



# AI Implementation Guidelines

Standards for taking an AI use case from idea to a monitored production system that people trust and use.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 02 / EMBER & CORAL

EXECUTIVE WORKING EDITION - 08 OCT 2026

# 02

## AT A GLANCE

Standards for taking an AI use case from idea to a monitored production system that people trust and use.

Most AI projects fail for ordinary reasons: unclear problems, weak data, no owner, no test and no plan for the people who must change how they work. These guidelines set the standards a team follows when it selects, builds or buys, tests, launches and runs an AI system. They include stage gates, a use-case intake form, readiness and launch checklists, vendor questions, evaluation rules, human oversight design, change management and monitoring. They are written to be adapted and adopted, and they are general guidance, not legal advice.

**WHO IT IS FOR** Product, operations and technology leaders, project managers, data and engineering leads, and business owners who sponsor or deliver AI projects inside a mid-sized or large organization.

**TIME TO USE** 70 minutes to read; one workshop to adapt the stage gates to your delivery process

## What you will have at the end

- A five-stage delivery path with an exit test at each gate
- A use-case intake form and selection rules that filter out weak ideas early
- Build-or-buy and vendor due diligence standards
- Evaluation, human oversight and launch readiness requirements you can audit
- A monitoring, change and retirement standard so systems do not drift unnoticed

## Contents

|           |                                                       |    |
|-----------|-------------------------------------------------------|----|
| <b>01</b> | What these guidelines standardize and when they apply | 03 |
| <b>02</b> | Seven delivery principles                             | 04 |
| <b>03</b> | Choosing what to build, buy or leave alone            | 05 |
| <b>04</b> | Five stages with an exit test at each gate            | 07 |
| <b>05</b> | Standards for buying AI and for preparing data        | 08 |
| <b>06</b> | Testing, human oversight and launch readiness         | 10 |
| <b>07</b> | Adoption, training and role changes                   | 11 |
| <b>08</b> | Monitoring, change control, incidents and roadmap     | 13 |
| <b>09</b> | A first quarter that produces evidence                | 14 |
| <b>10</b> | Terms used in this publication                        | 15 |

## PURPOSE AND SCOPE

# What these guidelines standardize and when they apply

01



The aim is repeatable delivery: every AI system goes through the same questions, scaled to its risk.

AI projects often start with enthusiasm and a demo. They stall when someone asks who owns the data, how accuracy will be measured, what happens when the model is wrong or whether the people expected to use it were consulted. These guidelines put those questions up front. They give delivery teams a common path so that a project in finance and one in customer support are judged by the same standards, with depth set by risk rather than by who shouts loudest.

They apply to any AI system your organization builds, configures, buys or embeds in a process: predictive models, generative assistants, document processing, search over internal content, automation and agents. They also apply to AI features that a vendor switches on inside existing software, because those change risk without a project ever being opened. Day-to-day staff use of general assistants is covered by the responsible use guidelines.

## SCOPE NOTE

This document is an operating template, not legal advice. Requirements on privacy, employment, consumer protection, sector regulation and AI itself, including the EU AI Act where it applies, depend on your jurisdiction and use case. Have counsel confirm current obligations before launch.



## WORKING QUESTION

Can your organization list every AI system in production today, with one named owner for each?



## PRINCIPLES

# Seven delivery principles

02

These principles settle trade-offs when schedules are tight and the demo looks better than the evidence.

| PRINCIPLE                     | MEANING                                                                 | WHAT IT PREVENTS                                                 |
|-------------------------------|-------------------------------------------------------------------------|------------------------------------------------------------------|
| Start with the problem        | A measurable business problem comes before the choice of model or tool. | Solutions in search of a use                                     |
| One accountable owner         | Each system has a business owner, not only a project team.              | Orphaned systems                                                 |
| Evidence before scale         | Pass a pre-agreed test on realistic data before expanding access.       | Launches based on demos                                          |
| Risk sets rigor               | Depth of review rises with the hazard of an error.                      | Shallow reviews for trivial uses, light reviews for serious ones |
| People before platform        | Users are involved early and trained before launch.                     | Tools nobody adopts                                              |
| Humans stay answerable        | Oversight is designed in, with real authority and time to use it.       | Rubber-stamp review                                              |
| Run it, do not just launch it | Monitoring, support and a retirement plan are part of the project.      | Silent degradation                                               |

Frameworks such as the NIST AI Risk Management Framework, which organizes work into govern, map, measure and manage, and ISO/IEC 42001, which describes an AI management system, support these principles. You do not need to certify against a standard to borrow its structure. The point is to have a consistent method, documented decisions and evidence that the method was followed.

