reliance.
Review time	The burden of responsible use.	Whether time moved rather than disappeared; queue conditions.
Feedback and help	Where people need support or encounter novelty.	Severity, responsiveness, psychological safety, duplicate issues.
Workaround	Where the designed system does not fit.	Whether the workaround protects work or creates new risk.

Exercise: diagnose low use

Harborline coordinators use the assistant on only 35% of eligible requests. Leaders call them resistant. Observation shows source retrieval often takes 20 seconds, and reviewers must reopen two systems to verify citations. What should happen next?

ANSWER: Treat non-use as evidence

Investigate eligible cases and queue conditions, confirm latency and verification burden, observe the fallback decision, and repair the workflow before increasing training or pressure. Report which non-use is rational. Do not use adoption targets to force people into an inferior or risky path.

Worksheet 9: people and adoption plan

PEOPLE AND ADOPTION PLAN
Affected roles, people, groups, and involvement points
Task, decision, knowledge, workload, accountability, and incentive changes
Meaningful human review and fallback in real conditions
Role-specific literacy objectives and practice cases

PEOPLE AND ADOPTION PLAN
Training, assessment, job aids, support, and escalation
Adoption measures: eligible use, completion, edits, overrides, time, help, workarounds
Reasonable causes of non-use and how they will be investigated
Feedback-to-decision path, owner, response time, and communication

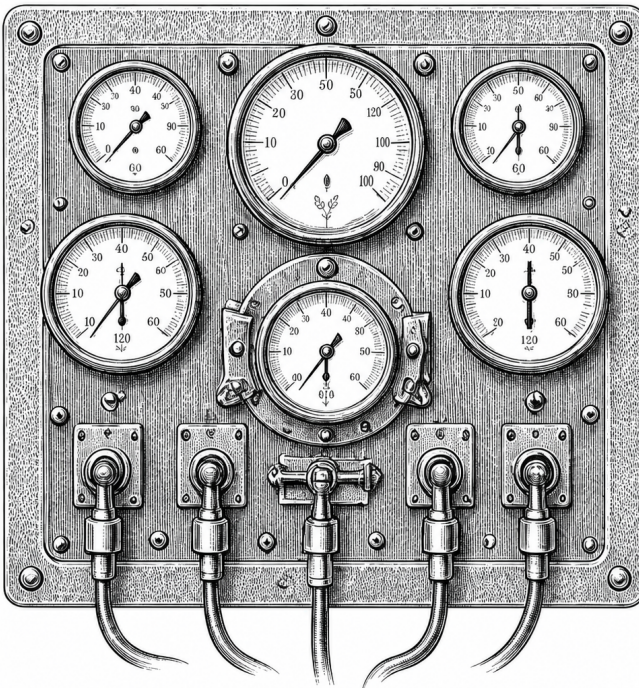
CLAIM: What this work supports

A people and adoption plan supports responsible pilot participation, role-specific preparation, and interpretation of workflow behavior. It does not prove acceptance, prevent job impact, or replace worker, accessibility, labor, or legal review.

Sources: C5, C9, C16, C19, C24, C26, C28. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Run the pilot as a controlled evidence loop

Use a stable cadence to see what changed, separate causes, control versions, and respond without damaging the evidence.



Learning objectives. You will operate a weekly evidence board, separate system, workflow, policy, data, and adoption causes, control changes, respond to incidents, and preserve enough evidence to make a defensible next decision.

Make the evidence visible while the pilot runs

A pilot is not a period of casual use followed by a survey. The team needs a regular view of exposure, versions, workflow outcomes, system behavior, risk, human behavior, economics, issues, changes, and stop conditions. Review exceptions and slices, not only aggregates.

WEEKLY BOARD	MINIMUM VIEW
Exposure	Eligible units, attempted units, users, queues, versions, data and action boundary.
Business/service	Baseline comparison, time, quality, rework, service and important slices.
System	Contract measures, failures, citations, reliability, latency, cost and regressions.
Risk/control	Hard blockers, High/Medium/Low findings, permission and control events, residual exposure.
Human/adoption	Eligible use, completion, edit, override, review time, fallback, help and workaround.
Issues/incidents	New findings, affected versions and units, containment, owner, status, decision impact.
Changes	What changed, why, authorization, expected effect, evidence invalidated, validation needed.
Decision	Continue, narrow, pause, stop, revise, or propose the next exposure.

Classify the cause before choosing the fix

CAUSE	DIAGNOSTIC QUESTION	EXAMPLE RESPONSE
Workflow	Does the target flow add delay, ambiguity, duplicate work, or a bad handoff?	Redesign the state, review, or integration.
Policy	Is the expected behavior unclear or disputed?	Resolve policy and update cases, training, and controls.
Data/knowledge	Is source content missing, stale, inaccessible, conflicting, or poorly indexed?	Fix ownership, access, freshness, retrieval, or provenance.
Model/context	Does the candidate fail despite correct inputs and contract?	Improve context, prompt, model route, or narrow task.
Tool/integration	Did a read, write, schema, timeout, or external state fail?	Fix contract, authorization, idempotency, recovery.
Measurement	Is the case, grader, baseline, or instrumentation misleading?	Repair the measure and rerun fresh evidence.
Adoption and operations	Are capacity, incentives, training, support, or trust shaping behavior?	Change operating conditions and reassess.

FIELD NOTE: Do not ask the model to repair an organizational disagreement

When domain experts disagree about the correct policy or acceptable consequence, prompt tuning hides the decision. Resolve the rule, document the owner and exceptions, then update the system and evidence.

Control pilot changes

- 1

Capture the issue and exact affected system, data, user, and workflow state.
- 2

Classify severity and whether the current exposure can continue safely.
- 3

Choose the smallest change that addresses the diagnosed cause.
- 4

Record every response-affecting version and which prior results remain applicable.
- 5

Run targeted regression plus any fresh representative, challenge, usability, or operational evidence required.
- 6

Authorize the new candidate and communicate the change to users and owners before resuming exposure.

Frequent improvement is compatible with a controlled pilot, but the evidence must remain attributable. If the system changes every day, report version-specific results and reserve a frozen period for the decision the pilot is meant to support.

Respond to an incident in six moves

MOVE	ACTION
Contain	Stop or narrow affected exposure; revoke or disable risky capability.
Preserve	Save prompts, context, sources, tools, policy decisions, traces, external state, versions, and timing.
Assess	Identify affected people, units, systems, rights, service, data, and obligations.
Correct	Repair records, notify required parties, reconcile state, and support affected users.
Learn	Find system and organizational causes; update risk, cases, controls, training, vendor and process.
Fresh gate	Require fresh evidence and named approval for the exact restart boundary.

Harborline mid-pilot incident

A stale manual passage leads to a wrong urgency cue in a draft. The coordinator catches it, but trust drops and usage falls. The source record shows the manual owner replaced the PDF without updating the approved index.

DECISION: Expected response

Pause the affected equipment class, preserve the draft and source lineage, verify whether other units used the stale version, correct the source inventory and freshness control, add regression and nearby variants, update the user-facing source state, explain the issue and fix, and require Service Safety, Data, Product, and Risk approval before resuming that class. Do not solve the incident only with a prompt warning.

Measure the learning system too

- Time from finding to triage, containment, owner assignment, fix, validation, communication, and closure.
- Repeat findings, reopened issues, stale backlog, unsupported workarounds, and overdue decisions.
- Coverage of production findings in regressions, risk controls, training, vendor review, and monitoring.
- Whether frontline feedback changes the product or disappears into an unowned channel.
- Whether model, source, policy, vendor, workflow, and organizational changes trigger appropriate revalidation.

Exercise: preserve the decision

A team improves the prompt after every bad case and reports one aggregate end-of-pilot score. What is wrong with the claim?

