Human oversight for AI-assisted work

Map where a person monitors, reviews, approves or stops an AI-assisted workflow.

Use one row per workflow. Start with the task, not the tool. Match oversight to the workflow's autonomy, stakes, data and possible impact if an output is wrong.

USE THIS AS A PRACTICAL OPERATING AID

It is not legal advice, a compliance test or a guarantee that a workflow is safe.

THREE OVERSIGHT MODES

MONITOR

A person samples outputs, checks for drift and records exceptions. Suitable only where errors are recoverable and the workflow does not act on its own.

HUMAN REVIEW

A person checks the output before it is used. They can edit, reject or request more evidence before the workflow continues.

MANDATORY APPROVAL

A named role must approve before a consequential action proceeds. Keep the decision, evidence and authorisation record.

HOW TO USE THE MATRIX

1. Define the task and intended output.
2. Record the data involved and who may be affected.
3. Choose monitor, human review or mandatory approval.
4. Name the accountable owner and reviewer.
5. Set triggers to pause, override, roll back or use a fallback.
6. Record evidence, incidents and review dates.

A responsible person remains accountable even where AI helps draft, recommend or act.

1. Define the work and the responsible person

Use short, factual entries. If you cannot describe the intended use or accountable owner, do not move to implementation.

Workflow / task	What recurring task is being supported?
Intended output	What should the workflow produce?
Intended use	What is the output allowed to do?
Foreseeable misuse	How could the workflow be misused or misunderstood?
Autonomy	Monitor / assistive draft / recommendation / action.
Stakes and impact	What could happen if the output is wrong, incomplete or delayed?
Data classification	Public / internal / confidential / personal / sensitive.
People affected	Who may be affected by the output or decision?
Accountable owner	Which role remains responsible for the workflow?
Oversight mode	Monitor / human review / mandatory approval.
Named overseer	Who checks, approves or intervenes?

2. Complete the control and make intervention explicit

Human oversight works best when the intervention point is specific. Record what causes a person to review, who can stop the workflow and how the work continues if the AI-assisted path is unavailable.

Pre-use check	What must be checked before the output is used?
Evidence / source	What source, record or test supports the output?
Review trigger	What requires human review or approval?
Accuracy check	How will the output be checked before use?
Approval threshold	Which outputs or actions cannot proceed without approval?
Pause / override	How does the authorised person stop or change the workflow?
Rollback / shutdown	How is the AI-assisted step disabled or reversed?
Fallback process	What manual or alternative process keeps work moving?
Monitoring record	What is checked, by whom and at what cadence?
Incident / escalation	Where are errors, concerns or suspected misuse reported?
Feedback / redress	How can an affected person ask questions, challenge or seek correction where relevant?

3. Keep evidence and revisit the workflow

Oversight does not end when a workflow begins. Record the evidence, escalation route and changes that make the original control no longer suitable.

Provider responsibility	What must a provider or contractor explain, support or notify?
Overseer training	What limitations and failure points must the reviewer understand?
Acceptance evidence	What testing or authorisation supports use?
Review trigger / date	What change or date causes the next review?
Decommission trigger	When should this workflow be retired or replaced?
Decision log	What changed, why, and who authorised it?

PRACTICAL RECORD-KEEPING

Keep prior approved versions and the reason for a change where auditability or safety matters. Escalate material privacy, security, legal or sector-specific uncertainty to qualified advice.

Read the control in a real workflow

ILLUSTRATIVE EXAMPLE - FICTIONAL ORGANISATION, NOT CLIENT EVIDENCE

Customer enquiry triage

AI drafts a suggested category and reply for staff review. It does not automatically send a message or make an eligibility decision.

Intended use	Draft support only; no automatic sending or eligibility decision.
Data rule	Do not enter personal or sensitive information into a publicly available GenAI tool.
Oversight	Mandatory human review and approval before any reply is sent.
Owner / reviewer	Customer Operations Lead / assigned service adviser.
Review triggers	Incomplete, unusual, sensitive, complaint-related or otherwise uncertain enquiry.
Intervention	Do not send the draft; edit or discard it; use the manual triage process.
Fallback	Standard manual triage template and escalation route.
Evidence	Approved response guidance and the original enquiry record.

GUARDRAIL

This example shows the kind of control to record. It does not establish that the control is sufficient for every business, sector, decision or legal obligation.