## Risk levels used in this document

- Low: internal use, reversible, no personal or confidential data, a person always reads the output.

- Medium: customer-facing or business-relevant output, confidential data, or moderate consequences of error.
- High: affects an individual's rights, finances, employment, health or safety, or is subject to specific regulation.
- If unsure, classify one level higher and let the reviewer lower it with a recorded reason.



WORKING QUESTION

Which of these principles does your last AI project show the least evidence of?

SELECTING USE CASES

## Choosing what to build, buy or leave alone

03



The cheapest failure is the project you decline in week one, so make selection explicit and unglamorous.

Every proposal enters through one intake form. The form is short enough to complete in thirty minutes and specific enough to reveal a weak case. Review requests on a fixed rhythm, such as weekly, and publish the outcome with reasons. Over time the decisions teach the organization what good proposals look like and reduce the load on reviewers.



### 01 Problem and owner

State the problem in one sentence, who experiences it, the current process, the named business owner and the sponsor.

### 02 Value and measure

Name the metric that should move, its current baseline, the target and how it will be measured. If no baseline exists, the first task is to establish one.



### 03 Data and access

List the data needed, where it lives, who owns it, its sensitivity tier, its quality and whether you have the right to use it for this purpose.



## 04 People and consequence

List the users and anyone affected by outputs. Describe what happens if the system is wrong and the proposed human check.

## 05 Approach and effort

Choose among no AI, an off-the-shelf feature, a configured product or a custom build, with a rough effort and a reason.

| SELECTION TEST  | PASS LOOKS LIKE                                                   | FAIL LOOKS LIKE                                            |
|-----------------|-------------------------------------------------------------------|------------------------------------------------------------|
| Problem clarity | A frequent, costly, well-understood task with a measurable output | A vague wish to use AI somewhere in the department         |
| Data readiness  | Accessible, reasonably clean data with a known owner              | Data scattered, undocumented or not permitted for this use |
| Value           | Savings or gains that clearly exceed build, run and change costs  | Benefits asserted but not measured                         |
| Risk fit        | Consequence of error is tolerable with a feasible review step     | Errors are severe and cannot be caught                     |
| Adoption        | Users want it and a manager will change the workflow              | Nobody owns the new way of working                         |

### USE THE INTAKE FORM FOR 'NO'

Not every request should become a project. Sometimes the right answer is a process fix, a better search tool or a spreadsheet. A good intake process says no early and leaves a record explaining why.

### WORKING QUESTION

What was the last AI idea you rejected, and is the reason written down anywhere?

## DELIVERY STAGES

# Five stages with an exit test at each gate

04

A system moves forward only when it passes the gate, and the gate is a decision, not a meeting.

Use five stages from idea to operation. The gates are lightweight for low-risk uses, often a checklist signed by the owner, and heavier for medium and high-risk ones, adding technical, privacy, security, legal and fairness review. Record every gate decision in the register with date, approvers, conditions and the next review date. Skipped records are the first thing an auditor or customer will notice.

| STAGE           | MAIN WORK                                                      | EXIT TEST                                                    |
|-----------------|----------------------------------------------------------------|--------------------------------------------------------------|
| 1. Frame        | Intake, owner, risk tier, success measure, budget              | Approved intake; owner and sponsor named; baseline known     |
| 2. Prove        | Small pilot on real data; compare with current process         | Meets pre-set accuracy and safety thresholds on the test set |
| 3. Build or buy | Select vendor or design; integrate security and privacy review | Review sign-off; contract terms agreed; oversight designed   |
| 4. Launch       | Training, controlled rollout, support fallback ready           | Launch checklist complete; users trained; rollback tested    |
| 5. Run          | Monitoring, feedback, change control, scheduled review         | Metrics on target; reviews on time; issues closed            |

## Build, buy or configure

- Buy or configure when the task is common and a mature product covers it; you inherit vendor risk but save build and maintenance cost.
- Build when the task is central to your advantage, the data is distinctive or the controls required cannot be met by a product.
- Avoid long custom builds before a pilot has shown the use case works with an off-the-shelf model.

- Plan for the model or vendor changing. Keep prompts, test sets and integrations documented so you can switch.



### Pilot design check

- A defined user group, time frame and success threshold
- A comparison with the current process, not with perfection
- A test set built from real cases, including difficult and rare ones
- A way for pilot users to flag bad outputs in one click
- A stop rule: the conditions under which the pilot ends early



#### WORKING QUESTION

Which gate does your organization skip when schedules slip, and what has that cost you?



#### VENDORS AND DATA

## Standards for buying AI and for preparing data

# 05

Most organizations buy more AI than they build, so vendor questions and data discipline carry most of the risk.