ANSWER: The candidate and evidence are mixed

The aggregate combines different systems and may adapt to inspected cases. Keep version-specific results, classify inspected cases as development or regression evidence, freeze the decision candidate, and evaluate it on protected representative and challenge evidence under the predeclared rules.

Worksheet 10: pilot evidence board

PILOT EVIDENCE BOARD
Week, exact candidate, population, exposure, users, data, and actions
Business and service results versus baseline and comparison
System behavior, reliability, latency, cost, and important slices
Risk, controls, human review, adoption, burden, help, and workarounds

PILOT EVIDENCE BOARD
Issues and incidents: evidence, severity, affected boundary, containment, owner
Changes: cause, authorization, versions, evidence invalidated, validation
Decision: continue, narrow, pause, stop, revise, or propose next exposure
Open questions and evidence needed next week

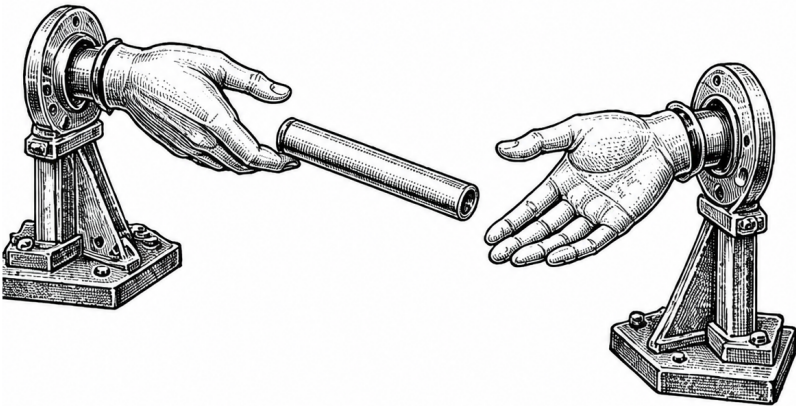
CLAIM: What this work supports

A controlled evidence board supports version-attributable learning and timely exposure decisions. It does not make the pilot representative or prove that findings will generalize beyond the tested workflow, population, period, and operating conditions.

Sources: C1-C3, C12-C18, C23. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Turn the evidence into the next decision

Interpret quality, risk, adoption, workflow impact, and economics without rounding mixed evidence up to scale.



Learning objectives. You will synthesize separate evidence tracks, calculate bounded economics, state uncertainty and claim limits, choose stop, revise, extend, or scale, and define the prerequisites for any wider exposure.

Start with the decision, then read every track

TRACK	DECISION QUESTION	DO NOT HIDE
Service impact	Did the end-to-end user or business outcome improve?	Slices, wait and review time, downstream rework, unintended effects.
System quality	Did the frozen complete system meet the behavior contract?	Counts, uncertainty, grader limits, version and conditions.
Risk and control	Did hard requirements and controls hold?	High-severity findings, residual risk, incidents, untested paths.
People and adoption	Could people use, challenge, support, and own the workflow?	Non-use reasons, burden, overreliance, workarounds, job impact.
Economics	Is the observed outcome plausible at full recurring cost?	Review, support, governance, incidents, integration, variability, exit.
Operating readiness	Can the organization run and change the system responsibly?	Owners, staffing, monitoring, vendor, recovery, backlog, literacy.

Use four next-action choices

DECISION	USE WHEN	REQUIRED OUTPUT
Stop	The outcome is weak, a blocker is disproportionate, or prerequisites make the path unattractive.	Reason, preserved learning, obligations, offboarding and affected-state correction.
Revise	The opportunity remains useful but system, workflow, policy, data, measure, or operating design must change.	Diagnosed cause, new candidate boundary, affected evidence and fresh gate.
Extend	Evidence is promising but duration, sample, slice, season, or operating condition is insufficient.	Exact uncertainty, additional exposure, controls, measures, stop rule and end date.
Scale	Every required track and hard gate supports a named wider boundary and ownership is ready.	Population/action increment, operating capacity, monitoring, rollback and next review.

DECISION: Scale one boundary at a time

Expand users, volume, data, geography, action, or autonomy only when evidence supports that specific change. A successful assistive pilot does not establish that the system may decide or act autonomously.

Build the economics from observed units

Use the pilot unit of work and a range, not a single projected ROI. Start with eligible volume, observed time or outcome change, loaded labor or service value, recurring system cost, human review, support, governance, monitoring, incidents, and change. Keep cash savings, capacity, avoided cost, revenue, quality, and risk reduction distinct.

LINE	PLAIN-LANGUAGE CALCULATION	CAUTION
Gross task capacity	available eligible units x use rate x mean net minutes / 60	If units already count uses, omit use rate. Use a mean, not a median gap. Observational differences do not prove causal savings.

LINE	PLAIN-LANGUAGE CALCULATION	CAUTION
Quality effect	change in rework or error units x evidenced consequence	Use credible consequence and include new error classes.
Service effect	change in throughput, delay, or attainment x evidenced value	Avoid assigning revenue without a causal path.
Recurring system cost	model + platform + tools + data + infrastructure + vendor	Include variability, peaks, minimums, and provider change.
Human operating cost	review + support + monitoring + training + governance + incident response	Do not assume human work disappears.
Change and exit	integration + migration + validation + contract + portability + retirement	Include refresh and switching, not only launch.
Net range	evidenced benefit range minus full cost range	Run low, expected, and high cases with explicit assumptions.

Harborline decision snapshot

EVIDENCE	FINDING	IMPLICATION
Workflow	Preparation time and complete-first-time improve in two queues; review remains material.	Assistive workflow has value; capacity conversion needs operating plan.
System	Overall draft and citation results are promising; contract citations miss their floor and the safety set is small.	Revise retrieval and collect fresh contract and safety evidence before expansion.
Risk	No unauthorized action observed after tool removal; stale-source incident exposed a governance gap now controlled.	Keep write actions excluded; audit freshness control during expansion.
Adoption	Later-candidate use is higher; different periods and denominators limit attribution. Supervisors need protected support capacity.	Fund support and monitor review burden.
Economics	Expected capacity value falls below first-year cost when source cleanup and internal ownership are included.	Tie expansion to eligible volume and review time; avoid booked savings.
Ownership	Operations, Product, IT, Data and Risk accept the assistive boundary; no owner accepts autonomous dispatch.	Keep the existing assistive boundary; expansion remains conditional.

DECISION: Expected recommendation

Revise the contract retrieval and review path within the existing two queues. Keep priority, assignment, customer commitments, and autonomous writes blocked. Hold expansion until the contract floor passes, protected outcomes are measured, vendor deletion and provider-change rehearsals close, and the full cost model supports a named next boundary.

Write the claim boundary

For candidate version ____, in population and period ____, compared with ____, the evidence showed ____ across workflow, system, risk, human, and economic tracks. This supports decision ____ within boundary ____.

Important uncertainty is _____. It does not establish ____, and wider exposure requires _____.

Exercise: refuse the rounded-up claim

A pilot reduces average preparation time by 30%, but the comparison period had lower volume, review time rose, and one High-severity safety case failed. The sponsor wants to say 'AI increased productivity by 30%.'

ANSWER: Keep the result and its limits together

Report the exact preparation-time result, volume difference, review-time change, design limitation, and safety blocker. Do not translate task time into productivity or scale readiness. Recommend containment and fresh evidence for the failure, then choose stop, revise, or extend under a bounded claim.