Ask the same questions of every vendor and keep the answers. Marketing claims are not evidence. Request documents, certifications, test results and contract language, and involve procurement, security and privacy early. Revisit vendors at renewal and when they announce material changes such as a new model version, a new sub-processor or a changed data use term.



### Vendor due diligence

- Is our data used to train or improve the vendor's models, and can that be switched off contractually?
- Where is data processed and stored, and which sub-processors are involved?

- What security attestations or audit reports are available and current?
- What accuracy and bias testing has the vendor done on tasks like ours, and can we see it?
- How are model changes announced, and can we test or pin versions?
- What are incident notification terms and response times?
- What are the terms on intellectual property, indemnity and ownership of outputs?
- How do we export our data, and what happens at exit?



## Data standards

- Record the owner, source, sensitivity tier, legal basis or permission and retention period for each dataset used.
- Check data quality before modeling: completeness, duplicates, outdated records and conflicting versions of the same fact.
- For assistants that answer from internal documents, maintain an owner and a review date for every source, and remove outdated material.
- Apply access controls so the assistant only retrieves what the asking user is allowed to see.
- Minimize personal data and apply the data protection principles that apply in your jurisdiction, with counsel's confirmation.



### COMMON MISTAKE

Connecting an assistant to a shared drive without cleaning it. The assistant will quote the old policy, the draft and the superseded price list with equal confidence. Clean sources first or limit the assistant to a curated set.



### WORKING QUESTION

Which vendor of an AI feature your people use daily has never answered these questions?



## EVALUATION AND OVERSIGHT

# Testing, human oversight and launch readiness

06

A system is ready when it passes a test someone wrote before seeing the results and a human can intervene meaningfully.

Write the evaluation plan before the pilot. Build a test set of real, representative cases with known correct outcomes, including hard and unusual ones. Agree the pass mark and what counts as an unacceptable failure, such as a fabricated source or a disclosure of restricted data. Test again after any change to the model, prompt, data or integration. Keep results, because they are the evidence for launch decisions and later audits.

| TEST TYPE         | WHAT IT CHECKS                                        | EXAMPLE METHOD                                         |
|-------------------|-------------------------------------------------------|--------------------------------------------------------|
| Accuracy          | Whether outputs are correct and complete              | Score against known answers; sample reviews by experts |
| Safety and misuse | Whether it can be pushed to harm or off-policy output | Adversarial prompts; red-team session; refusal checks  |
| Fairness          | Whether quality differs across groups affected        | Compare results across relevant groups; escalate gaps  |
| Security          | Whether data or actions can be exposed or hijacked    | Prompt injection tests; permission review; log review  |
| Usability         | Whether users can use it correctly                    | Observed sessions; error review; feedback              |

## Designing oversight that works

- Show reviewers the evidence behind an output, not only the output.
- Give reviewers authority and time to reject output without penalty.
- Sample for quality instead of asking for a rushed check of everything.
- Track override and rejection rates; near zero can mean excellence or inattention.
- For consequential actions, require human approval before irreversible steps such as sending, paying or deleting.
- Give affected people a way to contest outcomes and reach a person.

## Launch readiness

- Gate 3 sign-offs recorded, with conditions closed
- Test results meet the pre-set pass mark and are filed
- Owner, support contact and escalation route published
- Users trained and told what the system can and cannot do
- Disclosure wording approved where users or customers interact with AI
- Rollback or pause tested and the person authorized to use it named
- Monitoring live before the first user



### WORKING QUESTION

What was the pass mark for your last AI launch, and who set it before the results arrived?



### PEOPLE AND CHANGE

## Adoption, training and role changes

# 07



A technically sound system that people avoid delivers nothing, so plan the human side with the same rigor as the technical side.

Adoption is a management task. Involve the future users in the pilot and let them shape the workflow. Be honest about what the system will change in their roles. If tasks will move, say which and how people will be supported, whether by retraining, new responsibilities or changed targets. Fear fills silence. Tell people what the system is for, what it will not be used for and how their feedback will change it.



### 01 Map the workflow

Document the current process and the future one step by step, marking where AI acts and where a person decides.

## 02 Train for the real task

Run short, hands-on sessions using the team's own cases, covering strengths, known failure modes and how to check output.

## 03 Name champions

Pick respected users in each team to answer questions, collect feedback and model good review habits.

## 04 Update the surrounding rules

Revise job descriptions, targets, quality standards and handoffs so the new process is the official one.

## 05 Listen and adjust

Hold feedback sessions at two, six and twelve weeks and act visibly on what you hear.