Worksheet 11: outcome and decision record

OUTCOME AND DECISION RECORD
Decision, candidate, population, period, comparison, and exposure
Service and workflow evidence with slices and limitations
System quality, reliability, security, privacy, and control evidence

OUTCOME AND DECISION RECORD
Human review, adoption, workload, feedback, and unintended effects
Full economics: benefit, recurring cost, human cost, change, exit, sensitivity
Hard blockers, residual risk, uncertainty, and disconfirming evidence
Stop, revise, extend, or scale recommendation and exact next boundary

OUTCOME AND DECISION RECORD
Prerequisites, owners, evidence, stop rules, review date, and non-claims

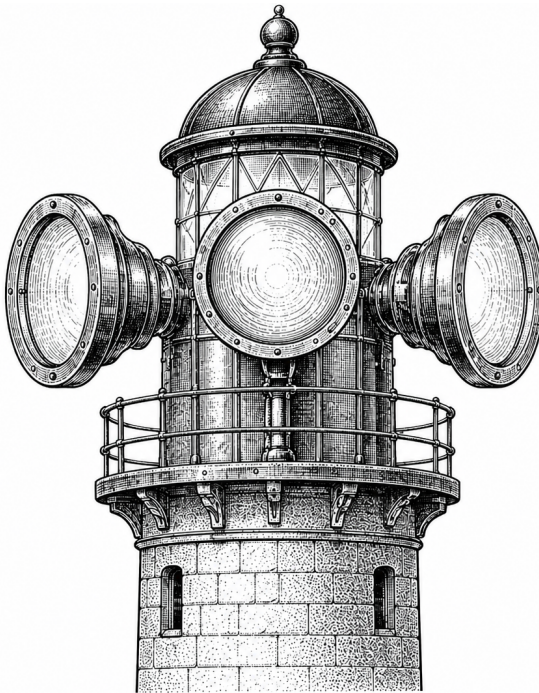
CLAIM: What this work supports

An outcome and decision record supports a named next action within the evaluated boundary. It does not establish causality beyond the design, guarantee future value, or authorize wider populations, data, actions, or autonomy without new evidence and approval.

Sources: C1-C5, C16-C19, C23. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Leave the client able to continue

Transfer accepted artifacts, decision rights, routines, and capability so the system no longer depends on the consulting team.



Learning objectives. You will define ownership across the system lifecycle, create an accepted handoff manifest and operating runbook, assess capability gaps, plan the first 90 days, and close the engagement without leaving invisible consultant dependency.

Handoff is an operating change

Document delivery is not handoff. The client has to know what exists, why it was designed this way, what evidence supports it, what remains uncertain, who owns each decision, how the system is monitored and changed, and how to stop or retire it. Owners should demonstrate those routines before the consultant leaves.

OWNERSHIP	ACCOUNTABLE WORK
Business/service	Outcome, workflow policy, service risk, priority, funding, affected-person consequence.
Product	Behavior contract, roadmap, boundary, evidence, versions, change and next decision.
Technical	Application, integration, model route, tools, reliability, telemetry, recovery, cost.
Data/knowledge	Source ownership, permission, quality, provenance, freshness, retention, deletion.
Security/privacy/risk	Controls, obligations, threat and impact review, residual risk, incidents, audit.
Operations/support	Users, queue, review, fallback, support, escalation, communication, reconciliation.
Vendor/commercial	Service and model changes, contract, performance, incident, renewal, portability, exit.
Executive	Exposure, funding, residual risk, scale or stop, and organizational accountability.

Transfer the complete record

PACKET	MINIMUM CONTENTS
Decision history	Briefs, options, evidence, rationales, dissent, scope and change records.
System	Architecture, manifest, code/configuration locations, data flows, sources, permissions, vendors.
Evidence	Cases, set purposes, methods, graders, results, limitations, incidents, acceptance and non-claims.
Operations	Runbook, monitoring, alerts, dashboards, support, incident, rollback, reconciliation, retirement.
People	Role design, training, job aids, access, review expectations, feedback and escalation.
Commercial	Contracts, renewals, costs, usage limits, service changes, contacts, export, deletion and exit.
Future work	Known limitations, risks, technical debt, research questions, backlog, decision dates and owners.

CAUTION: Remove hidden consultant dependencies

Check for personal accounts, undocumented scripts, private dashboards, external prompts, untransferred source files, consultant-owned vendor relationships, unexplained judgments, and review routines that only work when the project team is present.

Use acceptance demonstrations

- [1] Product owner identifies the exact production or pilot boundary and explains the next change gate.
- [1] Technical operator identifies the running version, traces a request, handles an outage, and executes rollback or degraded mode.
- [1] Data owner updates or retires a source, verifies permission and freshness, and explains lineage.
- [1] Operations handles a bad result, override, support request, fallback, incident and affected-state reconciliation.
- [1] Risk owner locates evidence, understands limitations, reviews residual risk, and knows stop and restart authority.
- [1] Vendor owner verifies service-change notice, cost, renewal, audit, export, deletion and exit path.
- [1] Executive owner can state the supported value, unresolved risk, full operating commitment, and next decision date.

Build the operating cadence

CADENCE	REVIEW
Continuous	Availability, errors, security, permission, cost, hard stop and incident signals.
Weekly	Workflow outcomes, quality samples, adoption, support, changes, issues, slices and owner actions.
Monthly	Trend, drift, full cost, vendor behavior, training, backlog, risk and control performance.
Quarterly or risk-based	Boundary, value, affected people, reclassification, external obligations, roadmap and retirement.
Event-triggered	Model/provider change, source change, incident, policy or law change, new population, data, action or location.

Plan the first 90 days

PERIOD	FOCUS
Days 0-30	Shadow the operating owners; close access and documentation gaps; run incident, rollback, source-update, and vendor-change drills.
Days 31-60	Client owners lead weekly evidence review and one controlled change; consultant observes only where agreed.
Days 61-90	Client runs review, change, revalidation, and executive decision cadence; resolve remaining capability gaps or narrow exposure.
Exit	Named owners sign acceptance; consultant access, accounts, data, tools, and support obligations are closed or transferred.

Retirement is part of ownership

- Define triggers: weak value, unsupported provider, control failure, new obligation, excessive cost, replaced workflow, or owner loss.
- Stop new use, preserve required evidence, reconcile external state, communicate, revoke access, export or delete data, and end vendor services.
- Retain the decisions, incidents, measures, and lessons needed for audit and future systems under the applicable policy.
- Verify that affected people have a working replacement or fallback and know how to challenge lingering outcomes.

Exercise: refuse a paper handoff

The final packet is complete, but the product owner has never run the evidence review, the data owner cannot update the source index, and the incident drill still depends on the consultant. Can the engagement close?

ANSWER: The artifacts are delivered; ownership is not transferred

Keep the system at the existing boundary, close the capability gaps, run acceptance demonstrations, and either extend a defined transition or assign an accountable internal or vendor operator. Do not label dependency as knowledge transfer.

Worksheet 12: handoff and closeout

HANDOFF AND CLOSEOUT RECORD
Business, product, technical, data, risk, operations, vendor, and executive owners
Accepted decision history, system, evidence, operations, people, commercial, and backlog packets
Operating runbook: monitor, support, change, evaluate, incident, rollback, reconcile, retire

HANDOFF AND CLOSEOUT RECORD
Acceptance demonstrations, results, gaps, and remediation owners
Continuous, weekly, monthly, periodic, and event-triggered cadence
30/60/90-day transition plan
Consultant access, accounts, data, tools, vendors, IP, support, and deletion closeout

HANDOFF AND CLOSEOUT RECORD

Final supported boundary, residual risk, non-claims, and next decision date

CLAIM: What this work supports

An accepted handoff record supports a controlled transfer of the existing system and its decision routines. It does not guarantee the client will operate them well; ongoing ownership, review, evidence, and leadership remain necessary.

Sources: C1-C5, C12-C14, C16, C18-C22, C24. The consulting sequence, worksheets, and Harborline case are author synthesis. Full citations appear in Further reading.

Guided capstone: Harborline decision review

Use the fictional evidence packet to make a bounded consulting recommendation and handoff plan. Allow two to three hours. The packet contains enough information to make a defensible decision, but some evidence is intentionally incomplete or in tension.

DECISION: Your job

Recommend stop, revise, extend, or scale. Name the exact workflow, population, data, actions, controls, and owners your recommendation permits. Identify the evidence that drives the decision, the uncertainty that remains, and the claims the packet cannot support.

1. Sponsor memo

Harborline maintains commercial refrigeration and food-service equipment across four regions. Peak season begins in twelve weeks. The COO believes AI dispatch could reduce response time and wants an enterprise recommendation. The VP of Service wants fewer incomplete work orders. The CIO wants one approved platform. The Safety Director will not accept automation that changes urgency, assigns a technician, or communicates a commitment without accountable review. Coordinators remember a prior trial that produced fluent but unsupported answers.

STAKEHOLDER	DESIRED OUTCOME	CONCERN	DECISION AUTHORITY
COO	Capacity and faster service before peak.	Slow transformation and fragmented tools.	Funds exposure and accepts enterprise risk.
VP Service	Complete work orders and predictable queues.	Review burden and disruption.	Owens workflow and service outcome.
CIO	Supported architecture and vendor model.	Security, integration, cost, lock-in.	Approves technical operating model.
Safety Director	Correct escalation and protected work.	Stale evidence, automation bias, hidden decisions.	Owens safety policy and stop/restart for safety exposure.
Coordinators	Less searching and duplicate entry.	Bad drafts, surveillance, loss of judgment, extra review.	Perform and accept the daily workflow.

2. Current-state evidence

MEASURE	BASELINE	IMPORTANT NOTE
Eligible service requests	About 2,400 per month across four regions	Volume varies 35% by season; pilot queues represent 38% of current volume.
Median active preparation	12.4 minutes	Includes intake, lookup, clarification, and work-order drafting.
Median elapsed to reviewable order	46 minutes	Waiting for customer or supervisor drives the tail.
Complete on first review	61%	Serial number, contract, and safety fields cause most rework.
Safety escalation	3.8% of requests	Rare but consequential; policy language varies by equipment.
Contract exception	11% of requests	Sources span two systems; 7 of 100 contracts sampled during discovery have inconsistent labels.
Coordinator systems	Five commonly used	Two lack single sign-on; source switching is a major burden.

Observations of 30 requests found that coordinators often start the work order before all evidence is available, then reopen it after checking manuals and contracts. Fourteen needed a second source lookup. Eight used a personal note or saved link. Supervisors described urgent safety cases as the point where expert judgment and phone escalation matter most.

3. Opportunity choices

OPPORTUNITY	EVIDENCE	MAIN DEPENDENCY OR RISK
Structured intake	Missing fields and inconsistent language drive rework.	Channel integration and field policy.
Approved evidence retrieval	Search and source switching add time; staff need attributable answers.	Access, ownership, source freshness, contract label cleanup.
Work-order drafting	Rewriting consumes time and follows a repeatable structure.	Unsupported facts, review design, measurement.
Priority recommendation	Leaders expect queue benefit; no clean historical ground truth.	Safety, policy disagreement, bias, meaningful oversight.
Technician assignment	Potential scheduling benefit is not baselined.	Certification, geography, availability, union rules, permissions, writes.
Customer communication	May reduce callbacks.	Commitments, tone, consent, contract and state reconciliation.

4. Data, architecture, and vendor packet

ITEM	FINDING
Requests	Contain customer contact, account, location, equipment, symptom, free text, attachments, and sometimes sensitive information.
Manuals	Approved central library exists; 9 of 100 manuals sampled during discovery lack a recorded owner or review date.
Contracts	Access is role-based, but labels and effective dates differ across systems.
Prior work orders	Useful for language and troubleshooting context; retention and cross-customer reuse require review.
Vendor service	Hosted model and retrieval platform; regional processing available; contract bars training on client content but change notice is only 14 days.
Vendor evaluation	High aggregate extraction score on 100 vendor-created cases; no Harborline safety or contract slice and no independent grader report.
Tools	Prototype exposes request read, source search, draft save, priority update, assignment, and outbound message tools.
Operations	Vendor provides uptime target; Harborline must own source freshness, user access, evaluation, monitoring, incident triage, and state reconciliation.

CAUTION: The prototype boundary is too wide

The vendor tool list includes priority, assignment, and outbound messages even though the proposed pilot is assistive. The runtime permission boundary, not a prompt instruction, must exclude those actions before pilot entry.

5. Proposed six-week pilot

DESIGN	PROPOSED CHOICE
Population	Two queues, weekdays, named coordinators, eligible intake and work-order tasks.
Intervention	Structure request, retrieve approved sources, list missing information, draft work order.
Comparison	Concurrent eligible requests handled with the current workflow, assigned by shift and queue where practical.
Exposure	Week 1 sandbox; week 2 shadow; weeks 3-6 human-approved drafts if entry gates pass.
Excluded	Priority, technician assignment, customer commitments, outbound messages, autonomous writes.
Primary workflow measures	Active preparation time, elapsed time, complete-first-review, rework, service-level attainment.
Protected outcomes	Safety and contract errors, cross-customer access, unsupported claims, review burden, incidents.
Stop rules	Unauthorized data/action, invented safety instruction, hidden write, repeated stale-source failure, uncontained incident.

6. Frozen candidate results

All numbers in this packet are invented for practice. Weeks 1-3 used HSP-0.7 for sandbox, shadow, and early assisted work. Its 29 incident-exposed drafts are reported separately below. After a fresh gate, HSP-0.8 froze the model, prompts, source index, freshness checks, permissions, and review interface for weeks 4-6. That cohort contains 612 eligible requests: 318 completed with assistance and 294 in the current workflow. Twelve coordinators participated. Assistance was available on 442 requests; 318 used it and 124 chose fallback. Those 124 are included in the 294 current-workflow requests alongside 170 requests without assistant access. This comparison therefore includes self-selection as well as nonrandom shifts.

TRACK	RESULT	LIMIT
Preparation time	Median 8.6 minutes assisted vs 12.1 comparison.	Queues were concurrent, but shift assignment was not randomized.
Elapsed to reviewable	Median 35 vs 44 minutes.	Customer waiting remains highly variable.
Complete first review	242/318 = 76.1% vs 185/294 = 62.9%.	Improvement smaller on contract-exception slice.
All required fields correct	298/318 submitted HSP-0.8 drafts = 93.7%.	Every assisted draft was audited; this is a draft-level measure, not field accuracy.
Citation validity	546/565 cited passages supported the field = 96.6%.	Contract: 89/100; routine manual: 392/400; other: 65/65. Passages cluster within drafts.
Unsupported safety claim	0 in 74 safety challenge cases after fix.	Small challenge set; does not establish zero production risk.
Cross-customer access	0 observed; exact authorization suite 180/180 passed.	Only configured pilot identities and sources were tested.
Review time	Median 2.9 minutes; 57/318 = 17.9% required substantive edits.	Review burden is included in preparation measure.
Eligible use	318/442 = 71.9%; final two weeks 168/200 = 84%.	The final 200 available requests are a subset of the 442. Fallback remains permitted.
Platform cost	\$8.10-\$13.60 platform cost per 100 available eligible units.	Add internal ownership, source cleanup, and any enterprise tier using the worked model below.