### EXAMPLE

A support team gets an assistant that drafts replies. Handling time falls, but quality complaints rise because agents send drafts unread. The owner adds a rule that agents must verify the cited policy, changes the target from speed alone to speed plus accuracy, and adds a weekly sample review.

### WORKING QUESTION

Which team will carry the most change from your next AI system, and have they been asked what would make it work?

## RUN AND IMPROVE

## Monitoring, change control, incidents and roadmap

08



Once live, an AI system needs the same operational care as any critical service, plus attention to quality drift.

Models and data change over time, vendors update their products and user behavior shifts. A system that passed testing in spring may behave differently by autumn. Monitoring therefore tracks quality, usage, cost, complaints and incidents on a fixed rhythm, and reports to the owner and sponsor. Any significant change to the model, prompt, data source or integration goes through change control and a retest proportionate to the risk.

| MONITOR                           | FREQUENCY                               | ACTION IF OFF TARGET                       |
|-----------------------------------|-----------------------------------------|--------------------------------------------|
| Quality sample reviewed by humans | Weekly at launch, then monthly          | Investigate cause, retest, adjust or pause |
| User feedback and complaints      | Continuous, reviewed weekly             | Triage, fix and communicate                |
| Usage and adoption                | Monthly                                 | Retrain, simplify or reconsider value      |
| Cost per use or task              | Monthly                                 | Optimize, change model or limit use        |
| Incidents and near misses         | On occurrence, summarized monthly       | Contain, correct, update controls          |
| Scheduled risk review             | At least annually, sooner for high risk | Re-approve, change or retire               |



### Incidents and retirement

- Anyone may pause a system they believe is causing harm; the owner and policy lead decide on restart.
- Report incidents using a short note: what happened, who was affected, data involved, action taken, cause and prevention.
- Counsel decides on notification duties early, because data protection and sector rules can set short deadlines.



- Retire systems that are unused, unowned or past review. Remove access, archive records per retention rules and update the register.

**WORKING QUESTION**

Which of your live AI systems has gone longest without a documented quality check?

30 / 60 / 90

A first quarter that produces evidence

09



Roll these guidelines out by installing the front end of the process first, then proving it on live projects, then tightening operations.

**First 30 days: ground it**

- Build the register of existing AI systems and name an owner for each
- Adopt the intake form and the three risk levels
- Name the reviewers for technical, privacy, security and legal checks
- Publish the vendor due diligence questions to procurement

**Days 31 to 60: test it**

- Run every new request through the intake form and the stage gates
- Write evaluation plans and pass marks for the two most important pilots
- Train project teams and managers on the guidelines
- Run a launch readiness review on one system about to go live



## Days 61 to 90: decide

- Review one live system fully against the monitoring and change standard
- Close or retire systems without an owner or review date
- Report gate times, declines, incidents and quality to the sponsor
- Revise the guidelines based on what slowed teams down or let risk through



### GLOSSARY

## Terms used in this publication

# 10



| TERM                 | MEANING                                                                                                           |
|----------------------|-------------------------------------------------------------------------------------------------------------------|
| Stage gate           | A decision point where a project must meet a stated test before moving to the next phase.                         |
| Use-case intake      | A standard form describing the problem, owner, data, people affected, consequence of error and proposed approach. |
| Pilot                | A limited, time-boxed trial on real work, measured against the current process.                                   |
| Test set             | A fixed collection of real cases with known correct outcomes used to evaluate a system repeatedly.                |
| Red teaming          | Deliberate attempts to make a system fail or behave harmfully, in order to find weaknesses before users do.       |
| Prompt injection     | An attack in which text in an input or document tries to override a system's instructions.                        |
| Model drift          | A change over time in model behavior or input data that degrades performance.                                     |
| Change control       | A process to review, test and approve modifications to a live system.                                             |
| Human oversight      | Design that lets an informed person review, challenge and override AI outputs.                                    |
| Vendor due diligence | Structured questions and evidence reviewed before and during use of a supplier's AI.                              |
| Register             | The maintained record of AI systems, owners, risk levels, approvals and review dates.                             |



| TERM      | MEANING                                                                                                             |
|-----------|---------------------------------------------------------------------------------------------------------------------|
| EU AI Act | European Union regulation taking a risk-based approach to AI, with duties that vary by category of system and role. |



## A note on use

These guidelines are general operating guidance and not legal advice. Confirm current legal, regulatory and contractual requirements with qualified advisers before launch.

### TAKE THE NEXT STEP

Take the free Halden assessment at [halden.company/assessments](https://halden.company/assessments), or talk to us at [halden.company/contact](https://halden.company/contact) about turning this into a plan for your organization.