Declared gates and missing outcomes

PREDECLARED EXPANSION GATE	HSP-0.8 EVIDENCE	DECISION
All-fields-correct drafts at least 93%	298/318 = 93.7%	Point estimate passes; review uncertainty.
Valid citations at least 95% overall AND in contract slice	546/565 = 96.6%; contract 89/100 = 89%	Contract floor fails. Hold expansion.
No unauthorized data/action or invented safety instruction	No event observed in 318 submitted units; 180/180 authorization tests; 0/74 safety challenges	Supports tested controls, not absence of rare risk.
Rework no worse; service and safety/contract outcomes measured	Rework 76/318 vs 109/294; service-level attainment and downstream contract-error audit incomplete	Outcome evidence incomplete. Hold expansion.
Owners ready; full cost acceptable	Deletion evidence and provider-change drill open; first-year expected net capacity value negative	Hold expansion; decide whether bounded repair is worth funding.

A passing point estimate is not a guarantee. The 74 safety challenges are selected stress cases, not a production prevalence sample. Keep the zero count and its selection limits together.

7. Mid-pilot incident and response

A stale compressor manual passage produced a wrong urgency cue in one draft. The coordinator caught it before submission. Investigation found that the document owner replaced the PDF without updating the approved index or review date. The team paused that equipment class, searched affected units, corrected the index, added owner and freshness gates, displayed source status in review, created regressions, and required Safety, Data, Product, and Risk approval before resuming. No submitted work order contained the wrong cue.

EFFECT	EVIDENCE
Containment	Affected class paused in 18 minutes; all 29 HSP-0.7 drafts reviewed; 28 had correct required fields, one had the caught wrong urgency cue, and none committed it.
Eligible use	Early HSP-0.7 use fell from 71% to 49% after the incident. The later HSP-0.8 cohort ended at $168/200 = 84\%$; different periods and denominators prevent treating this as a controlled effect of the fix.
Control	Source owner and review date now required; stale status blocks draft fields that depend on the source.
Residual risk	Other manual classes were sampled, not exhaustively verified; source governance remains an operating dependency.

8. Ownership and capability

AREA	CURRENT STATE
Service Operations	Owens review policy; current support uses 16 hours/month. Expansion would require 0.3 FTE, assumed 48 hours/month.
Product	Named owner accepts behavior contract, evidence board, version and change process.
IT/Engineering	Can operate integration and rollback; provider-change rehearsal remains incomplete.
Data	Owners assigned for pilot sources; enterprise source inventory is incomplete.
Safety/Risk	Accepts only current two-queue exposure during approved repair; expansion awaits contract and outcome evidence.
Vendor	Will add 30-day model-change notice at enterprise tier; export test passed; deletion evidence pending.
Economics	Expected first-year net capacity value is negative in the worked model; wider support requirements worsen it.

9. Reproduce the monthly economics

For practice, assume a month has 912 available eligible requests in the two queues (2,400 x 38%). Use 72% as the planning use rate, not as a replacement for 318/442 in the cohort report. Observed HSP-0.8 group mean preparation times are 8.8 and 11.8 minutes, including review, a 3.0-minute association. Nonrandom allocation prevents a causal savings claim. The median gap of 3.5 minutes is not used to calculate total capacity.

Loaded labor value is an assumed \$45/hour. Current supervision is 16 hours/month plus 8 hours for data, product, monitoring, and risk, at the same rate: \$1,080/month. Source cleanup costs \$6,000 once; spreading it over the first twelve months adds \$500/month. Expected platform cost uses \$10.85 per 100 available requests. No staffing reduction, added billable work, or cash saving has been demonstrated.

SCENARIO	AVAILABLE; USE; MINUTES	CAPACITY VALUE	NET/MONTH, YEAR 1
Low	400; 50%; 2	\$300.00	\$300 - \$54.40 - \$1,080 - \$500 = -\$1,334.40
Expected	912; 72%; 3	\$1,477.44	\$1,477.44 - \$98.95 - \$1,080 - \$500 = -\$201.51
High	1,400; 84%; 4	\$3,528.00	\$3,528 - \$113.40 - \$1,080 - \$500 = \$1,834.60

Capacity value = available requests x use rate x mean net minutes / 60 x \$45. Scenario assumptions are not additional observed pilot results. With one comparable 456-request queue added, expected capacity value is \$2,216.16/month, but 48 supervisor hours plus 8 ownership hours cost \$2,520 before platform, cleanup, or enterprise support. Wider volume does not automatically solve the economics. Ask Finance to value service improvements separately and avoid double-counting them as time and revenue.

Capstone tasks

- 1 Rewrite the executive request as the decision the packet can support.
- 2 Choose stop, revise, extend, or scale and name the exact next boundary.
- 3 Identify the decisive evidence, hard blockers, measurement limits, and residual risks.
- 4 Explain why the recommendation does or does not include priority, assignment, customer communication, or autonomous writes.
- 5 Write the next gate: prerequisites, owners, evidence, stop rules, and date.
- 6 Create a handoff plan covering Operations, Product, IT, Data, Safety/Risk, Vendor, and executive ownership.
- 7 State the supported claim and at least five non-claims.

Guided capstone decision worksheet

GUIDED CAPSTONE DECISION WORKSHEET
Decision and exact permitted workflow, population, data, actions, and exposure
Decisive evidence across service, system, risk, people, economics, and ownership
Hard blockers, measurement limitations, uncertainty, and residual risk
Stop, revise, extend, or scale recommendation and rationale

GUIDED CAPSTONE DECISION WORKSHEET

Prerequisites, owners, fresh evidence, stop rules, and next review date

Handoff and operating cadence

What this decision does not establish

Guided capstone answer key

The packet supports more than one carefully bounded implementation detail, but it does not support enterprise automation. A strong answer preserves the distinction between an assistive workflow and dispatch decisions or actions.

1. Reframe the decision

ANSWER: Decision the evidence can support

Whether Harborline should fund one bounded repair and fresh evidence period in the current two queues before considering any expansion of HSP-0.8. Priority, assignment, customer commitments, outbound messages, and autonomous writes stay disabled.

2. Read the evidence by track

- Workflow: preparation, elapsed time, and complete-first-review improved in the two pilot queues. Concurrent comparison helps, but nonrandom shifts and seasonal variation limit causal certainty.
- System: the draft-level point estimate passes, but contract citations are 89/100 against a 95% floor. Repair retrieval and review, then test fresh cases. Service-level and downstream contract-error evidence is missing.
- Risk: the reduced runtime tool set and exact authorization suite support the assistive boundary. They do not support any excluded action. The stale-source incident shows that knowledge governance is a material operating control.
- People: later-candidate use was higher, but different periods and denominators prevent attributing that change to the repairs alone. Support capacity and review burden belong in the operating model.
- Economics: the expected first-year capacity-value case is -\$201.51/month before an enterprise tier. It is not cash savings. Expansion needs more supervisor support. A positive high case does not close the funding decision.
- Ownership: acceptance is limited to current two-queue exposure during approved repair. Deletion evidence and a provider-change drill remain open. No owner accepts automated priority or assignment.

3. Make the bounded recommendation

DECISION: Expected recommendation

Revise within the existing boundary. Hold expansion while the contract floor, outcome evidence, vendor obligations, provider-change drill, and funding decision remain open. The COO may fund a four-week repair and fresh evaluation period. If the repair budget is refused or a hard blocker cannot be contained, stop the affected exposure. Extend is defensible only after naming the uncertainty and confirming that additional observation will answer it; scale is not supported by this packet.

Within ten working days of this review, Product brings the repair plan, Finance brings the cost sensitivity, and Risk brings the open controls to the COO. No reply means no new exposure. If approved, run four weeks on one frozen repaired candidate in the same queues, review weekly, and decide again on the next working day after week four. Contract citations must pass the declared floor on fresh cases; missing outcomes must be measured. Safety, Privacy, Security, and Operations retain immediate stop authority.

4. Keep excluded actions blocked

ACTION	WHY THE PACKET DOES NOT SUPPORT IT
Priority recommendation	No reliable baseline or agreed ground truth; safety consequence and policy variation remain material.
Technician assignment	Certification, geography, labor rules, availability, authorization, state, and recovery were not evaluated.
Customer communication	Commitment, consent, contract, tone, channel, approval, and reconciliation were outside the pilot.
Autonomous write	The successful boundary depends on the action being unavailable and on coordinator approval through the existing system.
Enterprise rollout	Other regions, sources, users, volume, season, support, contracts, and governance were not represented.

This gate repairs the existing boundary. It does not admit an additional queue.

5. Define the next gate

- 1

 Complete vendor deletion evidence and provider-change rehearsal.
- 2

 Assign and audit source owners, review dates, access, and freshness for the repaired candidate in the current two queues.
- 3

 Protect supervisor support capacity and train users on sources, review, override, fallback, and incident reporting.
- 4

 Freeze the candidate and predeclare larger contract and safety slices plus service, review, adoption, cost, and control measures.
- 5

 Retain the same hard stop conditions, response path, reduced tool set, and manual workflow.
- 6

 Require Service Operations, Product, IT, Data, Safety/Risk, Vendor owner, and executive owner to accept only the current two-queue repair boundary. Review a separate expansion proposal after the repair gate.

Handoff plan for the current boundary

OWNER	ARTIFACT AND ACCEPTANCE DEMONSTRATION	DUE
Operations	Review policy, fallback, incident runbook; supervisor runs a bad-draft and outage drill.	Before repair exposure
Product	Candidate manifest, cases, decisions; leads weekly evidence review and one controlled change.	Before final review
IT	Runtime, permissions, rollback; rehearses provider change and restores prior version.	Before repair exposure
Data	Source inventory and freshness rules; updates a manual and verifies index, access, and denial.	Before repair exposure
Safety/Risk	Residual risks and stop/restart record; reviews safety and contract evidence and witnesses incident drill.	At every exposure gate
Vendor	Change, export, retention, deletion terms; supplies deletion evidence and tests exit.	Before repair exposure
COO / Finance	Full cost, service value, funding, claim limits; accepts only the demonstrated boundary.	Ten-day decision and week-four review

Meet weekly during repair. After accepted handoff, use the Chapter 12 continuous, weekly, monthly, and event-triggered cadence. Keep consultant access and support obligations explicit until each owner demonstrates the transferred task. Missing demonstrations keep handoff open even when the files are delivered.

6. State the non-claims

- [] The pilot does not establish that AI can automate dispatch.
- [] The pilot does not establish zero safety, privacy, access, or unsupported-claim risk.
- [] The pilot does not establish the same effect in other regions, seasons, equipment classes, contracts, or user groups.
- [] The pilot does not establish causal productivity improvement or cash savings.
- [] The pilot does not establish readiness for priority, assignment, communication, or autonomous writes.
- [] The pilot does not establish legal compliance for every jurisdiction or use.
- [] The pilot does not establish that future model, vendor, source, policy, workflow, or staffing changes remain covered by the evidence.

CLAIM: Correct proof boundary

The packet supports a bounded repair decision for a human-approved intake, evidence-retrieval, and drafting service in its existing two queues, if owners fund and authorize that repair. It does not support expansion, autonomous dispatch, booked savings, or a completed handoff.

Guided capstone self-assessment

Rate each category. Competence requires a defensible boundary in every category; a strong average cannot compensate for a missing safety, permission, evidence, or ownership decision.

LEVEL	DESCRIPTION
Absent	The decision, evidence, owner, control, or claim boundary is missing.
Developing	The right topic appears, but important distinctions, evidence, controls, or owners are vague or wrong.
Competent	The recommendation uses the correct evidence, preserves boundaries, and names practical owners and next gates.
Strong	The answer also challenges assumptions, anticipates failure and operating change, and designs proportionate disconfirming evidence.

CATEGORY	COMPETENT EVIDENCE
Decision frame	Names the owner, choices, workflow, population, date, evidence, and non-goals.
Current work	Uses observed facts, baseline, exceptions, affected people, and limitations.
Opportunity	Compares alternatives and classifies assist, recommend, decide, or act.
Boundary	Names permitted data, outputs, actions, exposure, human authority, and fallback.
Governance	Connects risk to controls, permissions, legal review, vendors, decision rights, and stop/restart.
Pilot	Uses comparison, separate evidence tracks, gates, slices, stop rules, and version control.
Delivery	Coordinates product, data, engineering, evaluation, governance, adoption, commercial, and operations.
Evidence	Interprets workflow, system, risk, people, economics, and ownership without averaging blockers away.
Decision	Chooses stop, revise, extend, or one supported scale boundary with prerequisites.
Handoff	Transfers accepted artifacts, owners, runbooks, cadence, capability, vendor and retirement paths.

Automatic non-passing misconceptions

- [1] Treating executive sponsorship or a workshop score as proof of value.
- [1] Recommending autonomous dispatch without safety, authorization, state, recovery, and ownership evidence.
- [1] Claiming ROI without a baseline, comparison, full operating cost, sensitivity, and conversion mechanism.
- [1] Averaging away a High-severity failure or using adoption to offset a control gap.
- [1] Calling low use resistance without examining latency, review burden, workflow fit, incentives, and trust.
- [1] Calling handoff complete without named owners, acceptance demonstrations, monitoring, rollback, vendor, and retirement paths.

Appendix A: consulting lifecycle

Use this as the front sheet for an engagement or review packet.

- [] **DECISION** - A named owner, choices, evidence standard, date, scope, non-goals, and integrity disclosures are visible.
- [] **WORK** - The unit, actors, evidence, decisions, tools, queues, exceptions, baseline, and affected people are observed.
- [] **OPPORTUNITY** - Process, rules, information, AI, and defer options are compared; unknown is not scored as neutral.
- [] **BOUNDARY** - Users, data, outputs, actions, exposure, time, human authority, fallback, and non-claims are explicit.
- [] **GOVERNANCE** - Risk chains, controls, permissions, legal review, vendor evidence, decision rights, stop and restart are owned.
- [] **PILOT** - Hypothesis, population, comparison, separate evidence tracks, gates, hard blockers, exposure, and versions are predeclared.
- [] **DELIVERY** - Product, data, engineering, evaluation, governance, adoption, operations, and commercial decisions are integrated.
- [] **READINESS** - The exact complete candidate has representative, challenge, control, recovery, usability, cost, and owner evidence.
- [] **PEOPLE** - Role and workload change, meaningful review, literacy, training, support, feedback, and non-use reasons are addressed.
- [] **OPERATE** - Weekly evidence, slices, issues, incidents, changes, state reconciliation, communication, and stop rules are active.
- [] **DECIDE** - Service, system, risk, people, economics, and ownership are interpreted separately; the next boundary is exact.
- [] **TRANSFER** - Accepted artifacts, owners, runbooks, cadence, capability, vendor, retirement, and consultant closeout are verified.
- [] **CLAIM** - The final sentence states what the evidence supports, the uncertainty, and what it does not establish.

DECISION: Smallest supported next step

When evidence is mixed, do not round up to the hoped-for transformation. Recommend the smallest workflow, population, data, action, and operating boundary actually supported, then name the evidence required for expansion.

Appendix B: discovery interview and observation guide

Sponsor and decision owner

- What decision must be made, by whom, and by what date?
- Which outcomes matter, how are they measured now, and what tradeoff is unacceptable?
- What answer do you currently expect, and what evidence could change your mind?
- Which people, rights, safety, service, data, contracts, and jurisdictions can be affected?
- What is explicitly outside this engagement, and who may change that boundary?

People who perform and receive the work

- Walk me through the last real unit from arrival to outcome. Where did you wait, switch, check, ask, or recover?
- Which cases are easy, ambiguous, rare, high consequence, or impossible?
- What evidence do you trust, where does it come from, and how do you know it is current?
- What would a proposed assistant need to show so you could review it responsibly?
- Which errors create rework, embarrassment, harm, rights impact, or hidden downstream cost?
- What would make you avoid the system even if leaders wanted it used?

Observation prompts

- [1] Start and end state; unit of work; eligibility and slice.
- [1] Actors, evidence, decisions, tools, copies, searches, handoffs, queues, waits, interruptions.
- [1] Normal path, exceptions, workarounds, escalation, fallback, recovery, and downstream correction.
- [1] Active time, elapsed time, review, rework, error, outcome, user and employee burden.
- [1] What the systems record, what they omit, and which personal artifacts fill the gap.
- [1] Fact, participant report, interpretation, proposal, and unresolved question recorded separately.

Appendix C: assist, recommend, decide, and act

PATTERN	GOOD FIRST USES	REQUIRED QUESTIONS	EXPANSION GATE
Assist	Drafting, transformation, retrieval, summarization, classification support.	Can a person verify? Are sources and uncertainty visible? Is fallback easy?	Representative quality, review burden, safe data, support and workflow value.
Recommend	Ranking, triage, suggestions, options, risk flags.	Is the decision policy agreed? Can the person challenge? Are slices and overrides understood?	Validated decision support, meaningful oversight, protected outcomes, audit and appeal.
Decide	Bounded low-consequence decisions with observable rules and appeal.	What rights, safety, money, access or service changes? Who is accountable?	Legal and risk review, strong evidence, control architecture, challenge and remediation.
Act	Reversible, narrow, authorized actions with exact confirmation and recovery.	Which resource, permission, side effect, external state, duplicate and rollback path?	Server authorization, idempotency, approval, state checks, incident, reconciliation and owner acceptance.

CAUTION: Do not confuse interface language with authority

A label such as copilot, assistant, recommendation, or human in the loop does not determine the real role. Inspect what data the system reads, what it changes, what the person can see and control, and which outcome follows in practice.

Appendix D: measurement and economics cheat sheet

MEASURE	PLAIN-LANGUAGE CALCULATION	COMMON MISTAKE
Eligibility	eligible units / all arriving units	Reporting usage against cases the system cannot handle.
Eligible use	used eligible units / eligible units available	Calling non-use resistance without investigating conditions.
First-pass acceptance	units accepted without rework / completed units	Ignoring downstream correction or weak review.
Active time	time spent doing the task and review	Calling reduced model or transaction time a workflow gain.
Elapsed time	end timestamp minus start timestamp	Ignoring queues, waits, scheduling, and customer response.
System pass rate	passing evaluated units / evaluated units	Hiding set purpose, slices, grader, uncertainty, and version.
Override/edit	overridden or materially edited units / reviewed units	Treating every edit as a model failure or every acceptance as correctness.
Incident rate	incidents / defined exposure unit	Comparing unlike severity, detection, population, or time windows.
Gross capacity	available eligible units x use rate x mean net minutes / 60	If volume already counts uses, do not multiply by use rate again. A median gap cannot total capacity; cash needs a conversion plan.
Full recurring cost	system + data + human review + support + governance + monitoring + incidents	Using provider token cost as total cost.
Net value range	credible benefit range minus full cost range	Using one optimistic point estimate and excluding change or exit.

CAUTION: No measure interprets itself

Attach the unit, population, period, comparison, version, sample, missing data, uncertainty, important slices, measurement method, protected outcome, and decision boundary.

Appendix E: governance and vendor question bank

AREA	QUESTIONS
System role	What is the intended and prohibited use? Who is provider, deployer, operator, affected person, and decision owner?
Data	What enters, leaves, persists, trains, logs, crosses regions, reaches subprocessors, or supports rights requests?
Models	Which service and version run? How are changes announced, constrained, tested, rolled back, and audited?
Security	How are identity, authorization, isolation, secrets, source-to-sink policy, vulnerabilities, incidents, and supply chain handled?
Evidence	Which representative, challenge, slice, security, human, reliability, cost, and monitoring evidence exists? Who produced and graded it?
Human oversight	What can the person see, decide, override, appeal, stop, and recover under real workload?
Operations	Who monitors, supports, triages, changes, reconciles, communicates, and retires the system?
Commercial	What are full price, minimums, limits, SLAs, support, audit rights, liability, renewals, portability, deletion, and exit?
Regulatory	Which jurisdictions, sectors, classifications, documentation, transparency, literacy, impact, and incident duties require qualified review?

This question bank supports diligence and issue spotting. It is not a complete security, privacy, procurement, employment, accessibility, sector, or legal review. Tailor it to the actual use and accountable reviewers.

Appendix F:

plain-language glossary

TERM	MEANING IN THIS PLAYBOOK
Slice	A meaningful subset of cases, such as contract exceptions or safety-related requests. Report its own count and result.
Grader	The person or rule that evaluates a result. Check the grader against a defensible human reference before using its score.
Ground truth	A reference answer established by a stated process, not an infallible label. Preserve expert disagreement.
Lineage	The trace from a draft field back to the request, source, version, and transformation that produced it.
Idempotency	Repeating the same authorized operation does not repeat its side effect. A retry must not create a second work order.
Reconciliation	Check the authoritative system after an uncertain operation and repair any mismatch before acting again.
Source-to-sink policy	Rules for where information may come from and where it may go. A customer's contract cannot enter another customer's draft.
AI system	Model plus application, instructions, data, retrieval, tools, interface, permissions, people, policies, monitoring, and external state.
AI boundary	The users, workflow, data, outputs, actions, exposure, time, and exclusions that define permitted use.
Behavior contract	Observable intended behavior, hard failures, fallback, and claims for the named system and use.
Decision owner	The person accountable for accepting an action and its residual risk; contributors may supply evidence without owning the decision.
Evidence ledger	Traceable record of claims, sources, observations, tests, results, limitations, versions, and decisions.
Exposure	Who and what can be affected: users, units, data, actions, volume, time, location, and consequence.
Hard blocker	A predeclared condition that prevents entry or expansion regardless of aggregate performance.
Human oversight	A designed ability to understand, challenge, decide, stop, and recover; not merely the presence of a person.

TERM	MEANING IN THIS PLAYBOOK
Pilot	A bounded test of a system and operating model under conditions chosen to support a named decision.
Representative evidence	Evidence sampled to estimate behavior for a named deployment population under stated assumptions.
Challenge evidence	Risk-enriched cases intended to expose boundaries, rare failures, attacks, outages, and severe conditions.
Residual risk	Risk remaining after treatment, considered by an accountable owner within a named boundary.
System manifest	Versioned identity of the workflow, application, model, data, evaluation, controls, people, and vendors under test.
Meaningful human control	Authority, information, time, competence, record, and fallback sufficient for real responsibility.
Adoption	Observable use and completion behavior within eligible work, interpreted with fit, burden, incentives, trust, and support.
Drift	Change in users, work, data, policy, model, vendor, environment, measurement, or outcomes that can weaken prior evidence.
Scale	A specific expansion of population, volume, data, location, action, autonomy, or duration, each requiring matched evidence.
Handoff	Accepted transfer of artifacts, access, decision rights, runbooks, capability, vendor management, cadence, and accountability.
Claim boundary	A statement of what the evidence supports and what it does not establish.

Appendix G: worksheet index

Print the worksheets or copy their fields into the client's governed document system.

WORKSHEET	MODULE	PURPOSE
Engagement decision brief	1	Name the decision, owner, scope, evidence, non-goals, date, and integrity boundary.
Current-state evidence	2	Map the real work, baseline, exceptions, evidence, facts, and unresolved questions.
Opportunity recommendation	3	Compare alternatives and select one bounded change worth testing.
Target workflow and boundary	4	Define future flow, behavior contract, permissions, exclusions, fallback, and value hypothesis.
Risk, permissions, governance	5	Connect affected people and failures to controls, ownership, legal review, and vendor evidence.
Pilot charter	6	Predeclare hypothesis, population, comparison, measures, gates, exposure, and next action.
Integrated delivery plan	7	Coordinate workstreams, gates, decisions, assumptions, dependencies, and definition of done.
Readiness evidence packet	8	Identify the complete candidate and collect behavior, control, recovery, and acceptance evidence.
People and adoption plan	9	Design role change, literacy, review, training, support, feedback, and adoption measures.
Pilot evidence board	10	Operate a versioned weekly learning, issue, incident, change, and exposure decision loop.
Outcome and decision record	11	Synthesize separate tracks into stop, revise, extend, or bounded scale.
Handoff and closeout	12	Transfer accepted ownership, artifacts, runbooks, cadence, capability, vendor, and exit.
Capstone decision	Capstone	Make and defend the complete Harborline recommendation and proof boundary.

Further reading and source notes

The playbook favors governing texts, standards bodies, regulators, official public-sector guidance, and direct technical documentation.

Source status matters: law changes; living pages and vendor documentation can change; NIST frameworks are voluntary; OECD recommendations are non-binding; ISO summaries do not replace the paid normative text; a useful procurement pattern can outlast its legal references. Refresh fast-moving sources before publication or a client decision.

SOURCE GROUP	HOW IT IS USED	STATUS CAUTION
C1-C3	Risk-management backbone and operational actions.	Voluntary frameworks; select actions for the actual use and evidence.
C4-C5	Management system and trustworthy-AI principles.	ISO full requirements are paid; OECD recommendation is non-binding.
C6-C11	EU AI Act, 2026 amendment, implementation, literacy, privacy.	C8 supports the attributed timeline. C6-C7 legal bodies and C9 FAQ were unavailable to this check; review current law.
C12-C15, C21	Cyber, secure development, AI threats, supplier diligence.	Adapt to sector, architecture, provider and current threat model.
C16-C19, C24-C26	Problem framing, impact, accountability, adoption, literacy, worker considerations.	Public-sector context needs localization; adoption data may lag. C24 and C26 bodies were unavailable.
C20, C22	Contract and procurement diligence patterns.	Older or pre-2026 legal references require updating.
C23	Evaluation design principles.	Living platform documentation; product details and deprecation timelines can change.
C27-C28	Responsible AI diligence and accessible web workflows.	Voluntary diligence and testable web criteria; neither establishes every legal duty.

[C1] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 26, 2023. Voluntary framework. <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>

[C2] NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, July 26, 2024. <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>

[C3] NIST. AI RMF Playbook, complete version 2023. Suggested actions rather than a mandatory checklist. <https://airc.nist.gov/airmf-resources/playbook/>

[C4] ISO. ISO/IEC 42001:2023 AI management systems, free official summary. <https://www.iso.org/standard/42001>

[C5] OECD. Recommendation of the Council on Artificial Intelligence, adopted 2019 and amended 2024. <https://www.oecd.org/en/topics/ai-principles.html>

[C6] European Union. Regulation (EU) 2024/1689, the Artificial Intelligence Act. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=en>

[C7] European Union. AI Omnibus Regulation, July 2026. Consult the final legal text alongside the Commission timeline in C8. <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32026R1744>

[C8] European Commission. AI Act implementation timeline, accessed October 1, 2026. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

[C9] European Commission. AI literacy questions and answers. Further reading; page body unavailable at the October 1, 2026 check. <https://digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-answers>

[C10] European Union. Regulation (EU) 2016/679, General Data Protection Regulation. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

[C11] European Data Protection Board. Opinion 28/2024 on AI models and personal data. https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en

[C12] NIST. Cybersecurity Framework 2.0, 2024. <https://www.nist.gov/publications/nist-cybersecurity-framework-csf-20>

[C13] NIST SP 800-218A. Secure Software Development Practices for Generative AI, 2024. <https://csrc.nist.gov/pubs/sp/800/218/a/final>

- [C14] CISA and UK NCSC. Guidelines for Secure AI System Development, 2023. <https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development>
- [C15] OWASP. Top 10 for LLM Applications 2025, released November 2024. <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
- [C16] UK Government. Artificial Intelligence Playbook for the UK Government, 2025. <https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html>
- [C17] UK Government. Guidance for evaluating the impact of AI tools, 2025. <https://www.gov.uk/government/news/new-guidance-for-evaluating-the-impact-of-ai-tools>
- [C18] U.S. Government Accountability Office. AI Accountability Framework, 2021. <https://www.gao.gov/products/gao-21-519sp>
- [C19] OECD. The Adoption of Artificial Intelligence in Firms, 2025. https://www.oecd.org/en/publications/the-adoption-of-artificial-intelligence-in-firms_f9ef33c3-en/full-report.html
- [C20] European Commission. Updated EU AI Model Contractual Clauses, March 5, 2025. Adapt these examples to the current law and contract; they are not a compliance certificate. <https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/updated-eu-ai-model-contractual-clauses>
- [C21] NIST SP 1326. Cybersecurity Supply Chain Risk Management: Due Diligence Assessment Quick-Start Guide, July 2026. <https://csrc.nist.gov/pubs/sp/1326/final>
- [C22] UK Government. Guidelines for AI Procurement, 2020. <https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement>
- [C23] OpenAI. Evaluation best practices, living documentation accessed October 1, 2026. <https://developers.openai.com/api/docs/guides/evaluation-best-practices>
- [C24] U.S. Department of Labor. Artificial Intelligence Literacy Framework, 2026. Further reading; source body unavailable at this edition check. <https://www.dol.gov/agencies/eta/advisories/ten-07-25>

[C25] Government of Canada. Algorithmic Impact Assessment tool, current page accessed October 1, 2026. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/automated-decision-making/algorithmic-impact-assessment.html>

[C26] U.S. Department of Labor. Artificial Intelligence and Worker Well-being: Principles for Developers and Employers, 2024. Further reading; source body unavailable at this edition check. <https://www.dol.gov/newsroom/releases/osec/osec20240516>

[C27] OECD. OECD Due Diligence Guidance for Responsible AI, February 19, 2026. https://www.oecd.org/en/publications/oecd-due-diligence-guidance-for-responsible-ai_41671712-en.html

[C28] W3C. Web Content Accessibility Guidelines (WCAG) 2.2. Testable web criteria; evaluate the actual workflow. <https://www.w3.org/TR/WCAG22/>

DECISION: You are finished when ownership and the claim are bounded

A strong AI consulting engagement does not end with a roadmap, a demonstration, or a maturity score. It ends with a named decision, a supported boundary, credible evidence, unresolved risk, accountable owners, an operating cadence, and a client able to continue without the consultant.

Changelog

Version 1.0.1. Compared with edition 1.0.0, published October 1, 2026.

Previous PDF SHA-256: 4a7055b53ec4e4db7dbafacb6895513b055b4adeb02354c9f38fe53999ed92b2

Correction or clarification

Add a fictional support-pilot example before the blank decision summary and twelve numbered questions for completing it.

Show what a completed summary contains and let readers work through each decision before filling the template.

Author synthesis; all example figures are fictional. Context: AI Consulting Playbook.